

THÈSE

en vue de l'obtention du : **DOCTORAT**

Centre de Recherche : Sciences et Technologies

Structure de Recherche : Laboratoire de Recherche en Informatique et
Télécommunications

Discipline : Sciences de l'ingénieur

Spécialité : Informatique et Intelligence Artificielle

Présentée et Soutenue le : 18 / 12 / 2024

par :

Hamza BOUZID

Generative AI for Modeling Spatiotemporal Dynamics in 2D/3D Human Emotions and Actions: Analysis and Synthesis

Devant le JURY :

Faissal EL BOUANANI	PES, Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Université Mohammed V de Rabat	Président
Stefano BERRETTI	PES, Département d'ingénierie de l'information, Université de Florence, Florence, Italie	Examineur/Rapporteur
Mohammed EL HASSOUNI	PES, Faculté des Lettres et des Sciences Humaines, Université Mohammed V de Rabat	Examineur/Rapporteur
Sanaa EL FKIH	PES, Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Université Mohammed V de Rabat	Examinatrice/Rapporteuse
Aladine CHETOUANI	PH, Ecole d'Ingénieurs Sup Galilée, Université Sorbonne Paris Nord, France	Examineur
Ghizlane ORHANOU	PES, Faculté des Sciences Rabat, Université Mohammed V de Rabat	Examinatrice
Naima OTBERDOUT	PA, Centre International d'Intelligence Artificielle du Maroc, Université Mohammed VI Polytechnique, Rabat	Invitée
Lahoucine BALLIHI	PH, Faculté des Sciences Rabat, Université Mohammed V de Rabat	Directeur de thèse

Année Universitaire : 2024 - 2025

Dedicace

To the gentle hands that shaped my world,

To my parents El Habib and Malika,

To my brothers Abdelillah, AbdelMoughit, and Tarik,

To my sisters Samira and Naima,

To all my beloved ones.

I am deeply grateful to my friends for their unwavering support. Your camaraderie and encouragement have been enduring sources of motivation and inspiration. Our stimulating discussions, diverse perspectives, and shared aspirations have profoundly enriched my journey and work.

To my family, no words can fully convey my heartfelt gratitude. You have been my unwavering foundation through every challenge, offering endless support, patience, and love. My parents, Malika and El Habib, your boundless sacrifices, prayers, and belief in me have shaped the person I am today. I owe you everything. Thank you for nurturing my dreams and standing by my choices with unwavering faith.

To my younger brothers, Abdelillah and Abdelmoughit, thank you for your patience, your humor, and the countless cups of tea that brightened my days. To my eldest brother, Tarik, thank you for your steadfast support and encouragement throughout these years of study. To my older sisters, Samira and Naima, thank you for your constant support and love. Your presence has been a source of immense joy, and I am forever indebted to you.

This dedication is a testament to all of you, who have been my strength, my guiding light, and my inspiration. Thank you from the depths of my heart.

Acknowledgment

My years as a PhD student have been a remarkable journey, enriched with knowledge and unforgettable experiences. Throughout this period, I have had the privilege of meeting exceptional individuals whose support, guidance, and mentorship have been instrumental in bringing this work to fruition. I would like to extend my heartfelt gratitude to all of them.

First and foremost, I express my deepest gratitude to my PhD supervisor, **Professor Lahoucine Ballihi**, at the Faculty of Sciences (FSR) in Rabat. Since my master's thesis, he has been an exemplary mentor. I am immensely thankful for the opportunity to undertake my PhD under his guidance and for his unwavering support, insightful advice, and dedicated supervision. His profound expertise, breadth of knowledge, and thoughtful mentorship have profoundly shaped my research and fostered my passion for this field. I am forever grateful for his encouragement and commitment to my academic and professional development.

I am also profoundly grateful to the esteemed members of my PhD committee, who have graciously agreed to evaluate my thesis. Their dedication, expertise, and constructive feedback have significantly enhanced the quality of this work. I extend my sincere thanks to **Professor Faissal El Bouanani**, at the National School of Computer Science and Systems Analysis (ENSIAS) in Rabat, for presiding over the committee and for his valuable insights and leadership throughout this process.

I express my heartfelt appreciation to **Professor Stefano Berretti**, at the Department of Information Engineering (DINFO) in Florence, Italy, for his meticulous review and detailed examination of my thesis. His time and expertise have greatly contributed to the rigor of this research.

My sincere gratitude also goes to **Professor Mohammed El Hassouni**, at the Faculty of Letters and Human Sciences (FLSH) in Rabat, for his thorough evaluation and constructive feedback, which have added immense value to this work.

I am equally thankful to **Professor Sanaa El Fkihi**, at the National School of Computer Science and Systems Analysis (ENSIAS) in Rabat, for her careful review, insightful comments, and valuable contributions to this thesis.

I also wish to express my gratitude to **Professor Aladine Chetouani**, at the Sup Galilée School of Engineers (Sup'Galilée) in Paris, France, for his expert evaluation and constructive suggestions, which have deepened my understanding of the subject matter.

Furthermore, I extend my thanks to **Professor Ghizlane Orhanou**, at the Faculty of Sciences (FSR) in Rabat, for her valuable feedback, constructive discussions, and for generously participating in my PhD defense.

Special acknowledgment goes to **Professor Naima Otberdout**, at the International Center for Artificial Intelligence (AI Movement) in Rabat. As my master's project supervisor, she played a pivotal role in shaping my early research career. Her presence on this committee as an invited member brings invaluable perspective, and her guidance throughout both my master's and PhD journeys has been instrumental in my growth as a researcher. I am deeply appreciative of her mentorship and unwavering support.

I would also like to acknowledge the entire teaching staff at the LRIT laboratory for their dedication to high-quality training and their inspiring commitment to research. Their expertise and enthusiasm have greatly enriched my academic journey. My heartfelt thanks go to **Professor Khalid Minaoui**, at the Faculty of Sciences (FSR) in Rabat and President of LRIT, for fostering a dynamic and supportive research environment and for his genuine care for the well-being and success of PhD students. I am equally grateful to my colleagues and friends at LRIT, whose camaraderie and collaboration have made this journey enjoyable and memorable.

Abstract

acial expressions play a crucial role in non-verbal human communication, and their integration into computer vision systems is essential for improving human-computer interactions. The automatic analysis of facial expressions has long attracted the attention of the computer vision community, opening up potential in fields such as security, robotics, and AR/VR. This thesis explores two intersecting tasks: facial expression recognition and generation, with some solutions extended to human actions. Our goal is to develop deep learning models that effectively capture both spatial information and temporal evolution, while optimizing resource usage. First, we present FEV-GAN, a model that uses spatio-temporal convolutional GANs to generate facial expression videos from neutral images while preserving identity. Although effective, this model produces low-resolution videos. To address this limitation, we introduce FacEnhance, which leverages a diffusion model to enhance the resolution and quality of the generated videos. Next, we propose LMDM, a latent diffusion model for generating facial expressions and human actions, by separating content and motion spaces to produce varied movements. Finally, we explore the 3D domain with SpATr, a model based on spiral convolutions and Transformers for classifying facial expressions and human actions in 3D mesh sequences. We also present LM3DM, a latent diffusion model for generating 3D facial expressions, which directly applies diffusion within the latent space to capture the dynamics of expressions. This work deepens the understanding, synthesis, and application of human non-verbal communication in computer vision. Future efforts will focus on improving the generation of high-resolution videos and 3D scans.

Keywords: Generative AI; Generative Adversarial Networks; Diffusion Models; Transformers; Facial expression analysis; Deep learning; 3D mesh Generation;

Résumé

Les expressions faciales jouent un rôle crucial dans la communication non verbale humaine, et leur intégration dans les systèmes de vision par ordinateur est essentielle pour améliorer les interactions homme-machine. L'analyse automatique des expressions faciales a depuis longtemps attiré l'attention de la communauté de la vision par ordinateur, ouvrant des perspectives dans des domaines tels que la sécurité, la robotique et l'AR/VR. Cette thèse explore deux applications aux défis interconnectés : la reconnaissance et la génération d'expressions faciales, avec des extensions vers les actions humaines. Notre objectif est de développer des modèles d'apprentissage profond qui capturent efficacement à la fois les informations spatiales et l'évolution temporelle, tout en optimisant l'utilisation des ressources. Nous présentons d'abord FEV-GAN, un modèle qui utilise des GAN convolutifs spatio-temporels pour générer des vidéos d'expressions faciales à partir d'images neutres tout en préservant l'identité. Bien que performant, ce modèle produit des vidéos de basse résolution. Pour pallier cette limitation, nous introduisons FacEnhance, qui exploite un modèle de diffusion pour améliorer la résolution et la qualité des vidéos générées. Nous proposons ensuite LMDM, un modèle de diffusion latente pour générer des expressions faciales et des actions humaines, en séparant les espaces de contenu et de mouvement afin de produire des mouvements variés. Enfin, nous explorons le domaine 3D avec SpATr, un modèle basé sur des convolutions spirales et des Transformers pour la classification des expressions faciales et des actions humaines dans des séquences de maillages 3D. Nous présentons également LM3DM, un modèle de diffusion latente pour la génération d'expressions faciales en 3D, qui applique directement la diffusion dans l'espace latent afin de capturer les dynamiques des expressions. Ce travail approfondit la compréhension, la synthèse et l'application des communications non verbales humaines en vision par ordinateur. Les travaux futurs viseront à améliorer la génération de vidéos haute résolution et de scans 3D.

mots clés : Intelligence Artificielle Générative; Réseaux Antagonistes Génératifs (GAN); Modèles de Diffusion; Transformers; Analyse des Expressions Faciales; Apprentissage Profond; Génération de Maillages 3D;

Résumé Détaillé

Les communications humaines non verbales, incluant les expressions faciales et les actions humaines, jouent un rôle important dans les interactions sociales. Ces formes de communication sont des outils essentiels pour transmettre des émotions, des intentions et des réactions, et sont ainsi indispensables à l'interaction sociale et à la compréhension. L'exploitation de ces signaux non verbaux dans les interactions homme-machine et leur intégration dans les systèmes de vision par ordinateur sont devenues un domaine de recherche passionnant et prometteur. La compréhension et la synthèse des comportements humains sont des prérequis essentiels pour des systèmes d'interaction homme-machine efficaces, ouvrant la voie à diverses applications dans des domaines tels que la santé, la psychologie sociale, la sécurité et la surveillance, le divertissement, et la robotique.

Malgré un intérêt considérable, l'analyse des communications humaines non verbales présente des complexités implicites, notamment dans des environnements réels. Les défis proviennent de la diversité culturelle, des conditions de santé et des attributs personnels influençant les comportements humains, intensifiés par les conditions de capture des données telles que les angles, les poses, les illuminations et les occlusions. Ces facteurs contribuent à des similitudes inter-classes et à des variations intra-classes importantes, rendant l'analyse et la synthèse précises particulièrement difficiles. Les approches basées sur l'apprentissage profond ont montré un potentiel prometteur pour adresser ces défis, mais des lacunes subsistent. Un problème critique est la modélisation efficace des dynamiques spatio-temporelles implicites aux comportements humains très complexes et diversifiés, ce qui nécessite des modèles profonds à haute complexité capables de capturer à la fois les informations spatiales et les dépendances temporelles avec précision et robustesse. De plus, la modélisation de ces comportements complexes nécessite des bases de données larges et variées. Le manque de données limite la généralisation des modèles d'apprentissage profond à des populations et contextes diversifiés, soulevant des préoccupations telles que le surapprentissage et la dégradation des performances lorsque les modèles sont entraînés avec des données insuffisantes.

Dans cette thèse, nous étudions deux tâches présentant des défis communs : la compréhension et la synthèse des communications humaines non verbales. Notre principal intérêt se porte sur les expressions faciales, avec certains modèles étendus aux actions humaines. Dans nos travaux, notre objectif est de développer des modèles efficaces qui optimisent l'utilisation des ressources.

Tout d'abord, nous introduisons FEV-GAN (Facial Expression Video GAN), un modèle qui utilise des GANs Convolutionnels Spatio-temporels pour synthétiser des vidéos cohérentes d'expressions faciales basiques à partir d'images neutres, tout en préservant l'identité. Cette méthode améliore la qualité des vidéos en intégrant des caractéristiques d'identité et spatiales, démontrant son efficacité dans la génération de vidéos de haute qualité par rapport à l'état de l'art. Cependant, les vidéos générées sont de basse

résolution. Ensuite, nous présentons FacEnhance, une approche qui répond aux limitations des modèles d'expressions faciales à faible résolution. En exploitant des méthodes basées sur la diffusion, FacEnhance augmente la résolution des vidéos générées tout en améliorant la qualité et en intégrant des détails d'arrière-plan. Cette approche trouve un équilibre entre l'efficacité computationnelle et la résolution et la qualité d'image, avec des expériences approfondies montrant sa capacité à améliorer les sorties des modèles à faible résolution pour atteindre une qualité de pointe. Nous proposons ensuite LMDM (Latent Motion Diffusion Model) pour la génération d'expressions faciales et d'actions humaines. Ce modèle exploite des techniques de diffusion latente, connues pour leur capacité à apprendre des espaces latents complexes et à intégrer une stochasticité contrôlée pour diversifier le contenu généré. En séparant l'espace d'embedding en domaines de contenu et de mouvement et en appliquant la diffusion uniquement à l'espace latent de mouvement, LMDM assure des mouvements variés tout en préservant les détails du contenu à travers différentes motions.

Ensemble, ces études approfondissent notre compréhension, synthèse et application des communications humaines non verbales dans la vision par ordinateur. Les efforts futurs se concentreront sur l'amélioration de notre capacité à générer des vidéos et des scans 3D en haute résolution. Cela implique le développement de modèles sophistiqués pour des systèmes haute performance, ainsi que de modèles efficaces optimisés pour des environnements à ressources limitées, élargissant ainsi leur potentiel de déploiement sur des systèmes à faible spécification.

Conjointement, ces travaux approfondissent notre compréhension, synthèse et application des communications humaines non verbales dans la vision par ordinateur. Les efforts futurs se concentreront sur l'amélioration de notre capacité à générer des vidéos haute résolution et des scans 3D. Cela implique de développer des modèles sophistiqués pour des systèmes haute performance, ainsi que des modèles efficaces optimisés pour des environnements aux ressources limitées, élargissant ainsi leur potentiel de déploiement sur des systèmes à faible spécification.

Acronyms

ACD-I	A verage C ontent D istance - I ntity
AE	A uto- E ncoder
CGAN	C onditional G enerative A dversarial N etworks
DCNN	D eep C onvolutional N eural N etworks
DDPM	D enoising D iffusion P robabilistic M odels
DM	D iffusion M odels
FC	F ully C onnected layers
FEA	F acial E xpression A nalysis
FEG	F acial E xpression G eneration
FER	F acial E xpression R ecognition
FVD	F réchet V ideo D istance
HAG	H uman A ction G eneration
HAR	H uman A ction R ecognition
GAN	G enerative A dversarial N etworks
LDM	L atent D iffusion M odels
MoCap	M otion C apture
PSNR	P eak S ignal-to- N oise R atio
ReLU	R ectified L inear U nit
RNN	R ecurrent N eural N etworks
SSIM	S tructural S imilarity I ndex M easure
Tr	T ransformers
VAE	V ariational A uto- E ncoders
WGAN	W asserstein G enerative A dversarial N etworks

List of Figures

1.1	The overall process for automatic facial expression analysis and generation.	9
1.2	Examples of the six basic Ekman emotions from the MUG dataset.	12
1.3	Examples of basic upper face action units and their combinations.	12
1.4	An example of the facial action 'fear' decomposition using FACS [1].	13
1.5	Plutchik's Wheel of Emotions.	13
1.6	Russell's circumplex model of Affect.	14
1.7	Examples of Representations of Facial Expressions and Human Actions in Computer Vision.	16
1.8	An example of CNN architecture [2].	18
1.9	The most used activation function in DCNNs.	19
1.10	Generated faces using a GAN model. All examples are entirely synthetic and do not correspond to real individuals [3].	20
1.11	General overview of a generative adversarial network.	20
1.12	Generated faces using a diffusion model. All examples are entirely synthetic and do not correspond to real individuals [4].	21
1.13	A graphical representation of diffusion models, highlighting the noise diffusion process $q(x_t x_{t-1})$ and the denoising process $p_\theta(x_{t-1} x_t)$	22
1.14	Generated faces using a latent diffusion model [5].	24
1.15	Transformer model overview [6].	25
1.16	self-attention mechanism in Transformer.	26
1.17	Applications of Meshes in 3D Modeling and Computer Graphics. From left to right: a detailed 3D character model from a video game, an airplane model , and a collection of mechanical parts.	28
1.18	Elements of a triangular 3D mesh.	29
1.19	Spiral ordering on a mesh and an image patch [7].	31
2.1	Evolution timeline of facial expressions datasets.	44

3.1	Overview of the proposed model for generating facial expressions: FEV-GAN . The input image is mapped through two encoders, identity (E_{Id}) and spatial (E_S) encoders, into features F_{Id} and F_S . Both codes and the label L are fed to the decoder G_{dec} that transforms them into a video of the input identity performing the desired expression. The discriminator D has the purpose of classifying realistic scenes from synthesized scenes.	50
3.2	The generation of facial expression videos on the MUG database (left) and Oulu-Casia (right). The image sequences show the six basic facial expressions of the same subject on the test dataset. The presented images are sampled every two frames.	57
3.3	Examples of videos generated by our model, displaying the six basic facial expression performed by a person given in the input image.	57
3.4	examples of sequences generated by our proposed model. The image sequences in each group show the six basic facial expressions performed by the same subject on the MUG test dataset.	58
3.5	Qualitative comparison of sequences generated by FEV-GAN model and by the state-of-the-art models on the MUG database. The sequences generated by our model (a, c, e, g, i), by VGAN (b), C-VGAN (d) and by ImaGINator (h) are randomly selected from the test results. The sequence generated by MoCoGAN (f) and by motionGAN (j) are taken from the original papers. All the images are sampled with the time step of 4.	59
3.6	Qualitative comparison of sequences generated by FEV-GAN model and by the state-of-the-art models on the MUG database. The sequences generated by our model (a, c, e, g, i, k, l), by C-VGAN (h) and by ImaGINator (j) are randomly selected from the test results. The sequence generated by MoCoGAN (b), (Otberdout et al., 2019) (h), G3AN (d), RV-GAN (f), and by VDSM (n) are taken from the original papers. The absence of input identity indicate that the baseline does not preserve the input identity.	60
3.7	Qualitative comparison of sequences generated by our full FEV-GAN model (a), by FEV-GAN w/o E_S (b), and by FEV-GAN w/o E_{Id} (c) on the MUG database.	62
3.8	Examples of flawed video generation by our model.	63
4.1	Examples of videos generated by VideoGAN, ImaGINator, motioGAN, and FEV-GAN.	66
4.2	the integration of our proposed model as an extension to enhance generated low-resolution videos.	67

4.3	Overview of the proposed facial expression enhancement model. The diffusion model refines Gaussian noise iteratively, guided by a low-resolution frame f_l^n from input video v_l , a neutral high-resolution image I_{Id} , and the previously generated frame $\tilde{f}_h^{n-1}$, resulting in an improved higher-resolution frame $\tilde{f}_{high}^n$	70
4.4	Video examples generated by our model showcasing the six basic facial expressions. The right side features generated videos of the six facial expressions performed by the same individual, while the left side presents six different individuals, with each person expressing one specific expression.	76
4.5	Examples of videos enhanced by our proposed model showcasing diverse identities performing the six basic facial expressions. These visual examples offer insights into the model performance and highlight its efficacy in enhancing realistic and diverse facial animations.	77
4.6	Qualitative comparison of facial expression sequences before and after enhancement using our model. On the far left of the figure, we display the input images. Adjacent to them are videos generated by VideoGAN (a), ImaGINator (b), MotionGAN (c), and FEV-GAN (d). On the right side, we present the corresponding enhanced videos by FacEnhance.	78
4.7	Qualitative comparison of sequences generated by the FacEnhance model and state-of-the-art models on the MUG database. The sequences of our model (a, c, e, g), ImaGINator (b), and LFDM (f) are randomly selected from the test results, and the sequences of VDM (c) and FDM (e) are taken from [8].	80
4.8	Examples of flawed frames from videos enhanced by our model.	82
4.9	Examples of our validation results at 512x512.	83
5.1	Flowchart of the LMDM framework for facial expression generation. The model consists of two stages: spatial generation (autoencoder on the left) and motion generation (diffusion on the right).	88
5.2	Flowchart of FAE.	90
5.3	Flowchart of latent diffusion.	91
5.4	Flowchart of inference.	92
5.5	Examples of videos generated by our model showcasing the six basic facial expressions, given the input identity and the target expression, on MUG and Oulu-Casia datasets.	95
5.6	Qualitative comparison of sequences generated by the LMDM model and the baselines on the MUG database.	96

5.7	Examples of different generated videos by our model using the same input image and expression label.	98
5.8	Examples of sequences generated using LMDM, during our work on human action generation.	99
5.9	Examples of video generation with noticeable flaws and inaccuracies.	100
6.1	Overview of the proposed model for 3D human action recognition. Firstly, each mesh from the action sequence is mapped through the spiral auto-encoder, which extracts its embedding. Next, the sequence of embeddings is utilized by the temporal transformer, which employs self-attention to capture temporal dependencies, enabling the prediction of the action class.	107
6.2	Examples of mesh registration.	112
6.3	confusion matrices showing the classification results of our proposed model on the test set of (a) MUG and (b) BU-4DFE databases. We note that the model is trained on 24 frames.	113
6.4	Examples of failed FLAME mesh registration from BU-4DFE database.	113
6.5	Pipeline of using the proposed model for multitasking: facial expression recognition, facial recognition, and sex prediction.	114
6.6	confusion matrices showing the classification results of our proposed model on the test set of (a) MoVi, (b) BMLrub, and (c) Babel databases. We note that the model is trained on 24 frames.	117
6.7	Examples of correct and faulty predictions.	118
6.8	Class distribution of Babel dataset.	119
6.9	Pipeline of using the proposed model for multitasking: action recognition, identity recognition, and sex recognition.	122
6.10	Evolution of model accuracy based on the number of frames.	122
6.11	Evolution of model accuracy based on the number of multi-head self-attention heads.	123
7.1	Our conditional Spiral autoencoder overview.	130
7.2	Motion latent diffusion overview.	131
7.3	Overview of the LM3DM in the inference phase.	132
7.4	Examples of generated sequences for the same subject performing the same expression, in each two lines.	134
7.5	Examples generated by LM3DM trained on COMA.	135

7.6	Examples of the same expression generated by LM3DM.	136
7.7	Comparison of the facial expression generation contributions proposed in this thesis, and baselines.	140

List of Tables

2.1	An overview of Some facial expression datasets.	44
3.1	Quantitative comparison results of FEV-GAN and baseline models using PSNR, SSIM, ACD, and ACD-I metrics.	56
3.2	Subjective comparison of FEV-GAN and baseline models. The reported results are the mean of the raters preferences.	61
3.3	Performance comparison of our model without E_{Id} and E_S	62
4.1	Quantitative results for facial expression enhancement. The table presents results for videos generated by low-resolution models (VideoGAN, ImaGINator, MotionGAN, FEV-GAN) both before and after enhancement using our proposed model.	79
4.2	Quantitative comparison results of FacEnhance and baseline models.	80
4.3	Performance comparison of the full model against variants that exclude the expression encoder, previous frame (t-1), and the noisy previous frame in the generation process.	81
5.1	Quantitative comparison results of LMDM and baseline models using FVD, ACD, and ACD-I metrics.	96
6.1	FACIAL EXPRESSION RECOGNITION ACCURACY (%) USING THE SAME FEATURES ON MUG DATASET.	114
6.2	Facial expression/Identity/sex recognition accuracy (%) using the same features on MUG dataset.	114
6.3	Action recognition accuracy (%) results on MoVi, BMLrub, and Babel datasets.	118
6.4	Action recognition accuracy (%) and memory usage/batch on MoVi dataset.	120
6.5	Ablation study results.	120

Table of Contents

Dedicace	i
Acknowledgment	ii
Abstract	iv
Résumé	v
Résumé Détaillé	vi
Acronyms	viii
List of Figures	ix
List of Tables	xiv
Introduction	1
Problematic & Motivation	1
Thesis Contributions	3
Structure of the Thesis	5
List of Publications	7
1 Taxonomy and Background	9
1.1 Facial Expressions in Computer Vision	9
1.1.1 Introduction to Facial Expression Recognition (FER)	9
1.1.2 Introduction to Facial Expression Generation (FEG)	10
1.1.3 Categorizing Facial expressions in Computer Vision	11
1.2 Representations of Facial Expressions in Computer Vision	14
1.2.1 Image-Based Representations	14
1.2.2 Geometric-Based Representations	15
1.2.3 Motion-Based Representations	15
1.2.4 3D Representations	16

1.3	Relevant Concepts from Deep Learning	16
1.3.1	Deep Convolutional Neural Networks (DCNNs)	17
1.3.2	Generative Adversarial Networks (GANs)	19
1.3.3	Diffusion Models	21
1.3.4	Latent Diffusion Models	24
1.3.5	Transformers	25
1.3.6	Mesh Representation and Spiral Convolutions	27
1.4	Conclusion	32
2	Literature Review	33
2.1	Introduction	33
2.2	Facial Expression Recognition (FER)	33
2.3	Facial Expression Generation (FEG)	36
2.4	Facial Expression Databases	42
2.5	Conclusion	44
3	Dynamic Facial Expression Generation Based on Spatio- temporal Convolutional GANs	46
3.1	Introduction	46
3.2	Related Work	48
3.3	Proposed Approach	49
3.3.1	Generator E_{Id} , E_S , G_{dec}	50
3.3.2	Discriminator D	51
3.3.3	Loss Function	51
3.3.4	Training Algorithm	53
3.4	Experiment	54
3.4.1	Datasets	54
3.4.2	Implementation Details	54
3.4.3	Evaluation Metrics And Baselines	55
3.4.4	Experimental Results & Analysis	56
3.4.5	Discussion and Limitations	62
3.5	Conclusion & Perspectives	63

4	Facial Expression Enhancing with Recurrent DDPMs	65
4.1	Introduction	65
4.2	Related Work	68
4.3	Proposed Approach	69
4.3.1	Our proposed Model: FacEnhance	69
4.3.2	Training	72
4.3.3	Inference	72
4.3.4	Video Inference	73
4.4	Experiment	74
4.4.1	Dataset	74
4.4.2	Implementation Details	74
4.4.3	Evaluation Metrics	75
4.4.4	Experimental Results	75
4.4.5	Discussion & Limitations	82
4.5	Conclusion & Perspectives	83
5	Latent Motion Diffusion Models for Facial Expression Synthesis	85
5.1	Introduction	85
5.2	Related Work	87
5.3	Proposed Approach: LMDM	88
5.3.1	Flow Conditioned Autoencoder (FAE)	89
5.3.2	Proposed Flow Diffusion Model	90
5.3.3	Inference	92
5.4	Experiment	93
5.4.1	Datasets	93
5.4.2	Implementation Details	94
5.4.3	Evaluation Metrics	94
5.4.4	Experimental Results	95
5.4.5	Attempt for Extension to Human Action Generation	97
5.4.6	Discussion & Limitations	100
5.5	Conclusion & Perspectives	101

6	MoCap 3D Human Action Recognition Based on Spiral Auto-encoder and Transformer Network	102
6.1	Introduction	102
6.2	Related Work	105
6.2.1	3D Facial Expression Recognition	105
6.2.2	3D Human Action Recognition	105
6.3	Proposed Approach	107
6.3.1	Spiral Auto-Encoder	108
6.3.2	Temporal Transformer	109
6.3.3	Training Algorithm	110
6.4	Experiment: 3D FER	110
6.4.1	Datasets	110
6.4.2	Implementation Details	111
6.4.3	Experimental Results	112
6.5	Experiment: 3D HAR	115
6.5.1	Datasets	115
6.5.2	Implementation Details	115
6.5.3	Experimental Results	116
6.5.4	Ablation Study	120
6.5.5	Multi-Tasking Evaluation	121
6.5.6	Parameters Analysis	121
6.6	Discussion	123
6.7	Conclusion & Perspectives	124
7	Latent-Motion 3D Diffusion Model for 4D Facial Expression Generation	125
7.1	Introduction	125
7.2	Related Work	127
7.3	Proposed Approach	128
7.3.1	Spiral Auto-Encoder	129
7.3.2	Proposed Flow Diffusion Model	130
7.3.3	Inference	131
7.4	Experiment	132

7.4.1	Datasets	132
7.4.2	Implementation Details	133
7.4.3	Experimental Results	133
7.5	Conclusion & Perspectives	137
Conclusions and Perspectives		137
	Conclusion	138
	Perspectives and Future Works	141
References		143

Introduction

Problematic & Motivation

Nonverbal human behavior, including facial expressions, gestures, posture, and other bodily cues, plays a vital role in communication, often conveying emotions, intentions, and social signals more powerfully than words independently. The importance of understanding nonverbal communication has long been recognized. Darwin, for instance, highlighted its evolutionary significance: "We have also seen that expression in itself, or the language of the emotions, as it has sometimes been called, is certainly of importance for the welfare of mankind" [9]. Similarly, Mehrabian et al. [10, 11] found that, in emotional exchanges, facial expressions often take precedence over verbal content, with receivers placing greater trust in visual cues. Watzlawick et al. [12] further emphasized that "one cannot not communicate", reinforcing that all human actions carry communicative significance, and Wiener et al. [13] observed that even subtle bodily movements contribute to the communication process, revealing underlying emotions and intentions.

In computer vision, the ability to interpret and synthesize nonverbal cues has become a vital focus, aiming to enhance human-computer interactions. By enabling machines to understand and imitate human nonverbal behavior, we can develop systems that interact with users in a more intuitive and compassionate manner. This advancement has opened the door to various applications in different fields. Within the field of education, it can enhance personalized learning, monitor student engagement, support emotional well-being, improve classroom management, and provide comprehensive performance assessments. Additionally, it raises privacy and ethical considerations and enables the creation of Virtual Tutors and Teaching Assistants for online interactive learning. In medicine, it aids in diagnosing neurological disorders, assessing pain, and developing Virtual Therapists for managing social anxieties. In entertainment, it enhances character animation and NPC behavior, while in the metaverse, it enables realistic avatar interactions and helps monitor harmful behaviors. Overall, these applications, along with many others, highlight the significant impact of interpreting and generating nonverbal behavior in various fields.

Despite the growing interest in nonverbal communication analysis, particularly focusing on facial expressions and body actions, researchers in the field face numerous challenges. One primary issue is the variability in data capture conditions, such as differing body and head poses, occlusions, varying illumination, and misalignment. Additionally, cultural backgrounds and individual anatomical differences contribute to this variability. The complexity increases significantly with dynamic motions compared to static poses. Dynamic movements involve complex variations and increased individual differences,

making the learning process more complex. Understanding these dynamic motions requires handling temporal dependencies while preserving spatio-temporal content. To overcome these challenges, solutions must effectively capture both spatial information and temporal dynamics, ensuring accurate representation and preservation of the input data.

Machine learning has found it very difficult to address these challenges, often resulting in limited performances. This is due to the relatively low complexity of traditional ML models and the need for handcrafted feature extraction, where domain knowledge is used to define features relevant to the task. This approach can limit the ability to capture non-linear highly complex patterns and variations present in human behaviors. Additionally, handcrafted feature extraction struggles to generalize to variations in facial expressions and human actions and to unseen data.

With the emergence of deep learning, particularly Convolutional Neural Networks (CNNs) designed for image data, there has been remarkable progress in various computer vision applications, leading to significant advancements in the field. Therefore, extensive research has been conducted on human behavior, yielding interesting results. However, deep learning approaches have also encountered some challenges that impact their performance in human action and facial expression analysis. One major issue is the high complexity and diversity of nonverbal human behaviors, which necessitates models with considerable complexity and numerous parameters to capture fine-grained details. This, in turn, creates another challenge: the need for vast amounts of high-quality labeled data. Acquiring such extensive labeled datasets for effective training of deep models remains a non-trivial task. As a result, the focus has shifted to synthetically generated data. Beyond data augmentation, the ability to automatically synthesize human behavior has numerous applications in areas such as human-computer interaction, medical field, entertainment, and avatar impersonation.

Recently, the success of generative models like Generative Adversarial Networks (GANs) and Diffusion Models (DM) has led to remarkable progress in various image generation tasks, including the synthesis of static human actions and facial expressions. Despite this progress, most proposed solutions have primarily addressed static facial expression generation, while the dynamic synthesis of facial expressions is less explored. Given that human actions are more accurately described by dynamic motions rather than static poses, capturing realistic human behavior dynamics in video form presents a substantial challenge. This task is notably challenging due to the difficulty of modeling the continuous temporal evolution of human motions. Generating a facial expression or human action video from a single image is a one-to-many problem, where the output involves much more complexity and variability than the static input, which lacks temporal context. Moreover, the challenge extends beyond creating realistic images to synthesizing continuous frames that transition smoothly over time without sudden changes. Therefore, there is a critical need to define representations that can effectively capture the temporal evolution of human behaviors.

Generative deep learning algorithms have achieved remarkable success across various domains. Some advanced models have demonstrated the ability to produce high-resolution videos with plausible motions, reflecting their considerable potential. However, these models are still undergoing refinement and face several challenges that limit their applicability. High-resolution video generation models, while capable of impressive outputs, are often complex, requiring extensive data and **high-end hardware** for long training periods, which can last from weeks to months. This complexity limits their practicality and accessibility for many use cases. Conversely, more efficient models exist that can generate videos at lower resolutions and with fewer frames. While these models are less demanding in terms of computational resources, they often sacrifice video quality, resolution, and sequence length. A major challenge shared by both types of models is maintaining temporal consistency throughout the video generation process. This challenge involves ensuring smooth transitions between frames and preserving the identity of the input across the entire video sequence. Current methods frequently struggle with these issues, leading to videos that may display unnatural movements or inconsistencies in appearance over time.

Building on this background, leveraging deep learning models for human behavior analysis and synthesis has raised several critical research questions that this thesis aims to tackle. These questions include: (1) How to design efficient models capable of synthesizing coherent and high-quality videos of nonverbal human behavior from single images while preserving the subject identity? (2) Can we effectively enhance the results of state-of-the-art methods from low-resolution or low-quality videos to achieve higher-resolution and higher-quality outputs? (3) What strategies can be employed to obtain sufficient data for training these models effectively? (4) Is it possible to extend these techniques to encompass 3D mesh representations of the human body? These questions form the foundation of this research, focusing primarily on the recognition and synthesis of dynamic nonverbal human behavior. Particularly, we focus on facial expressions, with some proposed solutions extended to include human actions.

Thesis Contributions

This thesis introduces approaches designed to advance facial expression recognition and generation in both 2D and 3D domains, with some methods extended to include human action. The major contributions of the thesis are summarized as follows:

- **Comprehensive Review of State-of-the-Art Methods.**

A thorough state-of-the-art (SOTA) review has been conducted, covering facial expression recognition and synthesis. This review provides a comprehensive background overview, followed by an discussion of the datasets used in current SOTA research, highlighting their features and associated challenges. In addition, we cover

feature extraction techniques, including both hand-crafted feature descriptors and deep learning approaches, as well as recognition and analysis methods, ranging from classical models to the latest advancements. The synthesis methods are explored with a focus on generative models such as variational autoencoders, generative adversarial networks, and diffusion models, detailing their applications in both 2D and 3D domains.

- **FEV-GAN: Facial Expression Video Generation Using Spatio-temporal GANs.**

A novel approach for generating videos of the six basic facial expressions (happiness, sadness, anger, fear, disgust, and surprise). Starting from a single neutral facial image and a label indicating the desired expression, we synthesize a video of the individual performing the specified expression. Our approach, **FEV-GAN** (Facial Expression Video GAN), leverages Spatio-temporal Convolutional GANs, which are proficient at modeling both content and motion within a single network. Previous methods using similar networks have demonstrated the ability to generate coherent videos with smooth temporal evolution. However, these methods often face challenges in maintaining high image quality and input subject identity preservation across frames. To address this, our generator incorporates two image encoders: one pre-trained for facial identity feature extraction and the other for spatial feature extraction. The experimental results demonstrate the effectiveness of our approach in generating videos of the six basic facial expressions while preserving the input identity. The analysis also shows that utilizing both identity and spatial features enhances the decoder ability to better preserve identity and generate high-quality videos.

- **FacEnhance: Diffusion-Based Video Enhancement for Low-Resolution Facial Expression Videos.**

A novel diffusion-based approach, **FacEnhance**, for enhancing low-resolution facial expression videos (64x64 pixels) to higher resolutions (192x192 pixels), incorporating background details and improving overall quality. The approach leverages conditional denoising within a diffusion framework, guided by a background-free low-resolution video and a single neutral expression high-resolution image. FacEnhance generates a video that incorporates the facial expression from the low-resolution video performed by the individual, with background details from the neutral image. By complementing lightweight low-resolution models, FacEnhance strikes a balance between computational efficiency and desirable image resolution and quality. Experimental results demonstrate the efficacy of FacEnhance in enhancing low-resolution model outputs to state-of-the-art quality while preserving content and identity consistency.

- **LMDM: Latent-Motion Diffusion Model for Facial Expression Generation.**

We introduce a new approach for generating facial expressions from a single input image and a target expression label. Our method, referred to as Latent-Motion Diffusion Model (LMDM), leverages latent diffusion techniques known for their ability to learn complex latent spaces and integrate controlled stochasticity to diversify generated content. The main idea of LMDM is separating the embedding space into identity and motion domains, and applying diffusion to the motion latent space only. This strategy enhances our model capability to generate various facial expressions while ensuring that the identity details remain consistent across different expressions. The results highlight the competitiveness of the proposed model compared to state-of-the-art methods, demonstrating superior visual quality and notable diversity in the generated expressions.

- **SpATr: Spiral Auto-encoder and Transformer Network for 3D Facial Expression and Action Recognition.**

A novel approach for 3D facial expression and human action recognition, denoted as SpATr (Spiral Auto-encoder and Transformer Network), specifically designed for fixed-topology mesh sequences. SpATr model disentangles space and time in the mesh sequences. A lightweight auto-encoder, based on spiral convolutions, is employed to extract spatial geometrical features from each 3D mesh. These convolutions are optimized for fixed-topology mesh data, ensuring efficiency and scalability. Subsequently, a temporal transformer, based on self-attention, captures the temporal context within the feature sequence. The self-attention mechanism enables the capturing of long-range dependencies and parallel processing, allowing for scalability with long sequences. Our results demonstrate the competitive performance of our SpATr model in 3D facial expression and human action recognition while maintaining efficient memory usage.

- **LM3DM: Latent-Motion 3D Diffusion Model for Facial Expression Generation.**

We introduce a novel approach, referred to as LM3DM (Latent-Motion 3D Diffusion Model), for 3D facial expression generation. LM3DM directly operates within the latent space of a mesh generator, unlike traditional models that first generate landmarks and use them as conditions for mesh generation. Additionally, our model focuses applying diffusion models to learn the motion embedding space solely, introducing stochasticity for more varied and smoother motions while maintaining the subject’s identity throughout the mesh sequences. This is achieved by decoupling the embedding space into content and motion components using two separate encoders: one for capturing static features from an initial pose and another for processing landmark vectors that drive motion. After establishing these embedding spaces, diffusion is applied exclusively in the motion space to learn and generate new expression patterns.

Structure of the Thesis

The remaining of this manuscript is organized as follows:

- **Chapter 2:** This chapter provides the foundational background necessary to understand the thesis content. It covers the fundamental concepts related to facial expressions, including how they are represented and processed in computer vision. Additionally, it presents and explains the deep learning concepts used in the models and methods employed throughout the thesis.
- **Chapter 3:** In this chapter, we review and discuss the state-of-the-art research relevant to facial expression recognition and generation. The discussion covers techniques used for both 2D and 3D data, with a focus on the progress made in the field, the limitations of existing approaches, remaining challenges, and the research gaps that motivate the contributions of this thesis.
- **Chapter 4:** We introduce FEV-GAN, an approach for generating videos of the six basic facial expressions from a single neutral facial image and an expression label. This chapter details how we exploit Spatio-temporal Convolutional GANs to improve video quality and identity preservation, employing two image encoders to extract both facial identity and spatial features.
- **Chapter 5:** This chapter presents FacEnhance, a novel diffusion-based approach for enhancing low-resolution facial expression videos. FacEnhance improves video quality and resolution, incorporating background details and preserving content and identity consistency. By leveraging conditional denoising within a diffusion framework, it balances computational efficiency with high-quality output.
- **Chapter 6:** We introduce in this chapter the Latent-Motion Diffusion Model (LMDM) for generating facial expression videos from a single starting image and a target motion label. LMDM leverages latent diffusion to learn complex latent spaces, ensuring diversity in the generated content by separating identity and motion domains and applying diffusion to the motion latent space.
- **Chapter 7:** In this chapter, We introduce SpATr (Spiral Auto-encoder and Transformer Network) for 3D facial expression and human action recognition in fixed-topology mesh sequences. SpATr disentangles space and time in mesh sequences. It uses a lightweight auto-encoder with spiral convolutions to extract spatial geometrical features and a temporal transformer with self-attention to capture temporal context.
- **Chapter 8:** This chapter introduces LM3DM (Latent-Motion 3D Diffusion Model) for 3D facial expression generation. LM3DM disentangles space and time in mesh sequences. Then, It uses a diffusion models to learn the motion embedding space solely.

- **Chapter 9:** The final chapter offers concluding remarks and perspectives on the current work.

List of Publications

[1] **Hamza Bouzid** and Lahoucine Ballihi. "Facial expression video generation based-on spatio-temporal convolutional gan: Fev-gan." *Intelligent Systems with Applications* 16 (2022): 200139.

[2] **Hamza Bouzid** and Lahoucine Ballihi. "SpATr: MoCap 3D human action recognition based on spiral auto-encoder and transformer network." *Computer Vision and Image Understanding* 241 (2024): 103974.

[3] **Hamza Bouzid** and Lahoucine Ballihi. "FacEnhance: Facial Expression Enhancing with Recurrent DDPMs." *arXiv preprint arXiv:2406.09040* (2024). (Submitted to *Multimedia Tools and Applications*)

[4] **Hamza Bouzid** and Lahoucine Ballihi. "3D Facial Expression Recognition Using Spiral Convolutions and Transformers." *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 2023.

[5] **Hamza Bouzid** and Lahoucine Ballihi. "Motion-Oriented Diffusion Models for Facial Expression Synthesis." *2024 13th International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2024. (accepted)

[6] **Hamza Bouzid** and Lahoucine Ballihi. "Decoupling Spatial and Temporal Modeling in 3D Facial Expression Recognition." *Affinity Workshop NAML (North Africans in Machine Learning Workshop) of NeurIPS 2023*. 2023.

Chapter 1

Taxonomy and Background

In this chapter, we establish the foundational knowledge necessary to comprehend the content of this thesis. We begin by introducing the essential concepts of facial expressions within computer vision. We then explore the relevant concepts of deep learning, with a particular focus on Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), Diffusion Models, Transformers, and Spiral convolutions.

1.1 Facial Expressions in Computer Vision

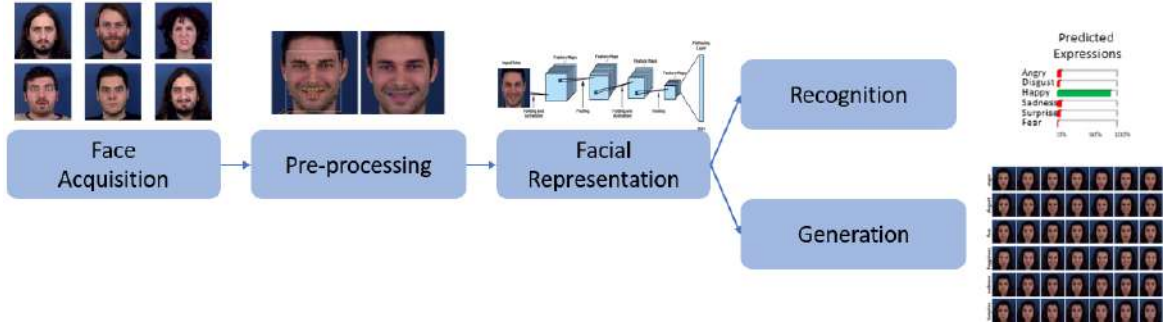


Figure 1.1: The overall process for automatic facial expression analysis and generation.

1.1.1 Introduction to Facial Expression Recognition (FER)

Definition

Facial Expression Recognition (FER) is a crucial part of human-computer interaction that enables computers to automatically identify and recognize human emotions based on facial expressions using computer vision and machine learning techniques [14]. This automated process **typically** involves several key steps:

1. **Face Detection:** Locating and isolating faces in an image or video frame.
2. **Feature Extraction and Analysis:** Identifying and extracting facial features manually or automatically, analyzing the configuration and movement of these facial features using machine learning or deep learning models.

3. **Expression Classification:** Categorizing the detected facial expression into predefined emotional categories.

Applications

FER leverages cutting-edge advancements in computer vision and biometric analysis to enable real-time detection and classification of emotions, finding diverse applications across various fields:

- **Human-Computer Interaction:** Assessing user experience and enhancing interactions by enabling systems to respond to users' emotional states.
- **Healthcare:** Assisting in mental health diagnosis and monitoring, detecting conditions like depression, anxiety, or stress, and providing feedback for therapies.
- **Education:** Personalizing learning experiences by monitoring student engagement and emotions, providing real-time feedback to educators.
- **Marketing and Advertising:** Analyzing consumer emotional reactions to products or advertisements, enabling companies to design their offerings based on emotional responses.
- **Security and Surveillance:** Enhancing security systems by identifying suspicious behavior or stress indicators in surveillance footage.
- **Automotive Industry:** Improving driver safety by monitoring drivers' emotional states and fatigue levels to prevent accidents.
- **Social Robotics:** Developing robots that can interact more naturally with humans by recognizing and responding to their emotional expressions.
- **Research and Psychology:** Studying human emotions and behaviors, providing insights into social and psychological processes.
- **Crime Detection:** Identifying potential criminal behavior (such as insurance fraud, shoplifting, and other suspicious activities) by analyzing emotional cues and stress indicators.

1.1.2 Introduction to Facial Expression Generation (FEG)

Definition

Facial Expression Generation is an essential aspect of human-computer interaction that allows computers to create and manipulate facial expressions on virtual avatars or images of real human faces. The goal is to generate realistic and expressive faces that can convey emotions dynamically or statically. The generated expressions can be conditioned based on various input data, such as emotion labels, action units (describing specific muscle

movements), or even user-provided expressions from another face.

This automated process **usually** involves several key steps:

1. **Input Feature Extraction:** Identifying and extracting facial features from a input image, either manually or automatically.
2. **Conditioning on Target Expression:** Encoding and incorporate the target expression information into the feature space.
3. **Expression Generation:** Modifying the configuration and movement of the facial features to create the target expression, using machine learning or deep learning models.

Applications

FEG utilizes advancements in computer vision and generative models to produce lifelike facial expressions, enabling various applications:

- **Entertainment and Media:** Generating dynamic facial expressions for digital characters in movies, games, and in virtual reality to enhance user experiences.
- **Healthcare and Therapy:** Helping individuals with disabilities communicate more effectively by generating expressions that convey emotions.
- **Marketing and Advertising:** Creating personalized advertisements that adapt to the viewer's emotional responses.
- **Education and Training:** Generating facial expressions for digital tutors to make the experience more realistic and interactive.
- **Social Robotics:** Designing robots that can generate appropriate facial expressions to enhance social interactions and improve the user experience.
- **Research and Development:** Advancing the field of affective computing by developing more sophisticated models for generating and understanding human emotions, and exploring new ways to integrate emotional expression into human-computer interactions for research purposes.

1.1.3 Categorizing Facial expressions in Computer Vision

The categorization of facial expressions in computer vision typically follows several key models, each offering unique perspectives on how emotions are represented and classified.

- **Ekman's Six basic facial expressions:** Psychologist Paul Ekman identified six fundamental emotional expressions that are universally recognized across different cultures. According to his research [15, 16], these expressions are biologically hardwired in humans and serve as universal markers of emotions. The majority of studies on facial

expressions have examined the perception of expressions in static photographs. Most of these studies suggest six "universal" categories of expressions that can be discriminated by people of all cultures. The six basic emotions—anger, disgust, fear, happiness, sadness, and surprise—are depicted in Fig.1.2.

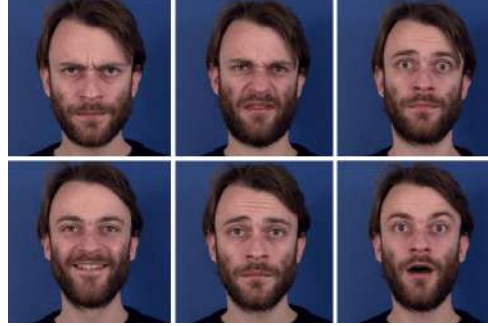


Figure 1.2: Examples of the six basic Ekman emotions from the MUG dataset.

These expressions are regarded as fundamental because they transcend cultural boundaries and are universally shared. However, they do not encompass the full spectrum of human emotions, as more complex or mixed emotions (e.g., trust, shame, or embarrassment) are not accurately captured by these basic categories.

- **Action Units from The Facial Action Coding System:** FACS, developed by Ekman and Friesen, is a widely used protocol to study and categorize facial expressions based on the movement of facial muscles, without relating them to emotional meaning [17]. The system divides any facial expression into small discriminable movements called Action Units (AUs). Each unit of actions is linked to a component of the face: 9 units in the upper part of the face and 18 in the lower part, in addition there are 11 AUs for the positions and movements of the head, 9 others for the positions and eye movements, 4 various action units, 9 full facial behaviors, 4 visibility and other additional codes [18]. Each Action Unit is uniquely identified by a numerical code, with several AUs capable of being activated simultaneously to form complex expressions. Fig.1.3 shows examples of eye-related AUs.

NEUTRAL	AU 1	AU 2	AU 4	AU 5	AU 1+2+4	AU 1+2+5
Eyes, brow, and cheeks are relaxed	Inner portion of brows is raised	Outer portion of the brows is raised	Brows lowered and drawn together	Upper eyelids are raised	Brows are pulled together and upward	Brows and upper eyelids are raised
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5	AU 1+6	AU 6+7
Cheeks are raised	Lower eyelids are raised	Inner and outer portions of the brows are raised	Medial portion of the brows is raised and pulled together	Brows lowered and drawn together and upper eyelids are raised	Inner portion of brows and cheeks are raised	Lower eyelids cheeks are raised

Figure 1.3: Examples of basic upper face action units and their combinations.

To fully label the entire expression, the set of AUs displayed are aggregated into a single numerical representation. This comprehensive labeling provides a detailed breakdown of facial expressions. For example, Fig.1.4 illustrates how the emotion of fear can be deconstructed into individual Action Units using FACS.

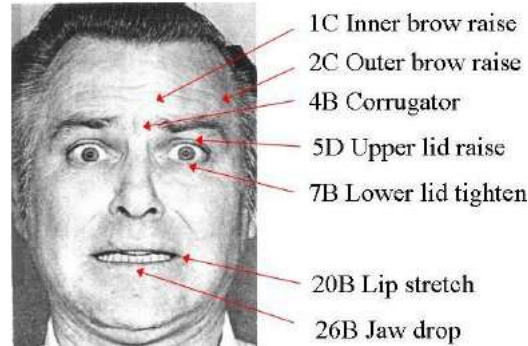


Figure 1.4: An example of the facial action 'fear' decomposition using FACS [1].

- **Plutchik's Wheel of Emotions:** displayed in Fig.1.5, is a model developed by psychologist Robert Plutchik to visually represent the relationships between primary emotions and their combinations. This model shows how emotions blend and interact to form more complex emotional states. The wheel helps to understand the spectrum of human emotions and their intensity [19].

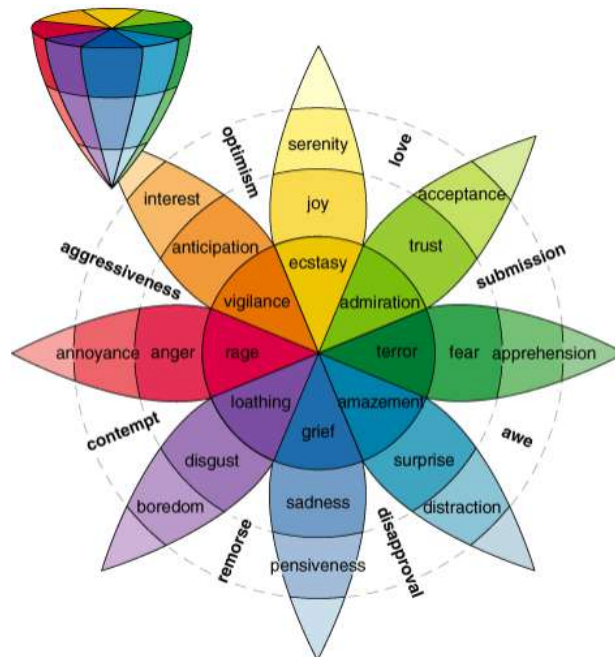


Figure 1.5: Plutchik's Wheel of Emotions.

The Wheel provides a comprehensive framework to understand how basic emotions interact and combine to form the rich and complex space of dynamic human emotions.

- **Russell's circumplex model of Affect [20]:** illustrated in Fig.1.6, is a psychological model that organizes emotions along two continuous dimensions: valence (pleasantness)

and arousal (activation). This model provides a two-dimensional framework to map the relationships between emotions and visualize how they vary in intensity and pleasantness.

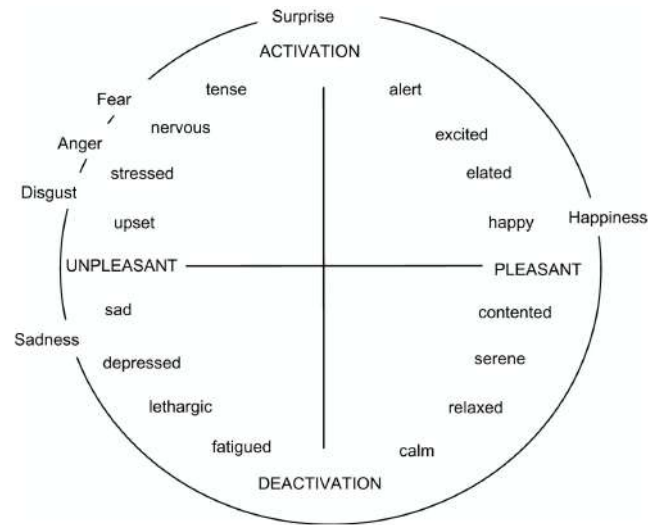


Figure 1.6: Russell's circumplex model of Affect.

By mapping emotions along these two axes, the model highlights fluid and dynamic nature of emotional experiences. It shows how emotions can be related to one another based on their levels of activation and pleasantness, offering a comprehensive view of emotional states and transitions.

1.2 Representations of Facial Expressions in Computer Vision

Facial expressions are a crucial aspect of nonverbal human communication, and in computer vision, they involve interpreting various facial movements captured in different types of data representations. These representations can be broadly categorized into multiple types, such as pixel-based, geometric-based, and 3D representations. Below, we explore some of the most common data types used for representing facial expressions in computer vision.

1.2.1 Image-Based Representations

Image-based representations use an image or a set of images to represent the visual content captured in nonverbal human behaviors. This approach involves using raw pixel values of an image as input for processing and analysis, leveraging the detailed visual information to understand and interpret Nonverbal human behaviors.

- **Advantages:** Image-based representations are of common use and easy to capture with cameras, smartphones, and other devices, making image-based approaches accessible

for many applications. In addition, they capture rich visual information, including texture, color, and shading, which are crucial for identifying Nonverbal human behaviors. This direct representation simplifies the processing pipeline and integrates well with convolutional neural networks (CNNs), which excel in handling pixel data.

- **Disadvantages:** However, they do not offer depth information, and they represent an object or scene from only one perspective, limiting the ability to capture or understand its full geometry or spatial layout. Additionally, these representations are sensitive to variations in lighting, pose, and occlusions, which can negatively impact the accuracy of expression recognition.

1.2.2 Geometric-Based Representations

Geometric-based representations use the spatial relationships and structural features of key points on the face (facial landmarks) or body (skeleton data) to represent nonverbal human behaviors. This approach involves identifying and analyzing landmarks or shapes to capture the geometry of facial expressions.

- **Advantages:** Geometric-based representations are robust to variations in lighting, pose, and expression intensity. They are efficient in terms of data storage and computational power, as they deal with a smaller set of key points. This method also provides a clear and interpretable representation of movements and deformations.

- **Disadvantages:** However, they may miss subtle texture and shading details that are important for recognizing fine-grained actions. The accuracy of these representations heavily relies on the precision of the keypoints detection algorithm and may struggle with capturing complex or blended actions.

1.2.3 Motion-Based Representations

Motion-based representations are essential for understanding temporal aspects of human behaviors, particularly in video sequences. These methods focus on tracking and analyzing the movement of facial features, capturing dynamic changes rather than static appearances. These approaches include Optical Flow and Motion Vectors, estimate the movement of pixels between consecutive frames in a video, generating a vector field that represents the direction and magnitude of motion.

- **Advantages:** Motion-based representations capture the dynamic aspect of human behavior, providing temporal information that is crucial for understanding actions and expressions over time. These methods are particularly effective for video-based applications and are robust to static variations in appearance.

- **Disadvantages:** Motion-based techniques, such as optical flow, can be sensitive to noise, motion blur, and occlusions. They also require temporal data (i.e., video sequences) and may struggle in environments with rapid changes in lighting or background.

1.2.4 3D Representations

3D representations capture the three-dimensional structure of the face or body to represent nonverbal human behaviors. This includes 3D meshes, point clouds, or depth maps that provide spatial information about the shape and contours of facial expressions and human actions.

- **Advantages:** 3D representations provide additional depth and contour data, leading to more accurate and detailed analysis of expressions and actions. They are more robust to changes in head pose and orientation, and are better suited for applications requiring high realism, such as virtual reality and 3D animation.

- **Disadvantages:** On the downside, 3D representations require complex hardware and algorithms for data capture and processing, which are computationally intensive and require significant storage. Capturing high-quality 3D data can be challenging, especially in uncontrolled environments.

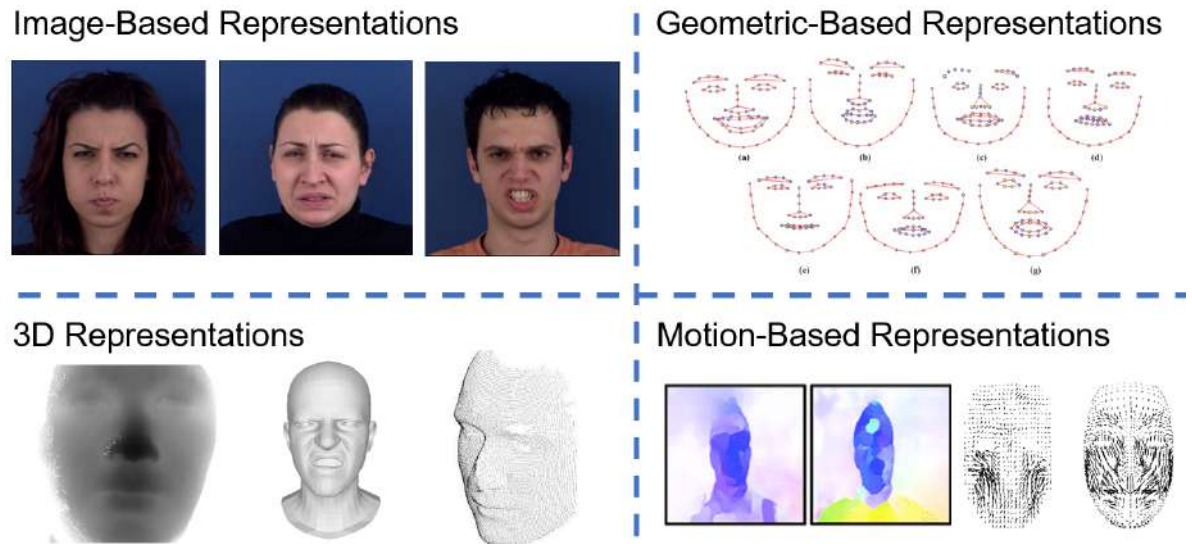


Figure 1.7: Examples of Representations of Facial Expressions and Human Actions in Computer Vision.

The highlighted data forms represent some of the most Common methods for representing facial expressions in computer vision. Each representation presents unique advantages and limitations, making the selection of the appropriate method highly dependent on the specific application and the problem at hand.

1.3 Relevant Concepts from Deep Learning

Over the past few years, computer vision researchers have concentrated on the development of representations capable of modeling facial expressions using different data forms. They have focused on handcrafted feature descriptors, which involve extracting a

set of relevant mathematical representations from the data for machine learning algorithms to correlate with the desired output. Nevertheless, hand-designed descriptors and traditional machine learning models face several limitations. They struggle in situations where human expertise is unavailable or when decisions are complex and difficult to explain, such as in language understanding, medical diagnoses, and action recognition. Machine learning also struggles in dynamic scenarios requiring frequent updates, such as price prediction and weather forecasting. Additionally, ML is insufficient for handling large datasets that surpass human engineering capabilities, such as in sentiment analysis and ad matching.

In contrast, Deep Learning (DL) presents several compelling advantages that address a range of challenges effectively. DL exhibits versatility across nearly all application domains, leveraging its capability to automatically learn optimized features without the need for manually crafted descriptors. This adaptability contributes to its robustness, enabling flexibility against variations in input data. Moreover, DL excels in generalization through techniques such as transfer learning, which allows a single model to be applied across diverse data types and applications, particularly valuable in scenarios with limited data. Additionally, DL is highly scalable, with architectures deployable at a supercomputing scale, demonstrating its capacity to handle large and complex tasks efficiently.

In this thesis, we utilize various forms of data, including images, spatial keypoints (landmarks and skeletons), and 3D mesh data. Our work is consistently focused on tasks involving temporal evolution for recognition and generation. This necessitates the use of diverse models tailored to different types of data and different goals.

- **Spatial Information in Images:** We employ Deep Convolutional Neural Networks (DCNNs) to effectively handle and interpret spatial information present in images.
- **3D Mesh Data:** To manage 3D mesh data, we use Spiral Convolutional Networks, which are specifically designed to process and analyze 3D geometric structures.
- **Temporal Data:** For handling temporal sequences and dynamics, we implement Transformer models, which are proficient at capturing long-range dependencies and temporal relationships.
- **Generation Tasks:** For generating data, we leverage Generative Adversarial Networks (GANs), Diffusion Models, and latent diffusion models all of which are designed for synthesizing new, realistic data based on learned patterns.

1.3.1 Deep Convolutional Neural Networks (DCNNs)

In the field of deep learning, Convolutional Neural Networks (CNNs) are among the most well-known and commonly used methods [21]. CNNs have revolutionized

the processing of grid-like data, such as images, videos, and audio. These models have demonstrated remarkable performance in various computer vision applications, including face recognition, animal classification, and object detection. One of the main advantages of CNNs is their ability to automatically identify relevant features without human supervision. Additionally, CNNs employ local connections and weight sharing, which reduces the number of trainable parameters, speeds up convergence, enhances generalization helping to avoid overfitting.

The hierarchical nature of CNNs allows for the extraction of features in a manner that breaks down complex concepts into simpler ones, organizes the features, and reduces dimensionality. The success of CNNs has been further boosted by the availability of powerful machines equipped with GPUs and large-scale annotated datasets, enabling the training of models with millions of parameters. Moreover, various architectural innovations, such as deeper networks (e.g., ResNet [22]), and efficient designs (e.g., MobileNet [23]), have pushed the boundaries of CNN performance in both accuracy and computational efficiency. Surveys [24, 25] offer further details about the evolution of CNNs.

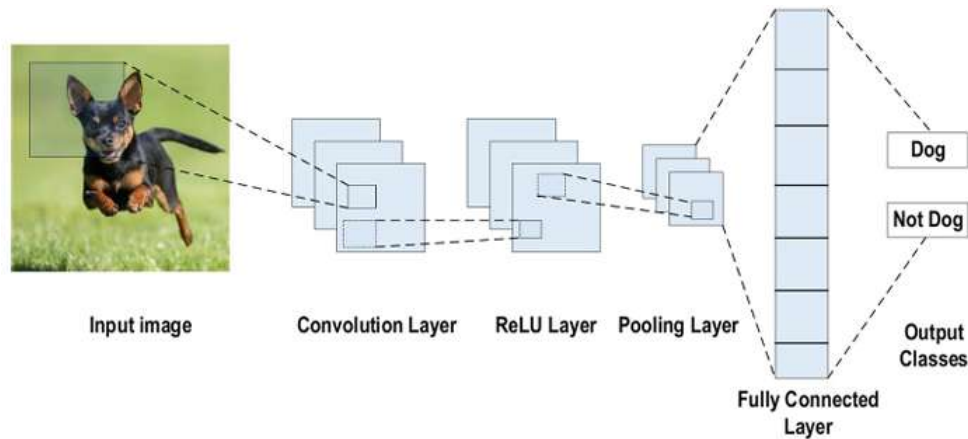


Figure 1.8: An example of CNN architecture [2].

As illustrated in Fig.1.8, DCNNs typically consist of several components:

- **Convolutional Layers:** These layers perform the core operation of Deep Convolutional Neural Networks (DCNNs) by applying convolution operations to the input data. Using various filters, convolutional layers detect different features and capture spatial hierarchies in the data, preserving the spatial relationships between pixels. Specifically, the convolution of input feature maps M with dimensions (W, H, C) , where C is the number of feature maps, with a kernel K of dimensions (w, h, C) results in an output feature map. The output is calculated as:

$$(K * M)(i, j) = \sum_{x=1}^w \sum_{y=1}^h \sum_{z=1}^C K(x, y, z) M(i+x, j+y, z)$$

The dimensions of the output feature maps are influenced by the filter dimensions,

padding, and stride parameters, which control how the filter convolves over the input data. This process allows CNNs to efficiently learn and recognize intricate patterns within the input data.

- **Activation Functions:** Since the convolution operation is linear, non-linear functions are applied element-wise after the convolution operations to introduce non-linearity into the model, enabling it to learn more complex patterns. The functions most commonly used to introduce non-linearity into DCNNs are the sigmoid, hyperbolic tangent, rectified linear unit (ReLU) activations, and Leaky ReLU. Fig. 1.9 provides a visualization of these functions.

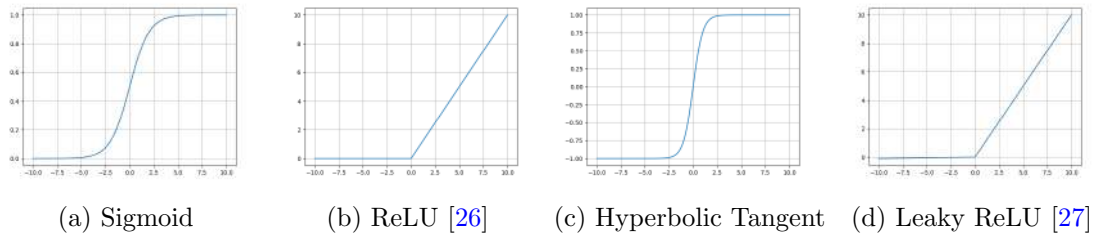


Figure 1.9: The most used activation function in DCNNs.

- **Pooling Layers:** These layers reduce the spatial dimensions of the feature maps, typically using max pooling or average pooling. Pooling helps in reducing the computational load and the number of parameters, and it also provides a degree of translation invariance.

- **Fully Connected Layers:** These layers connect every neuron in one layer to every neuron in the next layer, and they are typically used in the final stages of the network. They integrate the features extracted by convolutional and pooling layers to perform the final classification or regression task.

Despite their success, traditional CNNs face challenges with handling non-Euclidean data and large-scale variations in object appearances. This has led to the development of advanced techniques like geometric deep learning, which extends CNNs to irregular structures such as graphs and manifolds. Furthermore, combining CNNs with other techniques like adversarial training, attention mechanisms and Transformer architectures has opened new possibilities for tasks that require a more global understanding of data, such as image generation, object detection and segmentation and natural language processing [28, 29, 6].

1.3.2 Generative Adversarial Networks (GANs)

Generative models have had limited impact due to difficulties in approximating complex probabilistic computations and utilizing piece-wise linear units. However, recent advancements introduced by Ian Goodfellow et al. [28] have addressed these

challenges through Generative Adversarial Networks (GANs). GANs have demonstrated significant success across various applications, including image generation, face generation and manipulation (such as changes in hair color, gender, and aging), and image super-resolution. Fig.1.10 illustrates some realistic examples generated by GANs.



Figure 1.10: Generated faces using a GAN model. All examples are entirely synthetic and do not correspond to real individuals [3].

The framework of Generative Adversarial Networks (GANs) [28] is an innovative concept in machine learning and deep learning. It is based on a game theory scenario known as the Minimax game. In this framework, two networks are trained simultaneously through a competitive process: a generator G and a discriminator D . G learns to map input Gaussian noise to the given data distribution in order to produce samples that resemble real data, while D learns to distinguish between real data instances sampled from the given data distribution and fake instances sampled from G distribution (generated by G). An example of a GAN is presented in Fig.1.11.

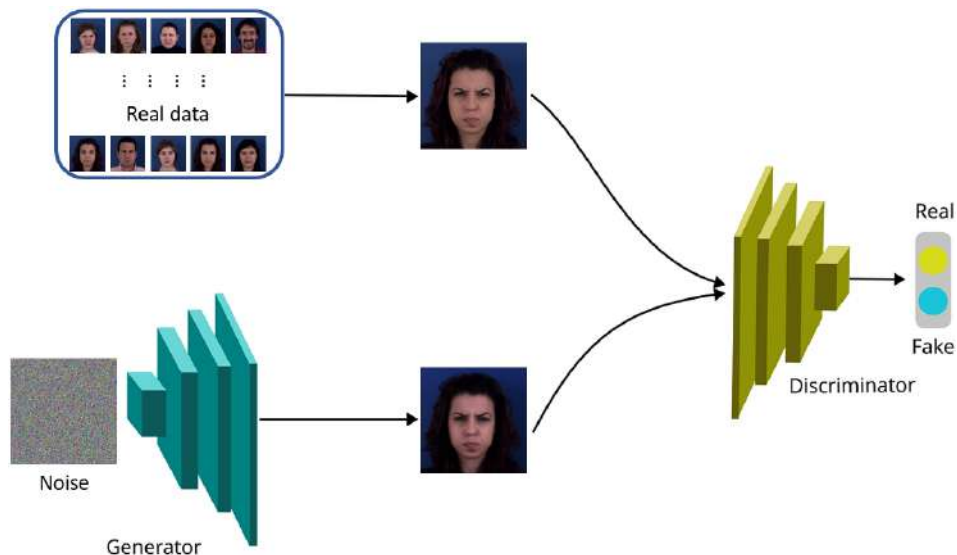


Figure 1.11: General overview of a generative adversarial network.

In easier words, the generator attempts to create convincing fake images, while the discriminator strives to continually differentiate between real and fake samples. And this competition between the two networks improves both of them.

To learn the generator distribution p_g over data x , a prior on the input noise variables $p_z(z)$ is defined. A mapping function $G(z; \theta_g)$ is used to transform this noise into data space, where G is a differentiable function represented by a multilayer perceptron with parameters θ_g . Additionally, a second multilayer perceptron $D(x; \theta_d)$ is defined to output a single scalar, representing the probability of x originating from the real data distribution p_r rather than p_g . Specifically, D is trained to minimize the quantity $\log(1 - D(G(z)))$, resulting in a two-player minimax game between D and G with the following minimax objective:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

Following the success of GANs, various architectural variants have been proposed, including Conditional GANs (cGANs) [30]. cGANs integrate additional information y to guide the generation process. In cGANs, both G and D are conditioned on this extra information, which could be class labels or data from other modalities. The objective function for cGANs becomes:

$$\min_G \max_D \mathbb{E}_{x \sim p_x} [\log D(x|y)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z|y)))]$$

1.3.3 Diffusion Models

Diffusion models, introduced in "Denoising Diffusion Probabilistic Models" (DDPMs) [31], have transformed image generation by framing it as a denoising process. Unlike methods that sample directly from a probability distribution, DDPMs iteratively refine a noisy image through a series of denoising steps, learning the underlying data distribution to generate high-quality samples. Fig.1.12 displays some faces generated using diffusion models.



Figure 1.12: Generated faces using a diffusion model. All examples are entirely synthetic and do not correspond to real individuals [4].

DDPM framework consists of two main components: the diffusion process $q(x_t|x_{t-1})$ and the reverse process $p_\theta(x_{t-1}|x_t)$, as illustrated in Fig.1.13.

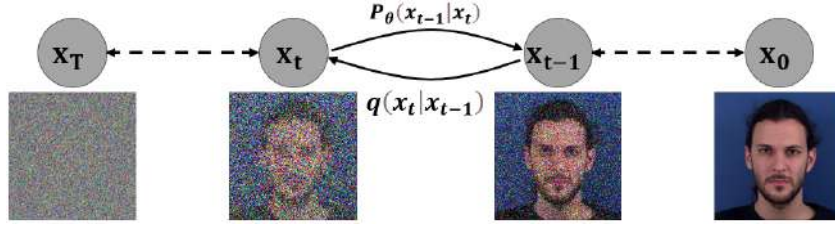


Figure 1.13: A graphical representation of diffusion models, highlighting the noise diffusion process $q(x_t|x_{t-1})$ and the denoising process $p_\theta(x_{t-1}|x_t)$.

- **Diffusion Process:** This process maps the initial data distribution $q(x_0)$ to a latent variable distribution $q(x_T) \sim \mathcal{N}(0, I)$ through a Markov chain with Gaussian noise ϵ and a predefined noise schedule $\beta_1, \dots, \beta_T$. The process is mathematically expressed as:

$$q(x_1, \dots, x_T|x_0) := \prod_{t=1}^T q(x_t|x_{t-1}),$$

$$q(x_t|x_{t-1}) := \mathcal{N}\left(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I\right),$$

where β_t is a small positive hyperparameter that controls the noise amplitude at each iteration. Defining $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$, the diffusion process enables closed-form sampling x_t at any time step t . Smaller β_t results in a larger α_t , indicating a slower diffusion process with less noise introduced at each step:

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I\right),$$

$$x_t(x_0, \epsilon) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, I).$$

- **Reverse Process:** This process maps the latent variable distribution $p_\theta(x_T) \sim \mathcal{N}(0, I)$ back to the data distribution $p_\theta(x_0)$, where θ represents the learnable parameters. This mapping is modeled as a Markov chain with Gaussian transitions, starting with $p(x_T) = \mathcal{N}(x_T; 0, I)$:

$$p_\theta(x_0, \dots, x_{T-1}|x_T) := \prod_{t=1}^T p_\theta(x_{t-1}|x_t),$$

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)^2 I\right).$$

- **Training:** The model is trained by maximizing the variational lower bound (ELBO) on the negative log-likelihood. The ELBO includes KL divergences between the diffusion process posteriors $q(x_t)$ and the reverse process mappings $p_\theta(x_t)$:

$$\mathbb{E}[-\log p_\theta(x_0)] \leq L := \mathbb{E}_q \left[\underbrace{D_{KL}(q(x_T|x_0)||p(x_T))}_{L_T} + \underbrace{\sum_{t>1} D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))}_{L_{t-1}} - \underbrace{\log p_\theta(x_0|x_1)}_{L_0} \right]. \quad (1.1)$$

Here, L_T denotes the KL divergence at the final timestep, L_{t-1} represents the KL divergence at each preceding timestep, and L_0 accounts for the log-likelihood of the initial step. By setting:

$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}x_t,$$

the posterior $q(x_{t-1}|x_t, x_0)$ is approximated as:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I\right). \quad (1.2)$$

With $\sigma_\theta^2 = \tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t$ for $t > 1$, $\tilde{\beta}_1 = \beta_1$, and a constant C , we have:

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \right] + C. \quad (1.3)$$

In practice, during training, the model minimizes a modified version of the ELBO:

$$\min_{\theta} L_{t-1}(\theta) = \mathbb{E}_{x_0, \epsilon, t} \left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|^2 \right], \quad (1.4)$$

where ϵ_θ is the parameterized noise predictor.

- **Inference:** During inference, a sample x_T is first drawn from a standard normal distribution $\mathcal{N}(0, I)$. Subsequently, samples x_{t-1} are drawn from the distribution $p_\theta(x_{t-1}|x_t)$ with the parameters:

$$\begin{aligned} \mu_\theta(x_t, t) &:= \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right), \\ \sigma_\theta(x_t, t) &:= \sqrt{\tilde{\beta}_t}, \quad t \in \{T, T-1, \dots, 1\}. \end{aligned} \quad (1.5)$$

Conditional Diffusion Models

Conditional diffusion models extend the framework by incorporating additional input y (e.g., a class label, text sequence, or guiding image) to model the conditional distribution $p_\theta(x|\mathbf{y})$. This extension enables the generation of data based on a specified conditioning signal.

During the diffusion process, noise is introduced to the input image independent of the conditioning label. However, the denoising model $p_\theta(x_{t-1}|x_t, \mathbf{y})$ takes both the image from the previous timestep x_t and the conditioning label $\mathbf{y}$ into account:

$$\begin{aligned} p_\theta(x_0, \dots, x_{T-1}|x_T, \mathbf{y}) &:= \prod_{t=1}^T p_\theta(x_{t-1}|x_t, \mathbf{y}), \\ p_\theta(x_{t-1}|x_t, \mathbf{y}) &:= \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t, \mathbf{y}), \sigma_\theta(x_t, t, \mathbf{y})^2 I\right), \end{aligned}$$

1.3.4 Latent Diffusion Models

Latent Diffusion Models (LDMs) [32] improve diffusion models by operating in a lower-dimensional latent space instead of the high-dimensional pixel space. This approach reduces computational complexity while effectively handling complex data distributions. Although LDMs follow the key steps of standard diffusion models, they are adapted for the latent space. Fig.1.14 presents examples of faces generated by latent diffusion.

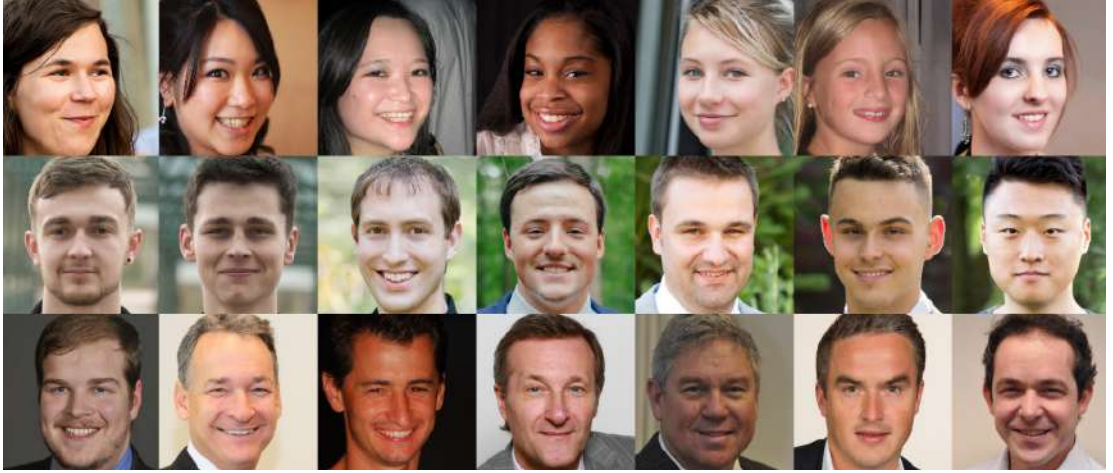


Figure 1.14: Generated faces using a latent diffusion model [5].

- **Latent Space Encoding:** The first step in LDMs involves encoding high-dimensional data (e.g., images) into a lower-dimensional latent space using an encoder. This encoder is typically a pre-trained neural network, such as an Autoencoder (AE) or Variational Autoencoder (VAE), which compresses the input data into a latent representation z . The latent space z captures the essential features of the input data while reducing its dimensionality. Let $E(x)$ denote the encoder function that maps an input image x to its latent representation z : $z = E(x)$.

- **Diffusion in Latent Space:** The diffusion process is then applied in the latent space rather than the pixel space. This involves all the diffusion steps explained in section 1.3.3, from mapping the initial latent distribution $q(z_0)$ to a latent noise distribution $q(z_T)$ through a Markov chain, to the reverse process that maps the latent noise distribution $p(z_T)$ back to the data distribution $p_\theta(z_0)$.

- **Reconstruction to Data Space:** After generating samples in the latent space, they are mapped back to the high-dimensional pixel space using a decoder. The decoder is the inverse of the encoder function. It reconstructs the data from its latent representation: $\bar{x} = D(z)$, where $D(z)$ denotes the decoder function that maps latent representations z back to the original data space.

1.3.5 Transformers

Transformers, introduced by Vaswani et al. [33], represent a major advancement in the field of sequence processing models. Traditional sequential models, like recurrent neural networks (RNNs) [34], long short-term memory (LTSMs) [35] and gated recurrent neural (GRUs) [36], struggle with capturing long-range dependencies due to their inherently sequential processing and limited receptive fields. In contrast, Transformers excel by using attention mechanisms that allow for direct connections between any two positions in the input, enabling effective modeling of long-range relationships. The Transformer model is built entirely on self-attention mechanisms, without using recurrence and convolutions altogether. This innovative architecture allows for significant parallelization, as the self-attention mechanism processes all positions in the input sequence simultaneously. This not only reduces training time but also improves scalability, making Transformers highly effective on large datasets.

Given an input sequence $X = (x_1, \dots, x_m)$, the Transformer architecture comprises an encoder and a decoder, each consisting of a stack of identical layers. The encoder maps the input sequence X to sequence of representations $Z = (z_1, \dots, z_m)$. The decoder then uses Z to generate an output sequence $Y = (y_1, \dots, y_m)$ one element at a time. At each step, the model operates in an auto-regressive manner, using the previously generated symbols as additional input for producing the next symbol. Fig.1.15 illustrates both the encoder (left) and decoder (right) of the transformer.

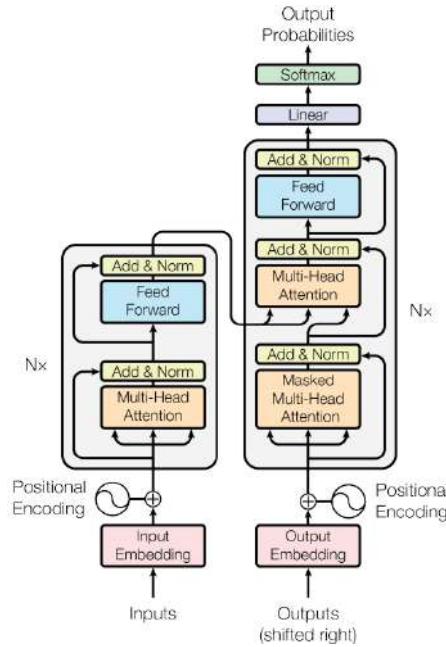


Figure 1.15: Transformer model overview [6].

- **Encoder:** The encoder consists of a stack of N identical layers, each containing two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Each sub-layer employs a residual connection followed by layer

normalization, resulting in outputs of the form $\text{LayerNorm}(x + \text{Sublayer}(x))$, where $\text{Sublayer}(x)$ is the function implemented by the sub-layer.

- **Decoder:** The decoder also consists of a stack of N identical layers. In addition to the two sub-layers in each encoder layer, the decoder includes a third sub-layer that performs multi-head attention over the encoder stack's output. Similar to the encoder, each sub-layer in the decoder uses residual connections followed by layer normalization.

Additionally, the self-attention sub-layer in the decoder is modified to prevent positions from attending to subsequent positions. This masking, combined with the offsetting of output embeddings by one position, ensures that the prediction for position i can depend only on the known outputs at positions less than i .

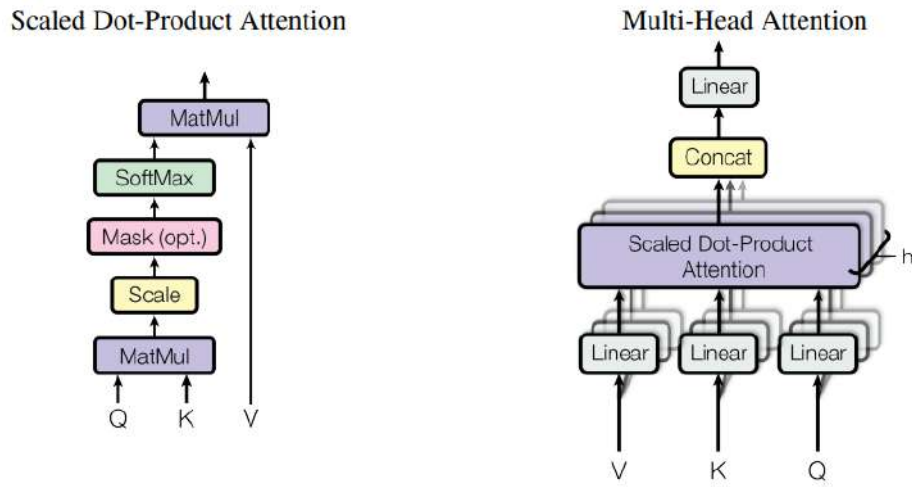


Figure 1.16: self-attention mechanism in Transformer.

Attention Mechanisms

An attention function maps a query and a set of key-value pairs to an output. The output is calculated as a weighted sum of the values, where each weight is determined by a compatibility function between the query and its corresponding key. For each token from the sequence, three vectors are created: Query (Q), Key (K), and Value (V) by multiplying the input token embedding by three learned matrices.

- **Scaled Dot-Product Attention:** (Fig.1.16, left side). The input comprises queries and keys of dimension d_k and values of dimension d_v . The dot products of the query with all keys are calculated, divided by $\sqrt{d_k}$, and then passed through a softmax function to generate the weights for the values. After finding the attention weights

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- **Multi-Head Attention:** (Fig.1.16, right side). Instead of using a single attention function with d_{model} -dimensional keys, values, and queries, it is advantageous to linearly

project the queries, keys, and values h times using different, learned linear projections to d_k , d_k , and d_v dimensions, respectively. Each of these projected versions of queries, keys, and values then undergoes the attention function in parallel, producing d_v -dimensional output values. These outputs are concatenated and projected again to obtain the final values.

Multi-head attention allows the model to simultaneously attend to information from different representation subspaces at various positions.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

where

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Here, $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ are the projection matrices.

- **Self-attention and Multihead-self attention:** is a variant of the general attention mechanism, specifically designed for capturing relationships within a single sequence. The key difference between general attention and self-attention lies in the scope of the relationships they capture. General attention mechanisms typically operate between different sequences, allowing the model to focus on relevant parts of an external sequence (such as in sequence-to-sequence tasks like machine translation). In contrast, self-attention operates within a single sequence, enabling the model to capture intra-sequence relationships and dependencies, which is crucial for understanding context within the sequence itself. In self-attention, each position in the sequence functions as a query, and the same sequence provides the keys and values. By allowing each position to attend to every other position, self-attention creates context-aware representations, enabling the model to understand dependencies and interactions within the sequence without relying on external information.

Multi-Head Self-Attention builds on self-attention by incorporating multiple attention heads. Each head performs self-attention independently with different learned projections of the input data. This approach allows the model to capture diverse aspects of the sequence context simultaneously. Each attention head focuses on different parts or types of relationships within the sequence, which collectively provides a more comprehensive representation.

1.3.6 Mesh Representation and Spiral Convolutions

While 2D images are widely used and have proven to be effective in numerous applications, they lack depth information and suffer from occlusion and perspective issues. These limitations can obstruct tasks that require spatial awareness, interactivity, and comprehensive analysis. To address these challenges, 3D data provides a more robust solution. By incorporating depth and spatial information, 3D data overcome the

constraints of 2D images, enabling more accurate modeling, and enhanced interactivity.

3D data comes in multiple forms, including point clouds, volumetric data, and mesh data. Fig.1.7 shows examples of 3D data forms. Point clouds represent objects as a collection of discrete points in 3D space, while volumetric data encapsulates 3D objects within a grid of voxels. Mesh data, which consists of vertices, edges, and faces to define the shape of a 3D object, offers distinct advantages. In Our work, we focus on mesh data due to its effectiveness in representing complex geometries. Meshes provide a continuous surface representation, which is particularly advantageous for tasks requiring detailed surface information, such as surface deformation analysis. Moreover, mesh data facilitates easier integration with computer graphics pipelines and is well-suited for real-time rendering applications. These benefits make mesh data a preferred choice for many applications in 3D modeling, animation, and computer vision. Fig.1.17 presents examples of mesh data applications.

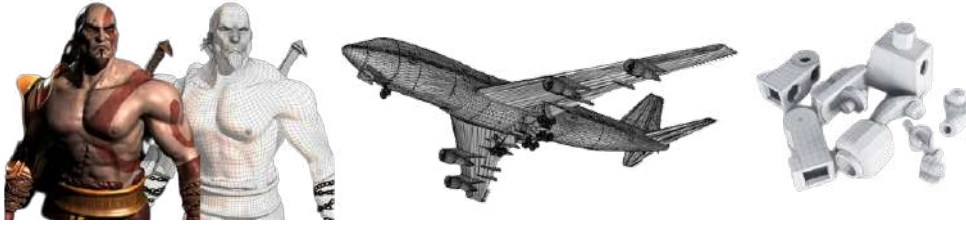


Figure 1.17: Applications of Meshes in 3D Modeling and Computer Graphics. From left to right: a detailed 3D character model from a video game, an airplane model , and a collection of mechanical parts.

Mesh representation

This representation is widely preferred over other surface models such as implicit surfaces or parametric surfaces. According to Botsch et al. [37], different types of 3D meshes can be distinguished based on the geometric shape of the faces, typically a triangle or a quadrilateral, rarely polygons with more vertices. In our work we have used triangular polygonal mesh. As illustrated in Fig.1.18, a triangular polygonal mesh M is defined by its geometry (vertices) and its connectivity, which describes the relationships between the vertices of the mesh. The components of a mesh M are:

- A set of vertices $S = \{s_1, \dots, s_n\}$
- A set of triangular faces $F = \{f_1, \dots, f_m\}$, where $f_m \in S^3$
- A set of edges $A = \{a_1, \dots, a_l\}$, where $a_i \in S^2$.

Mesh data finds use in a variety of applications, including computer graphics, medical imaging, and virtual reality. When applied to temporal data, such as sequences of meshes, they present significant challenges due to their changing topology. This means that each mesh may differ in the number of vertices and their arrangement. To address this, researchers often convert the data into point clouds or voxel grids, or

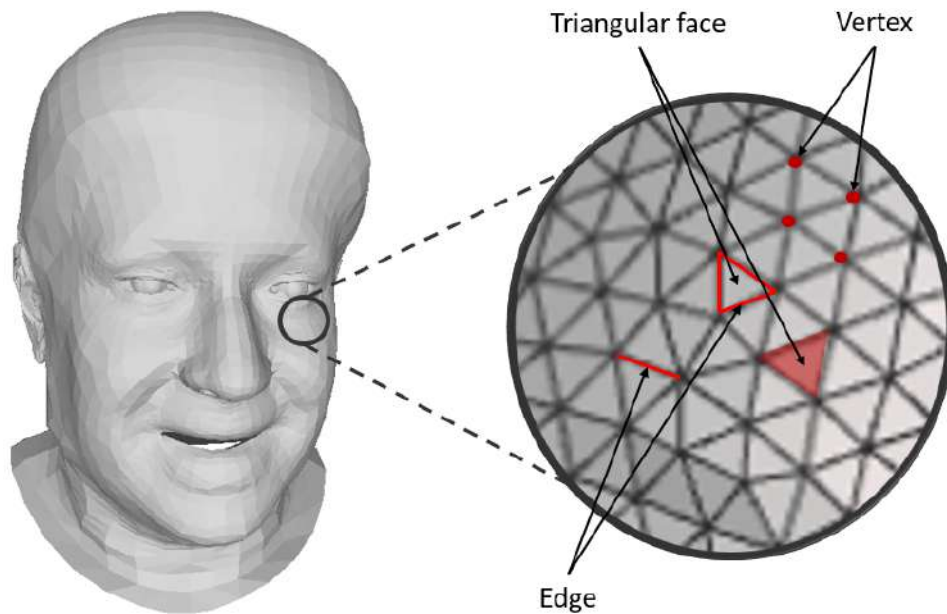


Figure 1.18: Elements of a triangular 3D mesh.

resort to complex techniques to manage the variability. This increases computational demands and complexity. However, certain applications require simple fixed topology mesh sequences. For instance, character animation, facial expression modeling, and biomechanical simulations in the metaverse benefit greatly from fixed topology meshes. These fixed-topology sequences facilitate consistent and efficient processing, leading to smoother and more realistic simulations. Despite these advantages, research has traditionally applied the same complex methods to fixed-topology meshes as to those with variable topology. Only recently have methods been developed that specifically address the unique needs of fixed-topology meshes, emphasizing the importance of specialized and efficient approaches in this area.

Fix topology Mesh

A fixed topology mesh is a type of 3D mesh where the connectivity and structure of the vertices, edges, and faces remain constant throughout the sequence or application. Specifically, 1) the number of vertices, 2) their connectivity, and 3) the order of these connections do not change over time. This consistency allows for simpler and more efficient processing, as the topology of the mesh does not need to be recalculated or adjusted for each frame or iteration.

Creating fixed topology meshes can be achieved through several methods and techniques:

- **Manual Design and Modeling:** Using software like Blender, Maya, and others to manually design meshes with a fixed topology or to adhere to standard templates.
- **Mesh Registration:** Aligning and deforming different meshes to a common reference template to ensure consistent topology across all meshes. Techniques such as SMPL

(Skinned Multi-Person Linear model) [38] and FLAME (Faces Learned with an Articulated Model and Expressions) [39] are notable examples that maintain fixed topology during alignment and deformation.

- **Mesh Morphing and Interpolation:** Gradually transforming one mesh into another while keeping the topology fixed and creating intermediate meshes with consistent topology between two given meshes.

- **Generative models:** Utilizing generative models that produce fixed topology meshes, thereby creating sequences with consistent topology following a predefined structure.

Spiral Convolutions for fixed topology meshes

Given a manifold discretized as a triangular mesh $\mathcal{M} = (\mathcal{V}, \mathcal{E}, \mathcal{F})$, where $\mathcal{V} = \{v_1, \dots, v_n\}$ represents the set of vertices, and $\mathcal{E}$ and $\mathcal{F}$ denote the sets of edges and faces respectively, let $f : \mathcal{V} \rightarrow \mathbb{R}$ be a function representing the vertex features.

A significant challenge in designing convolution-like operators for graphs or manifolds is the absence of a global coordinate system for each point. Early intrinsic mesh convolutional architectures, such as GCNN [40], ACNN [41], and MoNet [42], address this challenge by constructing a local coordinate system $u(x, y)$ around each vertex x of the mesh, without the requirement to any global coordinates system. This local system allows for the application of a set of local weighting functions $w_1, \dots, w_L$ to aggregate information from the neighboring vertices y in $N(x)$. This approach generalizes the concept of “patch operators”, which are analogous to sliding window filters in images:

$$(f \star g)_x = \sum_{\ell=1}^L g_{\ell} \sum_{y \in N(x)} w_{\ell}(u(x, y)) f(y)$$

Here, $\sum_{y \in N(x)} w_{\ell}(u(x, y)) f(y)$ represents the aggregation of “soft pixels” (L total), where f is similar to pixel intensity in images and g_{ℓ} denotes the filter weights. However, the lack of a global coordinate system equates to the absence of a canonical ordering of vertices. Therefore, patch-operator-based approaches can be viewed as attention mechanisms. Instead of a direct “1-1” mapping between neighboring features $f(y)$ and filter weights g_{ℓ} , an “all-to-all” mapping is achieved through learnable soft-attention weights $w_{\ell}(u(x, y))$. Moreover, the patch-operator-based methods are motivated by the need for insensitivity to variations in the meshing of the continuous surface. Ideally, each patch operator should remain effective regardless of the underlying graph topology.

These methods, however, come with significant computational complexity, a high number of parameters, and can be challenging to optimize. Additionally, they often require hand-crafting and pre-computing the local coordinate systems. To address these drawbacks, Bouritsas et al. [7] proposed a solution based on the observation that the challenges of lacking a global ordering and insensitivity to graph topology are less relevant

when working with fixed topology meshes. By locally ordering the vertices and maintaining that order, the graph convolution can be defined as follows:

$$(f \star g)_x = \sum_{\ell=1}^L g_{\ell} f(x_{\ell})$$

where $\{x_1, \dots, x_L\}$ denote the neighbors of vertex x ordered in a consistent way. Analogous to patch operators in image processing, each neighboring vertex serves as a patch operator. In the Euclidean setting, these vertices are ordered using a raster scan of pixels within a patch. For meshes, the authors proposed an intuitive and simple ordering method utilizing spiral trajectories.

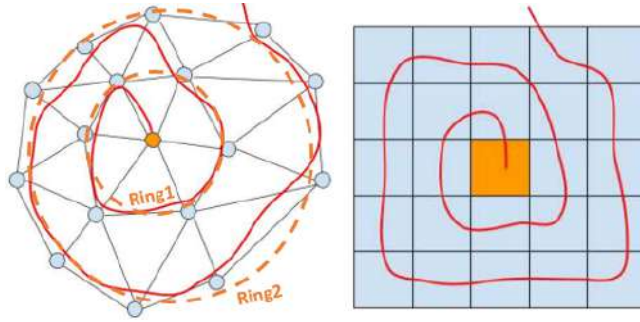


Figure 1.19: Spiral ordering on a mesh and an image patch [7].

Let $x \in V$ represent a vertex on a mesh, and let $R_d(x)$ denote the d -ring, an ordered set of vertices exactly d hops away from x along the shortest path. The notation $R_{d,j}(x)$ represents the j -th element in this d -ring. The spiral patch operator is defined as the ordered sequence:

$$S(x) = \{x, R_{1,1}(x), R_{1,2}(x), \dots, R_{h,|R_h|}(x)\}$$

where h denotes the patch radius, similar to the kernel size in classical CNNs. The spiral convolution is then expressed as:

$$(f * g)_x = \sum_{\ell=1}^L g_{\ell} f(S_{\ell}(x))$$

The uniqueness of the ordering is achieved by fixing two degrees of freedom: the direction of the rings and the initial vertex $R_{1,1}(x)$. The remaining vertices in the spiral are ordered inductively. The direction can be either clockwise or counterclockwise, while the first vertex is selected based on the underlying geometry of the shape, ensuring the method's robustness. A reference vertex x_0 is fixed on a template shape, and the initial point for each spiral is chosen in the direction of the shortest geodesic path to x_0 , formally defined as:

$$R_{1,1}(x) = \arg \min_{y \in R_1(x)} d_M(x_0, y)$$

where d_M represents the geodesic distance between two vertices on the mesh M . A fixed length L is chosen as a hyper-parameter, and each spiral is either truncated or zero-padded to fit this length. This ordering approach is applied for both local and global ordering in the mesh.

1.4 Conclusion

In this chapter, we have introduced the general background needed to understand the content covered in the following chapters. This includes key terminology related to facial expressions and the fundamental concepts of deep learning that we utilize throughout our thesis.

Chapter 2

Literature Review

2.1 Introduction

This thesis explores and contributes novel approaches for analyzing and synthesizing facial expressions through advanced deep learning techniques. To frame our contributions within the current state-of-the-art, this section provides an in-depth review of the relevant deep learning models, and datasets. It emphasizes their specific applications and effectiveness in both facial expression analysis and synthesis. This examination aims to identify key trends, challenges, and opportunities in the field, positioning our work within the context of state-of-the-art research.

2.2 Facial Expression Recognition (FER)

In recent years, the field of facial expression recognition (FER) has attracted considerable attention within the computer vision community due to its wide-ranging applicability. This interest has driven the development of numerous methods to address the inherent complexities of FER.

Earlier FER methods relied heavily on hand-crafted feature models, which extract facial expression features as either global or local attributes. A primary challenge in this approach is obtaining representative features that exhibit low within-class variance and high between-class variance. Commonly used hand-crafted features include geometric descriptors, such as distances between facial landmarks [43], and appearance descriptors like Local Binary Patterns (LBP) [44, 45, 46], Scale-Invariant Feature Transform (SIFT) [47, 48], and Histogram of Oriented Gradients (HOG) [49, 50]. These descriptors are utilized to create robust facial representations for effective analysis and recognition of facial expressions. However, hand-crafted features struggle to capture the full complexity of human facial expressions, leading to performance limitations in real-world scenarios.

More recently, the focus has shifted towards developing deep learning algorithms to overcome these limitations. Deep learning architectures, particularly those capable of modeling both spatial and temporal information, have shown significant promise in improving FER performance. One line of research involves leveraging 3D spatio-temporal convolutions to simultaneously capture spatial and temporal dimensions in video data

[51, 52, 53]. This approach enables a more comprehensive analysis of facial expressions. However, 3D convolutions are computationally demanding, requiring substantial memory and processing power. This poses scalability challenges, especially with long video sequences or high-resolution videos, where computational and memory demands escalate exponentially.

Another research direction explores the combination of deep architectures to model both appearance and motion information. For instance, various studies have successfully integrated deep convolutional neural networks (DCNN) with recurrent models, resulting in hybrid architectures such as CNN-RNN [54], CNN-BRNN [55], and RNN [56], as well as DCNN combined with transformers in [57, 58, 59], and DCNN with Graph networks and biLSTMs in [60]. These hybrid models have demonstrated efficacy in facial expression recognition by leveraging the strengths of both convolutional and recurrent layers to capture dynamic facial changes over time.

In addition to these methods, there has been significant interest in exploiting multimodal deep learning in FER. *Salama et al.* [61] and *Tan et al.* [62] utilized multimodal data by combining image data with features extracted from EEG signals. *Zheng et al.* [63] integrated text and audio inputs alongside image data, and *Zhang et al.* [64] combined facial landmarks with image features. These multimodal approaches offer robustness by incorporating multiple data sources but also introduce challenges, such as increased computational demands and data fusion complexities.

Several approaches have modeled the temporal evolution of facial landmarks as trajectories. *Taheri et al.* [65] employed landmark configurations on the Grassmann manifold, capturing facial deformations with geodesic velocities and classifying them using LDA and SVM. *Kacem et al.* [66] proposed mapping facial landmarks onto the Riemannian manifold of positive semi-definite matrices, using parameterized trajectories and aligning them with a specific SVM classifier, pairwise proximity function SVM (ppfSVM). *Otberdout et al.* [67] used deep covariance descriptors, which encode DCNN features into compact local and global covariance matrices on the symmetric positive definite manifold. Despite the strong performance of 2D FER methods, challenges such as pose variation, lighting sensitivity, and occlusions limit their effectiveness in real-world scenarios.

3D Facial Expression Recognition

To address these limitations, 3D FER has emerged as a more robust alternative by incorporating depth information, which improves robustness to pose variations, lighting, and occlusions, while better capturing the dynamics of facial expressions.

Existing 3D FER approaches can be categorized into two main groups: hand-designed feature descriptors and projection-based methods. The first group achieves FER through encoding the spatial and/or temporal information of the 3D video using hand-designed

feature descriptors, then using the extracted codes to train machine learning classifiers, such as HMMs, SVMs, or Random Forests, to label the expressions. The most popular examples of this group are based on descriptors that compute: distances between extracted keypoints of the face [68, 69, 70], distances between segments of level curves [71], LBP-TOP from consecutive frames [72], Free-Form Deformations (FFD) [73, 74], Deformation Vector Field (DVF) [75], Dense Scalar Field (DSF) [76, 77], Nebula features [78], three-dimensional discrete cosine transform (3D-DCT) [79], histograms of 3D gradient (HOG3D) [80], differential quantities, including coordinates, normals, and shape index values of each vertex [81]. Although hand-designed feature descriptor methods have demonstrated promising results on small datasets, they face several limitations. These methods often require extensive feature selection to select the most appropriate features for the task at hand. Scalability also poses a challenge for feature descriptor-based methods when confronted with variations in data size, distribution, or classes, often demanding human intervention to adapt the feature descriptors. Additionally, certain methods within this category only show high classification accuracy for the classification of three basic expressions and subject-dependent scenarios. Furthermore, their applicability is confined to brief sequences comprising only 6 to 12 frames.

To address the limitations of feature descriptor methods, projection-based approaches have emerged, utilizing 2D deep learning techniques to tackle 3D FER. *Abbasnejad et al.* [82] proposed projecting the 3D face scenes onto 2D to create videos of the frontal facial expressions, which are subsequently used to train a CNN model for FER. Unlike [82], which focuses on appearance-based features without any consideration to geometric information, *Li et al.* [83] suggested constructing geometrical images, such as Depth Images, three Normal Component Images, and Shape Index Images, from the 3D data. Then, a dynamic geometrical image network (DGIN) is employed for end-to-end facial expression classification, effectively fusing geometrical and temporal information. *Taha et al.* [84] suggested mapping geometric information from compressed 3D meshes into 2D texture images. These 2D texture images, containing a fusion of texture and geometric information, are then utilized by a CNN model for feature extraction and FER. In [85], *Behzad et al.* introduced a collaborative cross-domain dynamic image network that combines the concepts of multi-viewing and geometrical imaging, by fusing texture and geometrical information from 3 views (frontal, left, and right profiles), and feeding them to CNN model for classification. A refined version of this work includes incorporating features extracted from automatically detected landmarks [86]. The same authors [87] proposed a novel data augmentation technique and a sparsity-aware deep network for 4D FER. They start by projecting the input data into RGB and depth map images, followed by randomized channel concatenation to create new images. A sparsity-aware deep network is then used for deep feature extraction, and three orthogonal plans TOP-landmarks are computed to capture the encoded facial deformations stored in the 3D landmarks. Both features are then collaboratively used to train a BI-LSTM network for FER. While projection-based methods have shown impressive performance,

they are not applied directly to mesh data, and the process of projecting 3D data into the 2D domain can be time-consuming, potentially leading to a loss of geometric information.

2.3 Facial Expression Generation (FEG)

Static Facial Expression Generation

Facial expressions synthesis was initially achieved through the use of traditional methods, such as geometric interpolation [88, 89], parameterization [90, 91, 92], morphing [93, 94, 95], etc. These methods show success on avatars, but they fall short when dealing with real faces, and they are unable to generalize a flow of movement for all human faces due to the high complexity of natural human expressions and the variety of identity-specific characteristics. To face these limitations, neural networks-based methods have been applied on facial expressions generation, including RBMs [96], DBNs [97] and Variational Auto-Encoders [98]. These approaches offer improved data representation and flow between different data distributions compared to traditional methods. However, they still face challenges such as imprecision in controlling facial expressions, low resolution, and blurry generated images.

With the appearance of GANs, many of its extensions have been dedicated to facial expression generation. [99, 100, 101] exploit the concept of adversity with auto-encoders to present Adversarial Auto-Encoders. *Zhu et al.* [102] proposed CycleGAN, that employs Cycle-Consistency Loss to preserve the key attributes (identity) of the data. *Choi et al.* [103] introduced StarGAN, which addresses the inefficiency of creating separate generators for each transformation by using a single architecture capable of handling multiple transformations across different datasets. *Wang et al.* [104] suggested exploiting the U-Net architecture as GAN generator to increase the quality and resolution of generated images. *Arbush et al.* [105] proposed the US-GAN model that employs an "Ultimate Skip Connection" to link the input image with the output of the model, allowing the model to focus solely on learning expression-related details. While US-GAN improves identity preservation, artifacts still appear in areas related to expressions such as the mouth, nose, and eyes. The studies above established the task of generating facial expression classes (e.g., sad, happy, angry), but the intensity of the expression often limits understanding of the person's emotional state. *Ding et al.* [106] addressed this issue by introducing ExprGAN, which incorporates an expression controller module to adjust expression intensity continuously from weak to strong. Similarly, *Pumarola et al.* [107] and *Wu et al.* [108] used Action Units (AUs) to condition the generation process, offering greater diversity in the expressions produced. Other methods such as G2-GAN [109] and GC-GAN [110] exploited facial geometry as a condition to control facial expression synthesis. The objective of these models is to take as input a facial image and facial landmarks in the form of binary images or landmark coordinates, then learn to generate a

realistic face image with the same identity and the target expression. [111] and [112] utilize labels from the 2D Valence-Arousal Space, in which the valence represents how positive or negative an emotion is, while arousal represents the strength of emotional activation [20], to guide the process of facial expression generation, enhancing the variety and the control of the generated expressions. However, while these models have made progress in generating static facial expressions, they fail to account for the dynamic nature of real-life expressions.

Dynamic Facial Expression Generation

Facial expressions are naturally dynamic actions that contain more information and details than a single pose, e.g. the speed of facial expression transformation, head movements when displaying the expression, etc. This information can be significant in understanding the emotional state of a person.

To achieve facial expression transfer, methods such as [113, 114, 115, 116, 117, 118] utilize embeddings [119], action units [120], or facial landmarks [121, 109, 122] extracted from target video frames. These techniques focus on transferring facial expressions from a driver to a target face while maintaining the target’s distinct features, such as the shape of the eyes, nose, and overall facial structure, even in cases where significant differences exist between the source and target faces. Facial expression transfer offers several advantages, notably in preserving identity and enhancing realism. By leveraging actual expressions from real faces, expression transfer often produces more realistic and consistent results, leading to natural-looking facial movements compared to generative models that create expressions from scratch, potentially losing identity features. Furthermore, training models for expression transfer can be more efficient, concentrating on mapping expressions rather than generating new ones, resulting in faster training times and lower computational costs. However, facial expression transfer methods are limited by their reliance on pre-defined motion dynamics, as they do not learn these dynamics independently. Additionally, the quality and diversity of the driver expressions significantly impact the effectiveness of the transfer, and substantial differences in facial characteristics between the driver and target can cause artifacts or distortions. Generalization to unseen expressions remains challenging, often requiring a trade-off between preserving identity and accurately adopting the driver’s expression. Despite these challenges, recent advancements in methods utilizing generative adversarial networks (GANs) and variational models have made significant progress in balancing these trade-offs and enhancing the quality and applicability of facial expression generation.

A group of methods, such as VGAN [123], CVGAN [124], TGAN [125], G3AN [126], and ImaGINator [127], adopt a single step approach for generating the entire facial expression video by utilizing fractionally strided spatio-temporal convolutions. These convolutions simultaneously generate both appearance and motion, streamlining the process. For instance, *Vondrick et al.* [123] introduced VideoGAN (VGAN), which decomposes video

generation into two parts: a static background and a dynamic foreground. This method utilizes a dual-stream generator that processes background and foreground separately before combining them to form the final video output. In the same direction, *Saito et al.* [125] introduced Temporal GAN (TGAN), which generates video frames by learning from unlabeled videos. It uses a temporal generator to capture video dynamics and an image generator to transform latent variables into frames, producing a video sequence. *Wang et al.* [126] presented G3AN, which takes a different approach by modeling appearance and motion in a disentangled manner. It achieves this through a three-stream generator, where the main stream ensures spatio-temporal consistency, while two auxiliary streams provide multi-scale appearance and motion features. This separation aims to enhance the quality of the generated videos by addressing different aspects of video generation independently. Despite their advantages, these methods are unconditional and initiate the generation process from Gaussian noise, which simplifies the architecture but also introduces certain limitations. Specifically, they often struggle with preserving facial identity and offer limited control over the generated expressions. To address these challenges, *Wang et al.* [127] proposed ImaGINator, which integrates an auto-encoder architecture with spatio-temporal fusion, where low-level spatial features in the encoder are directly passed to the decoder, similar to the U-Net architecture. Moreover, ImaGINator employs two discriminators: one processes the entire video while the other processes frame-by-frame. This dual-discriminator setup aims to improve the spatio-temporal consistency of the generated videos. However, despite improving spatio-temporal coherence, these single-step approaches still face significant challenges, often resulting in lower video quality, higher noise, and reduced identity preservation compared to two-step methods. Additionally, the complexity of these models leads to high computational and resource demands, which restricts the resolution and quality of the generated images. The significant computational power and memory requirements can make these methods impractical for real-time applications or generating high-resolution videos. Furthermore, while models like VGAN and G3AN attempt to enhance video quality and consistency, they still struggle with achieving a high level of realism and maintaining the unique characteristics of the individual's facial identity. The challenge of balancing between generating realistic motion and preserving individual identity remains a significant hurdle.

To tackle these limitations, other works followed a different direction and investigated decoupling motion and content generation. In these works [128, 129, 130, 131, 132, 67, 133, 134] the motion is generated separately as codes in the latent space, these codes are then fed into the generator to generate frames of the video individually following the seq2seq [135] architecture. For example, *Tulyakov et al.* [128] introduced MoCoGAN, which decomposes the video into content and motion information, where the video motion is learned by Gated RNN (GRU) and the video frames are generated sequentially by a GAN. *Gupta et al.* [129] proposed RV-GAN, that uses transpose convolutional LSTMs (upsampling instead of downsampling) as the GAN generator to generate frames individually. However, both models exhibit content and motion artifacts and are limited to previously seen identities

and a finite range of expressions. In [131], *Fan et al.* extended the MoCoGAN architecture by incorporating an encoder to preserve identity and adding a control parameter for expression intensity. *Wang et al.* [130] utilized a Multi-Mode Recurrent Landmark Generator to learn to generate variant sequences of facial landmarks of the same category (e.g. different ways to smile), translated later to video frames. In [136], *Otherdout et al.* exploit a conditional version of manifold-valued Wasserstein GAN to model the facial landmark motion as curves encoded as points on a hypersphere. The W-GAN learns the distribution of facial expression dynamics of different classes, from which new facial expression motions are synthesized and transformed to videos by TextureGAN. Decoupling motion and content generation in facial expression synthesis offers notable advantages, including enhanced control over dynamic aspects and improved identity preservation. By separating motion from content, these methods enable fine-tuned manipulation of expressions and smoother transitions between different emotional states. This approach also facilitates the generation of diverse and varied expressions, contributing to more natural and expressive animations. Despite these advantages, significant limitations persist. These models frequently encounter spatio-temporal consistency issues, resulting in content and motion artifacts that can detract from the realism of the generated videos. They are generally constrained to identities encountered during training, which limits their ability to generalize to new or unseen faces.

Among the methods discussed, most facial expression generation approaches remain limited in scope, typically focusing on the face while excluding other important regions such as the hair, neck, and background. This narrow focus often results in lower resolution outputs, commonly around 64x64 pixels, which may compromise the overall quality and realism of the generated videos. Many researchers have attempted to overcome this limitation by adapting successful image generation models, such as StyleGAN [3], for facial expression tasks [137, 138]. These approaches involve leveraging the embedding space of pre-trained StyleGAN models and manipulating this space to produce video frames. While promising for generating high-quality expressions, these methods encounter significant challenges in maintaining content consistency and achieving realistic results when manipulating images beyond the pre-trained domain or in regions that are less covered by the generator. Additionally, these methods often struggle with integrating facial expressions smoothly with the surrounding context, resulting in potential inconsistencies in the final output.

To address the issues related to resolution and quality, diffusion-based [31] models have emerged as a promising alternative. These models aim to improve the detail and fidelity of the generated expressions by progressively refining images through a series of steps, thus enhancing the overall realism and coherence of the generated facial expressions. Some methods utilized diffusion to spatially enhance the resolution and quality of the generated frames [139, 140]. However, the iterative nature of diffusion models leads to significantly increased computational complexity and inference time. Video Diffusion Models (VDM) [141] propose extending the diffusion model from images to video by

using a denoiser composed of 3D spatio-temporal convolutions. Although promising, this approach suffers from high computational demands, limiting both the resolution and length of the generated videos (e.g., 32x64x64 or 16x128x128). Alternatively, other approaches [141, 142, 8, 143, 144] employ diffusion models to learn and manipulate the embedding space. *He et al.* [142] introduced a Latent Video Diffusion Model (LVDM), which employs a Variational Auto-Encoder (VAE) to compress RGB pixel data into a latent space, where diffusion processes are applied. In a similar direction, *Ni et al.* [8] developed a Latent Flow Diffusion Model (LFDM), which utilizes diffusion to synthesize optical flow sequences. These sequences are then used to warp the latent representation of the input image to align its content with the desired motion specified by the generated flow. LFDM has shown improved effectiveness in video synthesis and introduced a wide range of generated motions through stochasticity in the latent generation process. Diffusion models offer significant advantages in facial expression generation by enhancing resolution and quality through iterative refinement, resulting in highly detailed and realistic outputs. They are particularly effective at improving spatial details and producing smooth, coherent transitions in facial expressions over time. Additionally, the stochastic nature of diffusion models contributes to a richer variety of generated motions. However, these strengths come with significant trade-offs. The iterative refinement process of diffusion models significantly increases computational complexity and inference time, posing challenges for real-time applications. Furthermore, the stochasticity introduced can impact the consistency of deterministic features, such as identity and background details, potentially leading to variability and partial loss of input information. Thus, while diffusion models excel in generating high-resolution and dynamic facial expressions, their computational demands and variability must be carefully managed.

3D Facial Expression Generation

While 2D facial expression generation has achieved impressive results, its lack of depth and accuracy limits its use in immersive applications like the Metaverse, where realistic 3D expressions are crucial for lifelike avatars and enhanced social presence. The demand for 3D facial expression generation, especially using mesh data, is growing in virtual reality (VR), augmented reality (AR), and avatar-based communication, due to its ability to capture more natural interactions.

Despite its potential, 3D mesh-based facial expression generation (FEG) presents several challenges primarily related to the unstructured nature of mesh data. This lack of structure complicates critical processes such as vertex generation, mesh connections, and loss function computation. To handle this, precise and robust registration techniques are required to align or map 3D face scans from different subjects into a consistent template, ensuring consistency across various expressions and individuals. Without accurate registration, generated expressions may appear misaligned or unnatural.

The first **3D Morphable Face Model (3DMM)**, introduced by *Blanz* and *Vetter* in

1999 [145], represents 3D faces as a linear combination of basis shapes and textures derived from a set of 3D face scans. Each face is modeled using shape and texture coefficients, which are adjusted to create new, realistic 3D faces. By manually manipulating these coefficients, the model can generate various facial shapes and expressions, allowing for facial synthesis, reconstruction, and manipulation. Initially computed from 200 subjects, 3DMM has demonstrated its utility in a variety of applications, offering plausible shape generation with a simple underlying model. Subsequent studies [146, 147, 148, 149, 150, 151] have built on the previous work to improve the quality and accuracy of the generated meshes, by increasing the number of examples used to reconstruct the model, improving the template topology, and increasing the complexity of the used coefficients to achieve greater expressivity. However, the linearity of 3DMM limits its modeling capabilities, encouraging a shift towards non-linear neural encoder-decoder architectures. This comes at the cost of reformulating convolution and pooling/unpooling operations on the non-regular mesh structures [152]. *Ranjan et al.* [153] proposed an auto-encoder using spectral convolution operators, and *Bouritsas et al.* [154] proposed using spiral convolutions [155]. Despite their impressive modeling precision, a recent work [151] showed that they heavily suffer from poor generalization to unseen identities. This limits their practical use in tasks such as face fitting or expression transfer. The survey in [156] provide a detailed survey of 3D Morphable Face Models over the 20 years since they were first proposed.

Building on the 3DMM framework, and exploiting them to handle fixed topology meshes (same number and order of vertices, section 1.3.6), various methods have been developed for static 3D facial expression generation. Some approaches learn the parameters and deformation coefficients of morphable models from image inputs to generate 3D meshes [157, 158, 159, 160]. Others directly utilize deep neural networks to estimate discriminative 3D faces and expressions from images [161, 162]. While these methods have shown potential for generating static 3D meshes, they do not capture the dynamic nature of facial expressions, which are fundamentally associated with motion and temporal dynamics.

Although many researchers have tackled the challenge of 3D mesh deformation, the synthesis of dynamic 3D facial expressions remains a more complex task. Using the FLAME model [39], some studies have exploited audio features and speech signals for facial motion synthesis [163, 164, 165, 166, 167]. However, these methods rely on external information and do not explicitly model temporal dynamics.

To address this limitation, *Potamias et al.* [168] proposed modeling the motion with a temporal encoder based on a Long Short-Term Memory (LSTM) network, which generates a per-frame latent code starting from a per-frame expression labels. These codes are then fed into a mesh decoder, which generates per-vertex displacements that are applied to a neutral 3D face, resulting in expressive meshes. Despite promising results, this approach is deterministic and for a given label the same displacements are generated without variation, falling to mode collapse. Other methods focus on generating sequences of 3D landmarks to represent expression dynamics, which are then integrated with the FLAME

model to produce mesh sequences. *Zou et al.* [169] introduced a Transformer-based Variational Auto-Encoder to generate sequences conditioned on a desired expression, followed by landmark-guided mesh deformation to animate the neutral mesh. *Otberdout et al.* [170, 171, 172] presented a novel approach that uses a conditional manifold-valued Wasserstein GAN to model facial expression dynamics as curves on a hypersphere. The resulting motion is then transformed into a sequence of landmarks, which are converted into 3D mesh sequences using a spiral convolutional decoder. Additionally, *Zou et al.* [173] applied Diffusion Probabilistic Models (DDPMs), which have achieved significant success in other generative domains, to generate expression dynamics in the form of landmarks that are subsequently used to drive 3D mesh generation.

2.4 Facial Expression Databases

The deep learning field has witnessed a further leap in recent years thanks to the availability of large-scale datasets, which presents models with a variety of examples enabling them to learn and generalize effectively to unseen data. In the facial expression analysis domain, access to large and various labeled datasets is critical for achieving high performance and robust models. However, compared to fields like face recognition, facial expression datasets remain relatively limited in both size and diversity. This is caused by several challenges:

Limited Data Availability: One of the important challenges in facial expression analysis is the lack of training datasets that capture the full range of human emotional expression, including spontaneous expressions in real-world environments. These datasets must account for a variety of factors, including different head poses, occlusions, lighting conditions, and identity-specific characteristics. In addition, the complexity of natural human expressions, which oftentimes involve slight and intricate movements, requires datasets that represent this richness to train models effectively.

Complexity and Cost of Annotation: Labeling facial expressions is essentially subjective and expensive, often requiring expert annotators to ensure consistency and accuracy. The nuance of some expressions, particularly micro-expressions that occur rapidly and are of low intensity, makes them difficult to categorize and annotate consistently.

Cultural and Demographic Diversity: Several factors such as culture, age, and gender can influence the same facial expression, and they are not always interpreted in the same way across these demographics [174]. Most of the existing databases lack sufficient representation of this diversity, leading to models that work well on some populations but poorly on others. In order to develop more robust and general models, collected data must contain a diversity of demographic groups, with sufficient coverage of each group. This inclusivity is critical for avoiding biases during model training, and improving the model applicability to a global population.

Class Imbalance: Another common problem in most facial expression datasets is the representation of classes in an unbalanced manner. For instance, expressions like happiness and neutrality are overrepresented, while the opposite can be said for disgust, anger, and contempt. This imbalance can skew model training, leading to the ability to recognize frequent expressions, while struggling with less represented expressions. Addressing class imbalance requires strategic data collection, and may involve augmenting the underrepresented classes.

Concerns for Privacy and Ethics: Acquiring large-scale, diverse datasets human faces, especially with emotional expressions, raises ethical and privacy concerns. These concerns can be related to informed consent, anonymity, and security of the data collected, which is more of an issue given the sensitive nature of biometric data. Collecting real-world facial data requires following many ethical and legal measures, e.g. the General Data Protection Regulation (GDPR) in the European Union. Thus, these concerns can impede the collection of data, particularly in certain regions or contexts where privacy considerations are heightened.

The integration of **3D data** into facial expression analysis introduces additional complexities beyond those encountered with 2D data:

Complex Data Acquisition: Acquiring 3D facial expression data requires special equipment like multi-camera systems, structured light scanners, or depth sensing devices like LiDAR. These tools are usually costly, need precise calibration, and are used in controlled environments. The logistical challenges associated with setting up and operating such equipment limit the practicality of collecting large-scale, naturalistic 3D datasets. This limitation inhibits the development of models able to generalize across the variability present in real-world scenarios.

Alignment and Registration Difficulties: Accurately capturing and aligning 3D data from different subjects and expressions is a non-trivial task. The differences in facial geometry, expression intensity, and head pose require robust registration methods to map 3D facial scans into a standard template. Misalignment or registration errors can lead to distorted representations, which in turn affects model performance and the validity analyses.

Occlusions and Missing Data: 3D data acquisition is prone to occlusions and missing information due to various reasons, including blocking elements as hair, glasses or even certain head poses that hide some face parts. Also, scanning technology might have design limitations, which results in incomplete data capture, especially around sharp angles and deep corners. To address the problems of occlusions and missing data, it is important to apply complicated techniques such as interpolation or imputation for reconstruction of the lost areas, which may introduce artifacts or inaccuracies into the dataset.

Increased Computational and Storage Requirements: 3D data is more complex than its 2D counterparts, leading to more computational power demands for processing,

storage, and analysis. High-resolution 3D meshes, or even point clouds are vast amounts of information, necessitating efficient data management strategies. The computational can be a limiting factor to researchers with limited resources, which in turn slows down the development and testing new models. Additionally, training models on 3D data often requires more advanced hardware configurations, such as GPUs with larger memory capacities.

Addressing the challenges in facial expression data collection and analysis requires joint efforts from the research community. Recognizing these obstacles, researchers have engaged in various initiatives to overcome data limitations, enhance diversity, and improve the quality of datasets. Below, We show in Fig.2.1 the development timeline of facial expressions datasets, with their details presented in table 2.1.

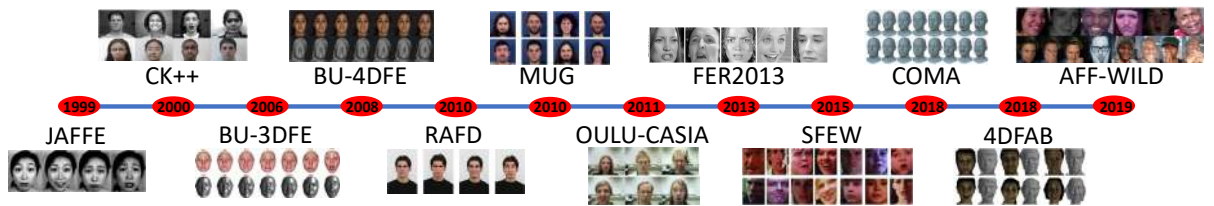


Figure 2.1: Evolution timeline of facial expressions datasets.

Table 2.1: An overview of Some facial expression datasets.

Dataset	#Exps	#Samples	#Subjects	Environment	D/S
2D Facial Expressions Datasets					
JAFFE [175]	7	Images: 213	10	Controlled lab	Static
CK+ [176]	7	Videos: 593	123	Controlled lab	Dynamic
RAFD [177]	8	Images: 1608	67	Controlled lab	Static
MUG [178]	7	Videos: 1462	86	Controlled lab	Dynamic
Oulu-CASIA [179]	6	Videos: 480	80	Controlled lab	Dynamic
FER2013 [180]	7	Images: 35887	N/A	Wild (web images)	Static
SFEW [181, 182]	7	Images: 1766	N/A	Wild (Movies, web)	Static
AFF-wild [183, 184]	7	Videos: 298	200	Wild (Movies, web)	Dynamic
3D Facial Expressions Datasets					
BU-3DFE [185]	7	scans: 2500	100	Controlled lab	Static
BU-4DFE [186]	6	3D seq: 606	101	Controlled lab	Dynamic
COMA [153]	12	3D seq: 144	12	Controlled lab	Dynamic
4DFAB [182]	7	3D seq: -	180	Controlled lab	Dynamic

2.5 Conclusion

In this chapter, we have conducted a comprehensive review of the literature relevant to our thesis, which focuses on exploring novel deep approaches to address existing challenges in facial expression analysis and synthesis. Our review was structured around two main areas: (1) facial expression recognition and (2) facial expression generation.

We provided an in-depth analysis of the current state-of-the-art methods in both FER and FEG, systematically categorizing them based on their fundamental methodologies. By exploring the motivations behind each approach, we highlight the driving factors that have shaped the field. Importantly, we identified and discussed the limitations and challenges within these areas, highlighting the critical aspects that require further investigation.

Our exploration throughout this chapter reveals that, despite significant advancements in facial expression recognition and generation, numerous challenges remain unresolved. This does not only present challenges but also opportunities for improvement and contribution. Motivated by these opportunities and the remarkable advancements in deep learning, we work on investigating novel methodologies for both FER and FEG in the following chapters.

Chapter 3

Dynamic Facial Expression Generation Based on Spatio-temporal Convolutional GANs

3.1 Introduction

Facial expressions are fundamental to human interaction, conveying emotions and intentions non-verbally. Integrating the ability to recognize and synthesize facial expressions into machines facilitates natural and smooth human-computer interactions. This integration opens the door to numerous exciting applications in different fields, including the movie industry, e-commerce, and even the medical field. While facial expression recognition has seen significant advancements, achieving high levels of precision, facial expression generation remains more challenging and less explored in the literature. Recently, the success of Generative Adversarial Networks (GANs) [28] in data generation, particularly in image synthesis, has inspired progress in generating static facial expressions.

However, dynamic facial expression synthesis, which involves generating sequences of expressions over time, is even less studied due to several inherent difficulties: **1)** learning the dataset spatial distribution (facial structure, background), **2)** representing a natural and smooth evolution of facial expressions (temporal representation), and **3)** preserving the identity of the input subject throughout the expression sequence. To address the high complexity of these three tasks, most existing methods tend to treat facial expression generation as a two-step process. One step is for the low dimensional temporal generation (motion), and the other focuses on spatial generation (content). Such methods [128, 130, 136] are typically based on the generation of motion as codes in a latent space, and then combining it with the input image embedding to generate frames individually, using an image-to-image translation network. While these methods are effective in learning facial structures and preserving identity, they struggle with modeling spatio-temporal consistency and maintaining appearance retention due to the independence between frame generations.

Motivated by the success of deep spatio-temporal neural network models in recognition and prediction tasks [187, 188, 189, 190], researchers have proposed a variety of one-step

methods [123, 191, 126, 127] that utilize fractionally strided 3D convolutions to generate the entire video in one step. Although these methods generate videos with improved spatio-temporal consistency, they often suffer from lower video quality, increased noise and distortions, and greater challenges in identity preservation compared to two-step methods. We claim this is due to the high complexity of the three tasks combined in a single network (learning **1.** the spatial representation (identity), **2.** the temporal representation, **3.** identity preservation). This requires a large network with high complexity and a large amount of data, which significantly increases the difficulty of model optimization.

To address these challenges, we introduce a novel dual-encoder architecture that encodes the input image into two distinct latent codes using two feature extractors: a pretrained identity feature extractor (E_{Id}) and a spatial feature extractor (E_S). By leveraging the high performance of state-of-the-art facial recognition systems, E_{Id} , derived from a pretrained facial recognition model, robustly captures and maintains the unique identity characteristics of the subject. Incorporating a pretrained identity encoder into the facial expression generation framework effectively remedies the high complexity associated with simultaneously learning spatial representations, temporal dynamics, and identity preservation. Through transfer learning, the model benefits from robust identity feature extraction, enhances identity consistency, and allows focused learning on expression dynamics. This strategic division of tasks not only improves the quality and realism of generated expressions without compromising identity integrity but also streamlines and accelerates the optimization process, resulting in faster convergence and more efficient training dynamics. Therefore, our dual-encoder approach enhances both the performance and scalability of facial expression generation models, paving the way for more sophisticated and reliable applications in areas such as human-computer interaction, virtual reality, and personalized animation.

In summary, our contributions include the following aspects:

1. We propose a conditional GAN with a single generator and discriminator that generates, at each time step, a dynamic facial expression video corresponding to the desired class of expressions. The generated videos present a realistic appearance and preserve the identity of the input image.
2. We investigate the influence of utilizing two encoders, E_{Id} and E_S , where E_{Id} is a facial identity feature extractor and E_S is a spatial feature extractor.
3. We exploit the high potential of state-of-the-art facial recognition systems by using a pre-trained face recognition model as our generator encoder E_{Id} , which ensures strongly related identity features. This facilitates the decoder's task by providing meaningful and structured features.
4. We thoroughly evaluate our model, both quantitatively and qualitatively, on two public facial expression benchmarks: the MUG Facial Expression Database and the

Oulu-CASIA NIR&VIS Facial Expression Database. We compare it with recent state-of-the-art approaches: VGAN [123], MoCoGAN [128], ImaGINator [127], and motionGAN [136].

The rest of this chapter is structured as follows. In section 3.2, we present state-of-the-art that have directly influenced our research. Section 3.3 introduces the proposed FEV-GAN model. Section 3.4 discusses the experimental configurations, encompassing both quantitative and qualitative analyses. The chapter concludes with section 3.5, offering insights into future research directions.

3.2 Related Work

In section 2.3, we provided a comprehensive overview of state-of-the-art techniques in facial expression generation. Building upon this, this section focuses on specific methods that have directly inspired our proposed model or generating results closely comparable to ours. By examining these approaches, we aim to contextualize our work within the existing literature and highlight the distinctions and advancements introduced by our methodology.

Dynamic facial expression generation methods can be broadly categorized into two main groups: the two-step facial expression generation framework and the one-step facial expression generation framework. The two-step framework generates motion separately as codes in the latent space, which are then fed into a generator to produce individual frames of the video sequentially. The one-step framework employs fractionally strided spatio-temporal convolutions to simultaneously generate both appearance and motion in a single step.

Within the two-step framework, MoCoGAN and MotionGAN are of particular interest. MoCoGAN (Motion and Content Generative Adversarial Network) [128] decomposes video synthesis into content and motion subspaces. The content, representing static aspects such as facial identity, is learned separately from motion, which is captured through a Gated Recurrent Unit (GRU). This separation allows MoCoGAN to generate video frames sequentially, preserving identity while producing a coherent motion sequence. By sampling from distinct latent spaces, MoCoGAN enables the generation of diverse facial expressions while maintaining consistent facial features.

On the other hand, MotionGAN [136] enhances the motion modeling process by employing a manifold-valued Wasserstein GAN. This method models facial landmarks as curves on a Riemannian manifold, specifically a hypersphere, capturing the geometric structure of facial motion. The GAN learns the distribution of expression dynamics for different classes, allowing it to synthesize new motion trajectories. These trajectories are transformed into realistic facial videos through the use of TextureGAN, which generates the texture based on the learned motion. This method effectively preserves natural motion

and anatomical constraints, producing realistic facial expressions with a high degree of dynamism.

In contrast, one-step approaches such as VGAN and ImaGINator aim to simplify the generation process by producing the entire video in a single step. VGAN (Video GAN) [123] decomposes the video into static and dynamic components. The static section represents the background, while the dynamic section represents the moving objects, in this case, the facial expressions. VGAN’s generator consists of two streams: one responsible for generating the background and the other for the foreground, which are combined to create the final video. This model effectively separates motion from static content, but when applied to facial expression generation, it faces limitations. The complexity of subtle facial movements, which cannot be easily split into static and dynamic sections, leads to less accurate representations and challenges in preserving identity across frames.

ImaGINator [127] addresses some of VGAN’s shortcomings by introducing a more sophisticated architecture. It employs a U-Net-like structure, blending an autoencoder with spatio-temporal fusion mechanisms. The spatial features extracted by the encoder are directly linked to the decoder through skip connections, preserving spatial details while ensuring temporal consistency. ImaGINator also uses two discriminators: one that evaluates the entire video sequence for spatio-temporal coherence and another that processes individual frames to maintain visual fidelity. While this approach theoretically improves the smoothness of the generated facial expressions and ensures temporal consistency, it suffers from lower video quality and noise, especially when compared to two-step methods. Moreover, identity preservation is less robust in ImaGINator due to the complexity of simultaneously modeling appearance and motion.

Motivated by the strengths and limitations of these models, our approach introduces a novel one-step framework for facial expression generation that leverages fractionally strided spatio-temporal convolutions. This model is inspired by the design of VGAN but addresses its limitations in capturing fine-grained facial movements and preserving identity. By incorporating spatio-temporal convolution layers and enhancing the generator’s architecture, our approach improves video quality and facial identity preservation while maintaining the simplicity of one-step generation. This model balances the need for spatial and temporal coherence, generating more realistic facial expressions in dynamic settings.

3.3 Proposed Approach

As stated in the introduction, our main aim is to develop a model that generates dynamic facial expression videos from appearance information and expression categories. We formulate our goal as learning a function $G : \{I, L\} \rightarrow \hat{Y}$, where I is the input image, L is the label vector, and $\hat{Y}$ is the generated video.

To achieve this objective, we propose a framework consisting of the following

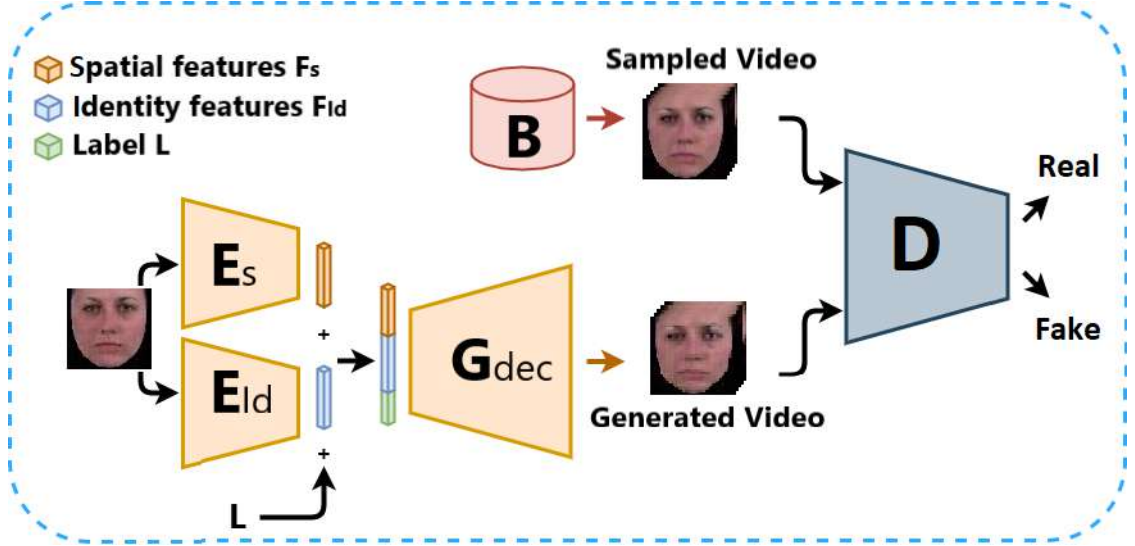


Figure 3.1: Overview of the proposed model for generating facial expressions: **FEV-GAN**. The input image is mapped through two encoders, identity (E_{Id}) and spatial (E_S) encoders, into features F_{Id} and F_S . Both codes and the label L are fed to the decoder G_{dec} that transforms them into a video of the input identity performing the desired expression. The discriminator D has the purpose of classifying realistic scenes from synthesized scenes.

components: a generator network G built on an encoder-decoder architecture. The encoders E_{Id} and E_S process the input image I to extract identity features F_{Id} and spatial features F_S , respectively. Specifically, E_{Id} captures the unique characteristics of the individual’s appearance to preserve identity across the generated video, while E_S focuses on extracting spatial information pertinent to facial expressions. The decoder G_{dec} utilizes the extracted features (F_{Id} , F_S) along with the label L to generate a realistic video $\hat{Y}$ that represents the target expression dynamics. By conditioning on L , the decoder can adjust the expression portrayed in the video. Finally, a discriminator D assists the learning of the generator. The discriminator evaluates the authenticity of the generated videos, ensuring they are indistinguishable from real videos in both appearance and expression categories. It provides feedback to the generator through adversarial training, guiding it to produce high-quality, expression-accurate outputs. An overview of our approach is illustrated in Fig. 3.1.

In the following, the architecture of our network is described, and details on the generator G and the discriminator D are provided.

3.3.1 Generator E_{Id} , E_S , G_{dec}

As shown in Fig. 3.1, our generator consists of three components: a pre-trained image identity encoder E_{Id} , an encoder E_S initialized with random weights, and a video decoder G_{dec} . The encoder E_{Id} is based on the well-known VGG-FACE model [192], a state-of-the-art facial recognition feature extractor. It takes a $(64 \times 64 \times 3)$ RGB

image I as input and transforms it into 1024 feature maps of (14×14) , capturing the facial identity features F_{Id} . The encoder E_S processes the same input image I to extract additional spatial features. Initialized with random weights (e.g., Gaussian initialization), it outputs 512 feature maps of (14×14) features F_S that contain spatial details beyond identity information. Next, the features F_{Id} and F_S , along with the label vector L , are concatenated and fed into the decoder G_{dec} to generate the new video. The decoder combines spatio-temporal convolutions and transposed convolutions (also known as fractionally strided convolutions) to transform the input tensor into the high dimensional generated video. Specifically, 3D convolutions capture both spatial and temporal information, enabling the modeling of dynamic facial expressions, while transposed convolutions effectively upsample the tensor to the desired video dimensions. This process results in a generated video of size $(32 \times 64 \times 64 \times 3)$.

3.3.2 Discriminator D

The discriminator is designed with three objectives: (i) to distinguish between real and synthesized scenes, (ii) to recognize realistic motion patterns, and (iii) to differentiate between various classes of motion. To achieve these tasks, we utilize a five-layer spatio-temporal convolutional network capable of learning both visual content and motion dynamics.

The network takes as input videos of size $(32 \times 64 \times 64 \times 3)$, either from the dataset Y or generated by the generator $\hat{Y}$, along with the corresponding labels vectors. It evaluates whether the video is realistic and checks for consistency between the motion and the provided label.

The discriminator architecture is designed to inversely mirror G_{dec} by replacing transposed spatio-temporal convolutions with direct spatio-temporal convolutions, effectively downsampling instead of upsampling. The final layer is adjusted to produce a binary classification output, indicating whether the input video is real or fake.

3.3.3 Loss Function

To train our model $G : \{I, L\} \rightarrow \hat{Y}$, both network D and G are optimized using the objective function defined in Eq 3.1.

$$\mathcal{L}_{total}(G, D) = \mathcal{L}_{adv}(G, D) + \lambda_1 \mathcal{L}_{rec}(G) + \lambda_2 \mathcal{L}_{id}(G), \quad (3.1)$$

which consists of three losses: Adversarial loss $\mathcal{L}_{adv}$ that helps the generator learn the database distribution. Reconstruction loss $\mathcal{L}_{rec}$ that captures the overall structure of the video, promoting temporal consistency and improving visual quality. Identity loss $\mathcal{L}_{id}$ which ensures facial identity details preservation throughout the generated video. Due to the significant differences in the magnitudes of the three loss values, λ_1 and λ_2 parameters

are used to help stabilize the training and balance the optimization of our model.

Our optimization objective is therefore:

$$G^* = \arg \min_G \max_D \mathcal{L}_{total}. \quad (3.2)$$

It is important to note that the parameters of E_{Id} are fixed during training, since it is a pre-trained model capable of extracting facial features effectively.

Adversarial Loss: Our conditional adversarial loss is a cross-entropy loss applied to both G and D . The goal is for $G(I, L)$ to learn generating videos $\hat{Y}$ that look similar to real videos Y , while D learns to distinguish between real samples Y and generated samples $\hat{Y}$. In this adversarial setting, G aims to minimize the function, while D aims to maximize it.

$$\begin{aligned} \min_G \max_D \mathcal{L}_{adv} = \min_G \max_D \mathbb{E}_{Y \sim P_{data}(Y)} [\log D(Y; L)] \\ + \mathbb{E}_{z \sim P_z(z)} [1 - \log D(G(I; L); L)]. \end{aligned} \quad (3.3)$$

Reconstruction Loss: Our reconstruction loss at the video level is defined by

$$\mathcal{L}_{rec} = [\|Y - \hat{Y}\|_1], \quad (3.4)$$

with the purpose of capturing the overall structure, video consistency, and helping preserve the identity details. This loss is a L_1 norm loss between the generated videos $\hat{Y}$ and the ground truth videos Y . Combining this loss with $\mathcal{L}_{adv}$, helps the generator G create more realistic videos and reconstruct smooth expression motion.

Identity Loss: To preserve the input identity throughout the generated video, we employ an identity loss. This loss is also based on the L_1 norm loss, operates in the feature space rather than the pixel space. We exploit the same architecture of our VGG-FACE encoder to extract the identity features from both input and output data. The identity loss is formulated as:

$$\mathcal{L}_{id} = \sum_{i=0}^N [\|F_{vgg}(I) - F_{vgg}^i(\hat{Y})\|_1], \quad (3.5)$$

where $F_{vgg}(I)$ represents the identity features of the input image I , and $F_{vgg}^i(\hat{Y})$ represents the identity features of the i^{th} frame of $\hat{Y}$, and N is the number of frames considered.

In practice, we apply $\mathcal{L}_{id}$ to only 4 frames out of the total 32 frames in the video. We rely on the spatio-temporal consistency provided by the 3D convolutions in our

network to generalize identity preservation across the remaining frames. This approach reduces computational overhead while maintaining identity integrity throughout the video sequence.

3.3.4 Training Algorithm

Algorithm 1 highlights the training process of the proposed FEV-GAN model. The training dataset (I, L, Y) is fed into Algorithm 1, where I represents the input images, L denotes the target expression labels, and Y consists of the ground truth videos. First, the model parameters are initialized. The parameters of E_{Id} are loaded from the pre-trained VGG-FACE model, while the rest of the parameters are initialized from a Gaussian distribution (1;2). The outer for loops are used to learn from the data for a given number of epochs and iterations in each epoch (3;4). In each iteration, the input image I_i is fed into the encoders E_{Id} (5) and E_s (6), that encode it to identity F_{Id} and spatial F_s features, respectively. These features and the given label L_i are then used by the decoder G_{dec} to generate the corresponding synthetic video $\hat{Y}_i$ (7). Next, the discriminator network D estimates the probability that a video is sampled from the dataset rather than the generator (8;9). The losses are computed based on the outputs of the generator and discriminator (10;12;13;14), and the model is optimized using backpropagation and the Adam optimizer (11;15).

Algorithm 1: Our Learning Algorithm for the proposed model FEV-GAN

Input: Ground-truth Video $Y = [Y_0, Y_1, \dots, Y_n]$; Input Image $I = [I_0, I_1, \dots, I_n]$;
Target expression label $L = [L_0, L_1, \dots, L_n]$;
Output: FEV-GAN model;

- 1 $E_{Id} \leftarrow$ initialized with non-learnable pre-trained VGG-FACE parameters;
- 2 $E_s, G_{dec}, D \leftarrow$ initialized with learnable parameters sampled from a Gaussian distribution;
- 3 **for** the number of epochs **do**
- 4 **for** the number of iterations in an epoch **do**
- 5 $F_{Id}(I_i) \leftarrow E_{Id}(I_i)$;
- 6 $F_s(I_i) \leftarrow E_s(I_i)$;
- 7 $\hat{Y}_i \leftarrow G_{dec}(F_{Id}(I_i), F_s(I_i), L_i)$;
- 8 $E_{real} \leftarrow D(Y_i, L_i)$;
- 9 $E_{recon} \leftarrow D(\hat{Y}_i, L_i)$;
- 10 $\mathcal{L}_D \leftarrow \log(E_{real}) + (1 - \log(1 - E_{recon}))$;
- 11 $D \leftarrow D - \alpha(\partial \mathcal{L}_D / \partial D)$;
- 12 $\mathcal{L}_{rec} \leftarrow ||Y_i - \hat{Y}_i||_1$;
- 13 $\mathcal{L}_{id} \leftarrow ||F_{vgg}(Y_i) - F_{vgg}(\hat{Y}_i)||_1$;
- 14 $\mathcal{L}_G \leftarrow \log(E_{real}) + \mathcal{L}_{rec} + \mathcal{L}_{id}$;
- 15 $G \leftarrow G - \alpha(\partial \mathcal{L}_G / \partial G)$;
- end**
- end**
- end**

3.4 Experiment

To comprehensively evaluate the performance of our model, we conducted an extensive series of experiments. In this section, we first describe the experimental setup used for training and evaluation. Then, we present both quantitative and qualitative results, comparing our model against state-of-the-art methods, including VGAN, MoCoGAN, ImaGINator, and motionGAN. Finally, we provide an ablation study to analyze the influence of each component of our proposed architecture.

3.4.1 Datasets

The evaluation of our method is performed on:

the MUG Facial Expression database (Multimedia Understanding Group Facial Expression database) [178]: contains videos of 86 people, performing seven facial expressions: happiness, sadness, surprise, anger, disgust, fear, and neutral. Videos in the database start and end with neutral expressions and display the peak of the expression in the middle. Each video consists of 50 to 160 RGB frames of 896×896 resolution. The data of 52 subjects is available to authorized users, 25 subjects are available on request and the remaining 9 subjects are available only in the MUG laboratory. In this work, we used the public data of 52 subjects, focusing only on the first half of the expression, which starts with the neutral expression and ends with the expression peak (apex).

Oulu-CASIA NIR&VIS facial expression database [179]: consists of 480 videos of 80 people, between 23 to 58 years old, performing six facial expressions: "happiness", "sadness", "surprise", "anger", "disgust" and "fear", going from neutral to apex expression. Each video consists of 9 to 72 RGB frames of 320×240 resolution, beginning with a neutral expression and end with the apex of the corresponding expression. The whole database is available to authorized users.

3.4.2 Implementation Details

Before using the data, we first split it in a subject independent manner into two sets, 80% for training and 20% for testing. We then cropped the face area and removed the background using OpenFace [193], sampled 32 frames from all videos using the framework proposed in [194], and resize each frame to 64×64 pixels. After preprocessing, we ended up with $32 \times 64 \times 64 \times 3$ videos of different expressions and subjects with black background.

The weights of the networks E_S , G_{dec} and D are initially sampled from a Gaussian distribution of the mean zero and a standard deviation 0.01. The weights of the network E_{Id} are initialized with the weights of the pre-trained VGG-FACE [192], and these weights are frozen during the training phase. Gradient descent is used for 400 epochs, in order to solve the optimization problem defined in Eq. 3.2. Binary cross entropy loss is used

for the adversarial loss, and L_1 norm is used for both reconstruction and identity losses. Additionally, Adam Optimizer [195] is used to train the model with a learning rate of 0.0002 and a momentum of 0.5. The pixels values of the input image and videos are scaled to the interval $[-1.1]$. Each layer of G_{dec} is followed by the ReLU activation function [196] and batch normalization [197], except for the output layer, which uses the tanh function. LeakyReLU [198] and batch normalization are used in the discriminator except for the input layer.

The implementation of the network is conducted using the TensorFlow Framework, building upon the implementation from [123]. Training on the MUG dataset took approximately 70 hours using an Nvidia GeForce GTX 1650 GPU with 4 GB of memory, while training on the Oulu-CASIA NIR&VIS dataset took approximately 40 hours on an Nvidia Titan V GPU with 12 GB of memory.

Regarding the state-of-the-art models used for comparison, we used the public codes of VGAN and ImaGINator provided by the authors, with some minor changes where necessary. Specifically, trained two versions of VGAN, the original version and a conditional version. we trained two versions of VGAN: the original and a conditional version. In the conditional version, an encoder transforms the input image into a latent code, which is then concatenated with the labels and fed into the generator to produce videos with the desired target expressions of the same input face. For MoCoGAN, we utilized the results presented in [127], and for motionGAN [136], we relied on the results from the original paper.

3.4.3 Evaluation Metrics And Baselines

To deeply evaluate our model quantitatively, we use several metrics:

- 1) **PSNR** $\uparrow$ (Peak Signal-to-Noise Ratio): measures the pixel-level similarity between the generated videos and the corresponding ground truth.
- 2) **SSIM** $\uparrow$ (Structural Similarity Index Measure): represents the structural similarity between real and reconstructed videos, emphasizing perceptual quality beyond pixel-level comparisons.
- 3) **ACD** $\downarrow$ (Average Content Distance) [128]: measures content consistency in generated videos, based on the average of L_2 distances between facial features of consecutive frames in a generated video. However, ACD only reflects identity consistency within the generated video, lacking information on the identity preservation from the input image.
- 4) **ACD-I** $\downarrow$ (Average Content Distance - Identity) [199]: an ACD extension that measures the identity preservation of the input face in the generated video. It computes the average L_2 distance between facial features from the input image and the frames in the generated video. Facial feature vectors are extracted using OpenFace [200], a deep learning-based facial recognition model that surpasses human accuracy.

Table 3.1: Quantitative comparison results of FEV-GAN and baseline models using PSNR, SSIM, ACD, and ACD-I metrics.

	Trained on MUG				Trained on Oulu-Casia			
Model	PSNR $\uparrow$	SSIM $\uparrow$	ACD $\downarrow$	ACD-I $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	ACD $\downarrow$	ACD-I $\downarrow$
VGAN [123]	16.32	0.41	0.14	1.55	15.09	0.61	0.27	1.37
C-VGAN [123]	22.30	0.83	0.09	0.73	15.98	0.61	0.25	1.26
MoCoGAN [128]	18.16	0.58	0.15	0.90	—	—	—	—
ImaGINator [127]	20.29	0.85	0.08	0.29	22.98	0.84	0.07	0.16
motionGAN [136]	25.9	0.90	—	—	24.44	0.89	—	—
FEV-GAN	27.10	0.91	0.09	0.23	25.61	0.89	0.12	0.19

Higher SSIM and PSNR scores indicate better generated videos quality, while lower ACD scores indicate similar faces in consecutive generated video frames, and lower ACD-I values indicate higher similarity between faces in the input images and the generated videos.

3.4.4 Experimental Results & Analysis

Quantitative results

To perform our quantitative analysis, we first generate 106 videos of 18 different subjects from the testing subset, each performing the six basic facial expressions.

First, we demonstrate that our model FEV-GAN offers better reconstruction capabilities than all baselines, as measured by **PSNR** and **SSIM**. Table 3.1 indicates that our model generates better quality videos with less noise, and preserves the general structure of the input through all the video. Next, we evaluate the content consistency using **ACD** metric. Our model achieves content consistency comparable to models utilizing spatio-temporal convolutions, such as VGAN and ImaGINator, which are known for their high content consistency. FEV-GAN also surpasses MoCoGAN, which relies on Image-to-Image translation techniques.

In terms of identity preservation, our model significantly outperforms VGAN, C-VGAN, and MoCoGAN, and achieves competitive results compared to ImaGINator. While VGAN does not prioritize identity preservation, its results serve as a reference point for scenarios where identity consistency is not maintained.

We note that the ACD and ACD-I comparison does not include the motionGAN, for the reason that the authors have used a different approach to calculate the metrics. This makes a direct comparison challenging, although motionGAN’s results in other evaluations suggest strong performance in related tasks.

Qualitative results

Fig. 3.2 presents examples of generated videos of the six facial expressions of the same identity. Fig. 3.3 demonstrate examples of generated videos of the six facial expressions of different given identities. Both figures show the results of our proposed model trained on the MUG dataset (left) and the Oulu-Casia dataset (right). In addition, Fig. 3.4 shows more examples of videos generated by our model. The generated videos are sampled randomly from the test results. We recall that the data used for learning and testing the model is subject independent. The synthesized videos generally display natural, smooth, and continuous facial variations that align well with the input label. They also preserve the features of the input images, such as identity, beard, and glasses. Our model shows slightly better results on MUG data than on Oulu-Casia data, likely due to the latter's fewer data instances with a variety of skin colors, lighting conditions and with accessories (glasses) on the face, which introduce additional challenges for model learning.

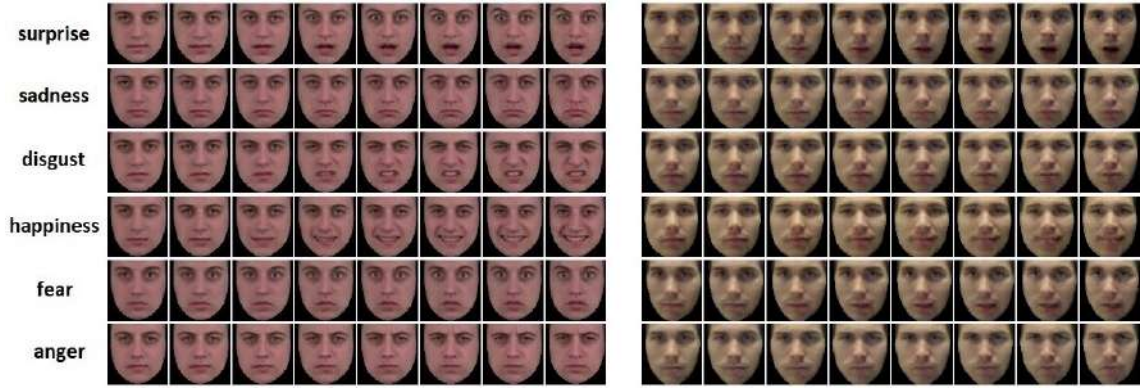


Figure 3.2: The generation of facial expression videos on the MUG database (left) and Oulu-Casia (right). The image sequences show the six basic facial expressions of the same subject on the test dataset. The presented images are sampled every two frames.

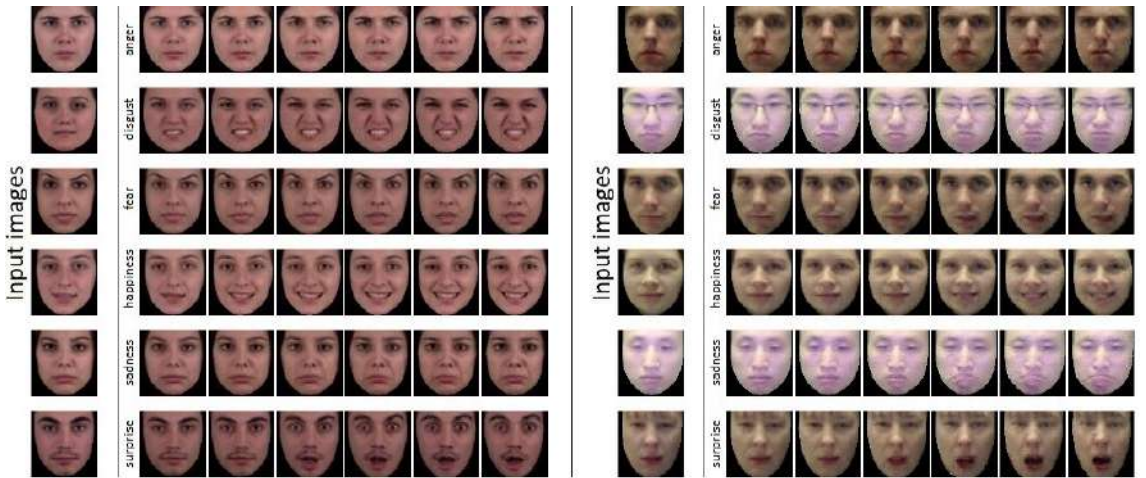


Figure 3.3: Examples of videos generated by our model, displaying the six basic facial expression performed by a person given in the input image.

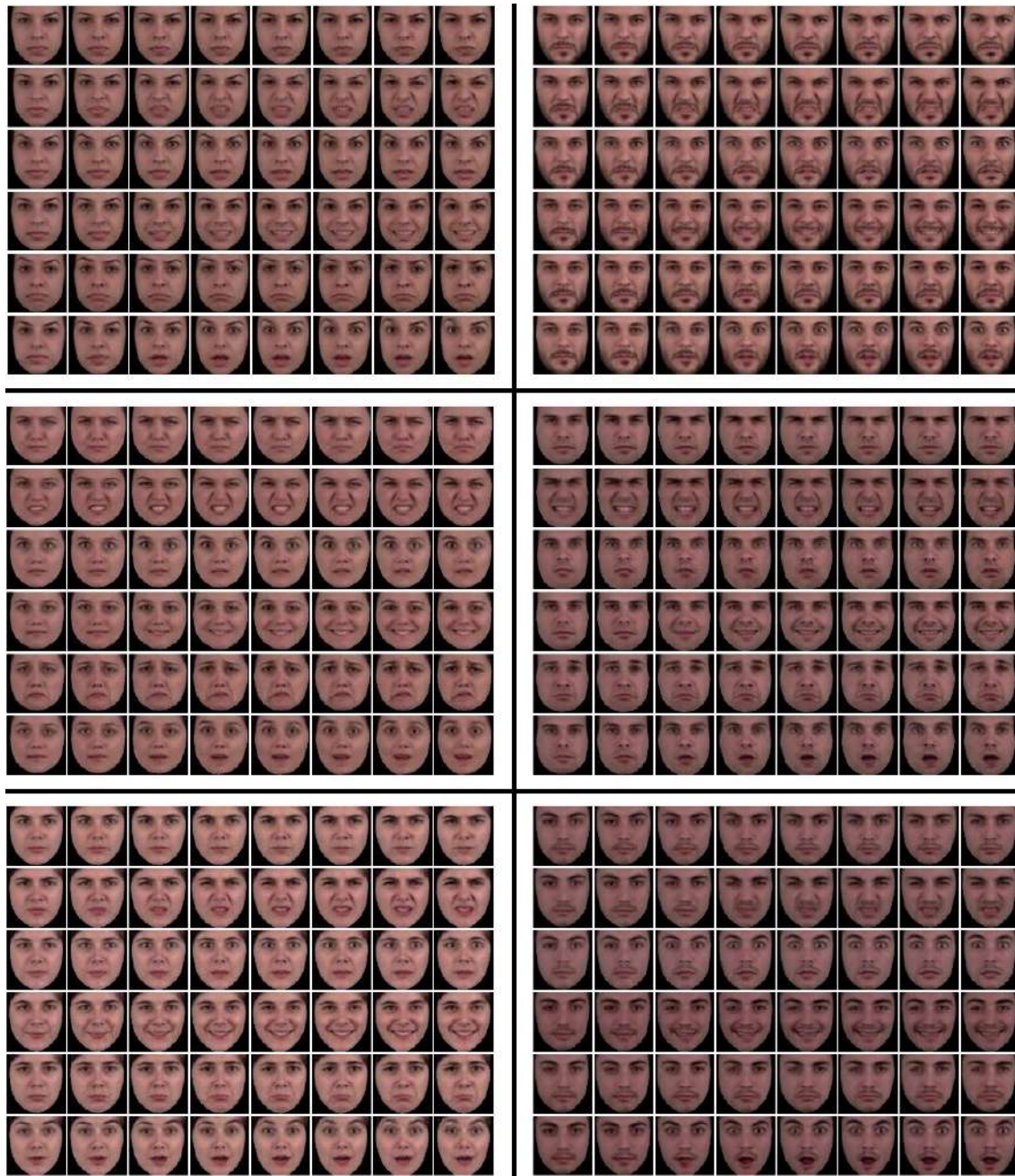


Figure 3.4: examples of sequences generated by our proposed model. The image sequences in each group show the six basic facial expressions performed by the same subject on the MUG test dataset.

In Fig. 3.5, a comparison between our results and the state-of-the-art on the MUG database is conducted. The figures illustrate videos generated by our model, VGAN, C-VGAN, MoCoGAN, ImaGINator and motionGAN. On the left side of the image, we show the input images utilized to generate the videos, in the middle, we show the videos generated by our model (a, c, e, g, i), and on the right side we show the videos generated by the baselines (b, d, f, h, j). In each comparison, each line, the videos generated by our model and the baseline contain the same identity and expression. In the first comparison (a,b), VGAN shows a natural smooth expression, but a total loss of the identity. In (c,d), C-VGAN generally preserves the input identity but with detail loss (like skin color), noise and artifacts. In (e,f), MoCoGAN offers sufficient identity preservation, but the identity is already used in the training set. MoCoGAN also suffers from the unnatural expression and distortions in the generated images, we suspect this causes the high ACD-I value even when the identity is preserved better than C-VGAN. In (g,h), ImaGINator preserves the input identity, but changes the skins color and displays an unnatural expression. In (i,j), motionGAN shows sufficient identity preservation and structure consistency, but displays artifacts near the mouth and nose area. The figure demonstrates that our model generally surpasses the baselines in terms of identity preservation and the quality of the generated videos. It also demonstrate that our model maintains the same content-consistency and expression naturalness as spatio-temporal models.

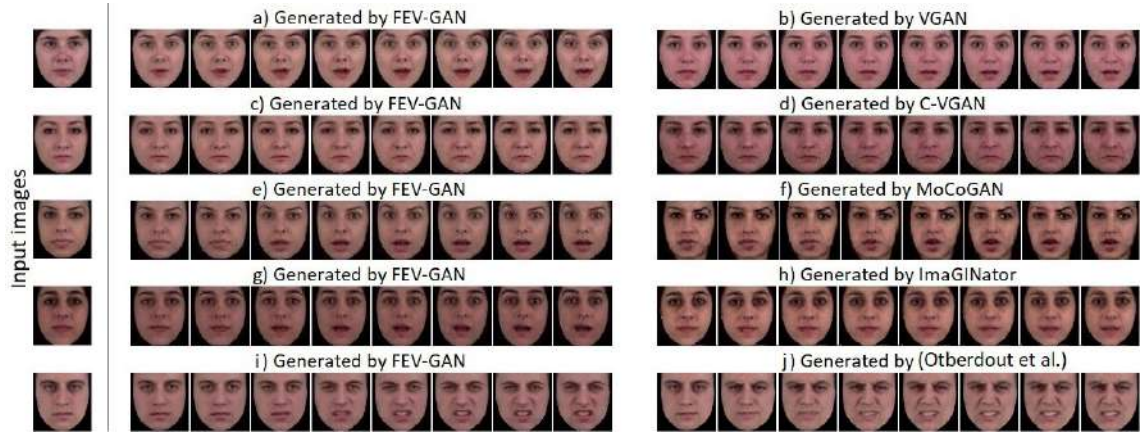


Figure 3.5: Qualitative comparison of sequences generated by FEV-GAN model and by the state-of-the-art models on the MUG database. The sequences generated by our model (a, c, e, g, i), by VGAN (b), C-VGAN (d) and by ImaGINator (h) are randomly selected from the test results. The sequence generated by MoCoGAN (f) and by motionGAN (j) are taken from the original papers. All the images are sampled with the time step of 4.

More qualitative comparison is displayed in Fig. 3.6. In this comparison, We observe that MoCoGAN (b) produces large artifacts and generates videos where the expression intensity fails to increase smoothly, resulting in immediate transitions. The video starts with a neutral expression and suddenly shifts to the target expression. In contrast, G3AN [126] (d) shows smoother transitions between expressions, but the frames during these transitions are noisy and contain noticeable artifacts. RV-GAN [129] (f) delivers better

overall video quality and smoother expression transitions compared to MoCoGAN and G3AN. However, like the others, it struggles with preserving the input identity, which is a crucial factor in facial expression generation. C-VGAN (h) offers smoother facial expressions and maintains the main structure of the input identity, but it sacrifices many important details in the process. ImaGINator (j), while successful at preserving identity, generates unnatural expressions and suffers from frame distortions. motionGAN (l) and VDSM [117] (n) display natural, smooth facial expressions with sufficient identity preservation. However, VDSM images are blurry, and motionGAN images contain artifacts near the mouth area. In comparison, our generated samples show smoother, more continuous facial expression evolution with strong identity preservation, while also exhibiting less noise and fewer artifacts than the state-of-the-art models.



Figure 3.6: Qualitative comparison of sequences generated by FEV-GAN model and by the state-of-the-art models on the MUG database. The sequences generated by our model (a, c, e, g, i, k, l), by C-VGAN (h) and by ImaGINator (j) are randomly selected from the test results. The sequence generated by MoCoGAN (b), (Otterdout et al., 2019) (h), G3AN (d), RV-GAN (f), and by VDSM (n) are taken from the original papers. The absence of input identity indicate that the baseline does not preserve the input identity.

Subjective comparison - we performed a subjective evaluation, in which we aim to obtain opinions of human raters on videos generated by our method compared to the state-of-the-art. We asked 17 volunteers, all PhD students and researchers in the field of deep learning from various university laboratories, to participate in the evaluation. Each volunteer was presented with more than 30 pairs of videos, where one video was generated by our model and the other by one of the baseline models. The volunteers were not informed of the video sources. Their task was to choose the video that performed better based on three criteria: expression naturalness, identity preservation, and video quality.

As shown in Table 3.2, our method largely outperforms the compared models by a large margin. The raters show strong preference for our method over C-VGAN (91.44% VS

08.56%), MoCoGAN (78.09% VS 21.91%), ImaGINator (71.07% VS 28.93%), motionGAN (68.05% VS 31.95%), which corresponds to the quantitative results.

Table 3.2: Subjective comparison of FEV-GAN and baseline models. The reported results are the mean of the raters preferences.

Models	Rater preference(%)
FEV-GAN/C-VGAN	91.44% / 08.56%
FEV-GAN/MoCoGAN	78.09% / 21.91%
FEV-GAN/ImaGINator	71.07% / 28.93%
FEV-GAN/motionGAN	68.05% / 31.95%

The volunteers also provided qualitative feedback. When compared to C-VGAN, our method displayed similar levels of expression naturalness and content consistency while surpassing it in identity preservation and the overall quality of generated videos. Against MoCoGAN, our model outperformed it in all evaluated criteria. In comparison with ImaGINator, our method achieved similar identity preservation, but generated smoother and more natural facial expressions. According to the raters, the closest competitor to our model is motionGAN, as it slightly outperforms us in identity preservation, but our model produces videos with less noise and fewer artifacts.

The results of the quantitative and the qualitative analysis prove that our model maintain the same content consistency and expressions naturalness as other spatio-temporal convolution models, while offering better quality and identity preservation.

Ablation Study - In this section, we focus on demonstrating the significance of the components used to build our model. We showcase the effect of the double encoder method and the effect of the pre-trained facial feature extractor. This is conducted by performing the ablation study on the MUG database. We follow the same evaluation metrics used previously and test multiple versions of our model, each with one key component removed, to analyze the effect. We first train two modified versions of our model. In the first, we discard the spatial encoder E_S , while in the second, we remove the identity features encoder E_{Id} . Both networks are trained and evaluated using the same dataset, parameters, losses, and the same number of epochs, as the full network.

Fig. 3.7 shows sequences generated a) by our full network, b) by the network w/o E_S , and c) by the network w/o E_{Id} .

a) shows that our full network generates videos where the input identity performs the target expression with high quality and enhanced facial details.

b) illustrates that removing the spatial encoder (E_S) results in videos of noticeably lower quality, with poorer identity preservation. The network struggles to retain essential details like skin color and facial features, leading to a less accurate portrayal of the input

identity.

c) the absence of E_{Id} generates videos with more natural expressions, but introduces greater noise and distortions. More importantly, these videos fail to preserve the identity of the input image, generating unrecognizable faces compared to the original.

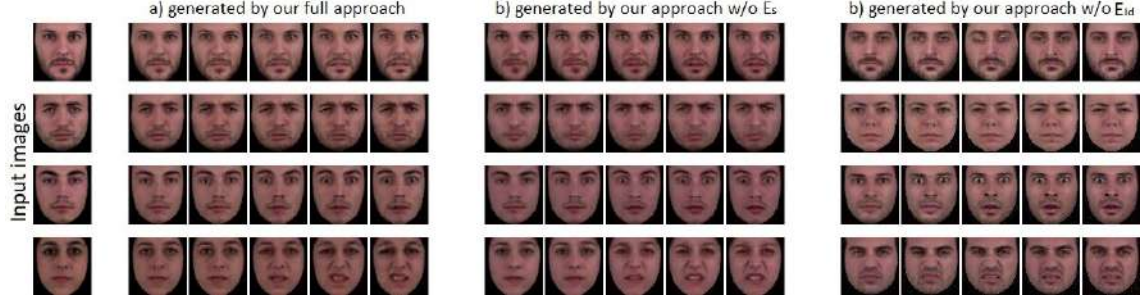


Figure 3.7: Qualitative comparison of sequences generated by our full FEV-GAN model (a), by FEV-GAN w/o E_S (b), and by FEV-GAN w/o E_{Id} (c) on the MUG database.

Table 3.3 summarizes this ablation study into quantitative metrics. The results clearly demonstrate that the full version of the network outperforms the modified versions in terms of video quality, content consistency, and identity preservation. This improvement can be attributed to the dual-encoder architecture. The E_{Id} encoder effectively captures identity-related features, reducing the burden on the E_S encoder, which can then focus on extracting complementary spatial details. This collaboration between the two encoders enhances the decoder’s ability to generate videos with superior quality and finer facial details.

Table 3.3: Performance comparison of our model without E_{Id} and E_S

Model	PSNR $\uparrow$	SSIM $\uparrow$	ACD $\downarrow$	ACD-I $\downarrow$
FEV-GAN w/o E_S	25.44	0.90	0.11	0.34
FEV-GAN w/o E_{Id}	19.71	0.69	0.13	1.22
Full FEV-GAN	27.10	0.91	0.09	0.23

3.4.5 Discussion and Limitations

From both the quantitative and qualitative comparisons, we conclude that our model significantly outperforms the benchmarks C-VGAN and MoCoGAN, across all evaluation criteria. As for the comparison to ImaGINator, our model exhibits similar content consistency, but ImaGINator results display unnatural expressions with clear distortions in the mouth and eyes area Fig. 3.5.f and Fig. 3.6.j. For motionGAN, Similarly, while it achieves high identity and structural preservation, it also introduces artifacts in the mouth and eye regions. We argue that our proposed model generates videos of higher quality, with more natural expressions and minimal noise or distortions.

However, there are some inaccuracies that flaw our FEV-GAN model. Fig. 3.8 illustrates some examples of imprecise video generation. For instance, the model can synthesize the

teeth region even when it is not visible in the source image but fails to generate accurate eye regions when they are not clearly presented (1st example of Fig. 3.8). Also, since the model is trained using neutral expression input images, when a non-neutral input is provided, the generated video fails to display the transition from neutral to the target expression (2nd example of Fig. 3.8). Additionally, the model occasionally generates videos with some minor distortions in the eyes or mouth regions (3rd and 4th example in Fig. 3.8).



Figure 3.8: Examples of flawed video generation by our model.

3.5 Conclusion & Perspectives

In this chapter, we introduced a novel Conditional GAN, namely FEV-GAN, designed to effectively generate dynamic videos of the six basic facial expressions, given only a single neutral input image and a target facial expression category. Specifically, we address the low quality and identity preservation issues encountered by facial expressions generation models using fractionally strided spatio-temporal convolutions.

Based on our state-of-the-art study, these issues are related to the difficulty of the task of generating dynamic facial expressions. Our FEV-GAN model remedies these issues by employing two distinct encoders E_{Id} (identity-encoder) and E_S (spatial-encoder) to extract identity features F_{Id} and spatial features F_S , respectively. These features are given as input to the decoder G_{dec} , which works to preserve facial identity and generate high-quality facial expression videos.

We have deeply evaluated our method on two well-known benchmark databases: The MUG facial expression database and Oulu-CASIA NIR&VIS facial expression database. We utilized a comprehensive set of metrics (PSNR, SSIM, ACD, ACD-I) for quantitative assessment, and expert human evaluation for qualitative analysis. Our results demonstrate that FEV-GAN significantly outperforms the current state-of-the-art methods in dynamic facial expression generation, affirming the effectiveness of our approach.

However, despite the performance of our FEV-GAN model, our tests revealed certain limitations that indicate areas for future improvement. One of the primary challenges is

the relatively low resolution of the generated videos, limited to (32x64x64x3). The model is also unable to incorporate important contextual elements such as background details, hair, and clothing, which are important to the overall visual clarity. Additionally, the model occasionally generates artifacts and noise, particularly in areas around the eyes and mouth, which can disrupt the naturalness of the generated expressions. These limitations motivate us to explore new directions for improving the generated video quality, increasing resolution, and eliminating artifacts and distortions. Future work will focus on enhancing these aspects to push the boundaries of facial expression generation even further.

Chapter 4

Facial Expression Enhancing with Recurrent DDPMs

4.1 Introduction

Non-verbal human communication, including facial expressions and human actions, plays a significant role in interpersonal interactions, with deep foundations in psychology, sociology, and cognitive science. Extensive research has focused on analyzing this communication form [193, 201, 202, 203], unlocking new possibilities across various application fields, including virtual reality, gaming, telemedicine, and emotional AI assistants. With the rise of generative models, static 2D facial expression generation has achieved significant success [103, 204, 205, 206], but it falls short in capturing the dynamic aspect of facial expressions, which involve continuous, smooth movements, rather than a static pose. Dynamic facial expression generation has received less attention due to the incorporation of the temporal dimension, necessitating a dual focus on spatial and temporal understanding, respecting spatio-temporal consistency, and learning identity preservation through all video frames.

Several strategies for facial expression generation have been explored, including transfer-based methods for expression transfer [207, 121, 110], the use of linear or non-linear coefficients to manage the expression intensity and temporal evolution [208], spatio-temporal representation learning for one-step video generation [123, 127, 209], and motion space learning for generating facial expression frames [136]. However, initial tests highlighted common limitations, such as low-quality and low-resolution results (32x32 or 64x64 pixels), difficulties in background generation, leading to distorted or absent backgrounds, and a lack of essential details like hair, neck, and clothing in the synthesized videos.

Several strategies for facial expression generation have been explored, including expression transfer methods [207, 121, 110], techniques for controlling expression intensity over time [208], spatio-temporal one-step video generation [123, 127, 209], and motion space learning for generating expression frames [136]. Despite these efforts, limitations such as low-quality, low-resolution outputs (e.g., 32x32 or 64x64 pixels), distorted or missing backgrounds, and a lack of essential details like hair and clothing persist. Fig.4.1

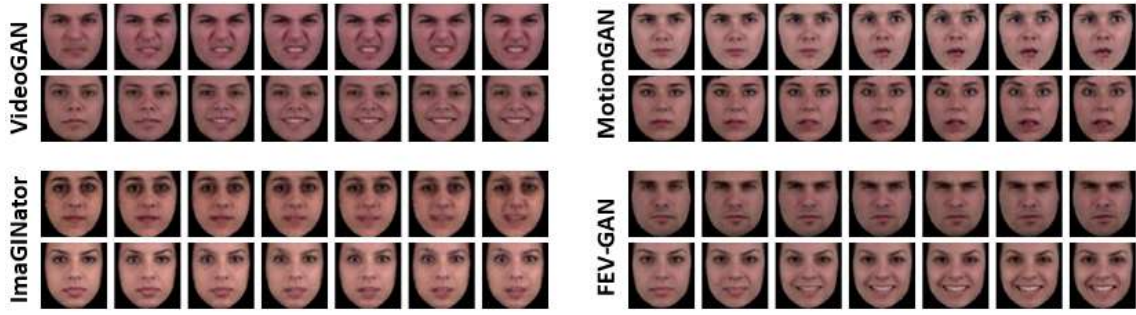


Figure 4.1: Examples of videos generated by VideoGAN, ImaGINator, motioGAN, and FEV-GAN.

provides examples of facial expression videos generated by these methods, demonstrating issues like blurry frames and inconsistencies in the background.

To address these limitations, recent approaches aim to utilize successful image generators like StyleGAN [3] for facial expression generation. These methods [210, 211, 212] involve learning to manipulate the embedding space of StyleGAN to generate high-quality facial images. However, they struggle with content consistency as the generation process operates frame-independently, limiting their ability to manipulate images outside the model’s domain or in areas less covered by the pre-trained model. Additionally, the lack of temporal coherence remains a critical challenge. Another approach addressing the low resolution and quality issue is the diffusion-based models (DM) [31]. These models have demonstrated remarkable efficacy in image synthesis [213, 32, 214, 215], excelling in capturing complex data distributions through a series of diffusion refinement steps. While diffusion models have shown notable success in facial expression generation [141, 8], their application to video generation presents new challenges. The integration of both spatial and temporal dimensions significantly increases computational demands, particularly with the use of 3D convolutions, which increases memory consumption and slows down both training and inference.

In response to the persistent challenges associated with low resolution and quality in resource-efficient facial expression generation methods, this chapter introduces a novel diffusion-based approach, namely **FacEnhance**, for enhancing low-resolution efficient methods videos to the state-of-the-art level. Specifically, the proposed model addresses deficiencies in efficient models, such as low resolution, low quality, and lack of background details. It operates on generated videos to accomplish facial super-resolution, quality enhancement, and the addition of background/details. Taking low-resolution videos (64x64) of an individual performing a facial expression and a higher-resolution neutral image (192x192) of the same individual as input, the model integrates facial expressions with background and identity details, yielding videos of (192x192) pixels that combine the facial expression from the low-resolution video with the identity details and background sourced from the higher-resolution image. We note that the proposed model can theoretically generate videos that potentially reach even higher resolutions. However,

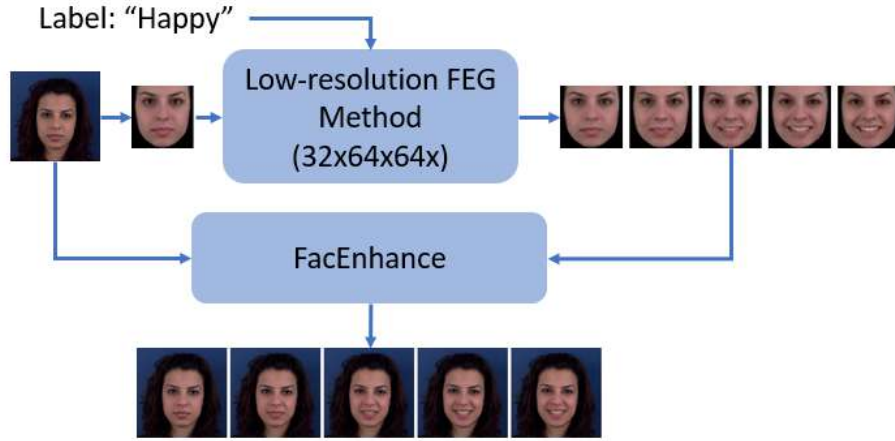


Figure 4.2: the integration of our proposed model as an extension to enhance generated low-resolution videos.

we limit our work to (192x192) resolutions due to constraints in computational power.

The proposed model complements existing low-resolution facial expression generation models by enhancing their results spatially. By applying our model to low-resolution generated videos, we achieve a balance between efficiency and quality. In this study, we utilize facial expression generation models introduced in [123, 127, 136, 209] to produce black background low-resolution facial expression videos. Our model then enhances these videos, resulting in (192x192) facial expression videos with improved quality and integrated background details. The synergy of both models ensures high-quality facial expression generation without excessive computational demands. Fig.4.2 demonstrates the integration of our model into the state-of-the-art pipeline as an extension for enhancing low-resolution methods. The process begins with preprocessing the input image to a low resolution with no background. This image is then passed through existing low-resolution models to generate a low-resolution video representing the facial expression. Finally, FacEnhance uses the generated video along with the input image to upscale the resolution and improve quality while maintaining content consistency.

We extensively evaluate our model quantitatively and qualitatively on the widely recognized MUG facial expression database. Our comparison includes recent state-of-the-art approaches such as ImaGINator [127], LDM [32], VDM [141], and LFDM [8]. The results of our experiments demonstrate the effectiveness of our approach in enhancing low-resolution facial videos to state-of-the-art quality while maintaining content and identity consistency.

The rest of this chapter is structured as follows. Section 4.2 presents state-of-the-art that have directly influenced our research. In Section 4.3, we present the proposed FacEnhance model. Section 4.4 presents the experimental configurations, encompassing both quantitative and qualitative analyses. The chapter concludes with section 4.5, offering insights into future research directions.

4.2 Related Work

In Section 2.3, we presented a detailed review of state-of-the-art techniques in facial expression generation. Expanding on that, this section focuses on specific methods that have directly influenced our research or produced results closely aligned with ours. By analyzing these approaches, we position our work within the context of the existing literature and underscore the key innovations and improvements brought about by our proposed model, FacEnhance.

We first address existing low-resolution methods, which typically generate facial expression videos at resolutions of 64×64 pixels or less, lacking detail in regions such as the hair, neck, and background. These methods include [208, 128, 216, 123, 127, 126, 209]. Specifically, we focus our evaluation on VideoGAN [123], ImaGINator [127], MoCoGAN [128], and FEV-GAN [209], as these methods have provided insights for our work.

VideoGAN [123] utilizes a GAN framework to decompose video generation into two parts: a static background and a dynamic foreground. This dual-stream generator architecture allows VideoGAN to generate videos that capture both spatial and temporal aspects, making it well-suited for producing continuous sequences of frames. Building on VideoGAN, FEV-GAN [209] enhances its architecture with a dual-encoder system that extracts both identity and spatial features from a static facial image. These features are then combined with expression labels to generate smooth and consistent facial expressions over time, improving both the temporal coherence and identity preservation of generated videos. ImaGINator [127] introduces a spatio-temporal fusion architecture, utilizing a U-Net-like structure that passes low-level spatial features from the encoder directly to the decoder. This design preserves fine details while generating realistic and temporally consistent videos. MotionGAN [136] takes a different approach by explicitly modeling motion trajectories. This is achieved by learning a motion space and sampling from it to generate motion embeddings. By leveraging facial landmarks and integrating a manifold-valued Wasserstein GAN, MotionGAN captures the complex, non-linear motions involved in facial expressions. These motion embeddings are then used by an image-to-image generator to synthesize the facial expression frames, offering a more accurate representation of dynamic sequences.

However, these models are often limited in terms of resolution and quality when generating facial expression videos, and they fail at including details such as hair, neck regions, or backgrounds. To address these challenges, we introduce FacEnhance, a diffusion-based model aimed at enhancing the resolution and quality of generated facial expression videos. FacEnhance is designed to bridge the gap between low-resolution facial generation methods and high-resolution video generation, delivering better quality, more realistic results. We further compare our model with recently proposed diffusion-based methods, such as Video Diffusion Models (VDM) [141], Latent Diffusion Models (LDM) [32], and Latent Flow Diffusion Models (LFDM) [8].

VDM [141] extends diffusion models [31] into the spatio-temporal domain by incorporating 3D spatio-temporal convolutions that jointly model appearance and motion. While this allows VDM to achieve high content consistency across frames, the computational complexity and resource demands of diffusion models in the video domain are significantly high. LDM [32] remedy these computational challenges by compressing the input data into a lower-dimensional latent space before applying the diffusion process. By operating in this latent space, LDMs significantly reduce the computational challenge while maintaining the ability to produce high-resolution and high-quality images and videos. LFDM [8] separates the spatial and temporal components of facial expression generation. More precisely, LFDM employs a 3D-UNet-based diffusion model to synthesize optical flow sequences representing the motion in the latent space. These generated optical flow sequences are then utilized by a conditional auto-encoder to animate the input image.

4.3 Proposed Approach

The primary objective of our model is to enhance the quality, resolution, and background of facial expression videos generated by low-resolution models. Specifically, we aim to process a low-resolution (64x64) facial expression video in conjunction with a higher-resolution input image of the same person, featuring a neutral facial expression, background, and additional details. The objective is to generate an improved higher-resolution (192x192) video, with enhanced facial expressions, background, and finer details. This task is formulated as training a model to learn a function denoted as: $\text{FacEnhance}(v_{low}, I_{Id}) \rightarrow \bar{v}_{high}$, where $v_{low} = [f_{low}^1, f_{low}^2, \dots, f_{low}^N]$ represents the low-resolution video with N frames, I_{Id} is the higher-resolution input image, and $\bar{v}_{high} = [\bar{f}_{high}^1, \bar{f}_{high}^2, \dots, \bar{f}_{high}^N]$ is the generated higher-resolution video.

To elaborate, each n^{th} low-resolution frame f_{low}^n (sampled from v_{low}), combined with the neutral input image I_{Id} and the high-resolution frame generated from the previous step $\bar{f}_{high}^{n-1}$, is fed into a diffusion model. The diffusion model then generates the high-resolution frame $\bar{f}_{high}^n$ incorporating background details. This process is expressed as $\text{Diff}(f_{low}^n, I_{Id}, \bar{f}_{high}^{n-1}) \rightarrow \bar{f}_{high}^n$. Finally, all frames are aggregated to construct the complete high-resolution video $\bar{v}_{high}$. Fig.4.3 illustrates the overview of the proposed model.

4.3.1 Our proposed Model: FacEnhance

In our proposed model, we leverage conditional denoising to guide the facial image generation (diffusion models are detailed in section 1.3.3). The condition label is composed of 3 guiding elements: 1) n^{th} low-resolution frame f_{low}^n , which provides the expression details, 2) identity image I_{Id} , which supplies identity features and background information, and 3) previous generated high-resolution frame $\bar{f}_{high}^{n-1}$, which ensures content consistency between frames. The conditional distribution $p_{\theta}(x|I_{Id}, f_{low}^n, \bar{f}_{high}^{n-1})$ enables the generation

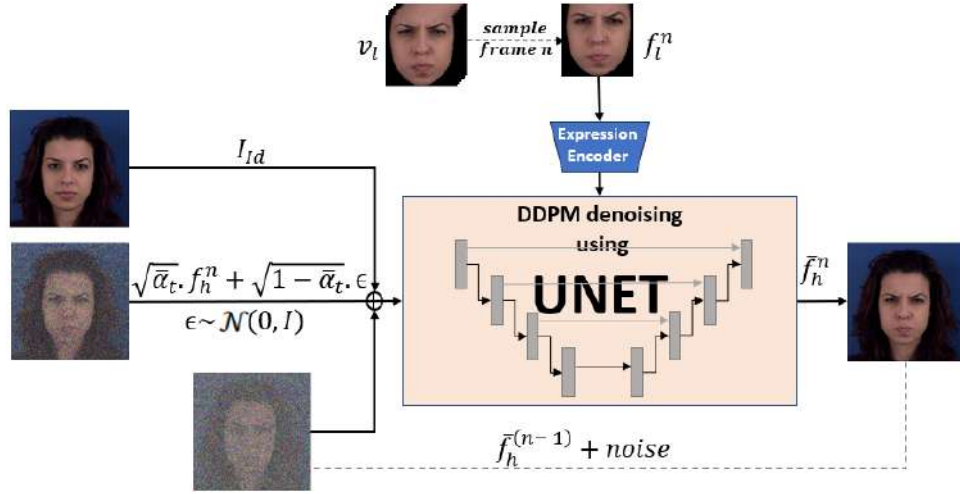


Figure 4.3: Overview of the proposed facial expression enhancement model. The diffusion model refines Gaussian noise iteratively, guided by a low-resolution frame f_l^n from input video v_l , a neutral high-resolution image I_{Id} , and the previously generated frame $\bar{f}_h^{n-1}$, resulting in an improved higher-resolution frame $\bar{f}_{high}^n$.

of data dependent on the specified conditioning images.

Throughout the diffusion process, noise is systematically injected into the input image, independent of the conditioning label $q(\bar{f}_{high,t}^n | \bar{f}_{high,t-1}^n)$. On the other hand, the denoising model $p_\theta(\bar{f}_{high,t-1}^n | \bar{f}_{high,t}^n, I_{Id}, f_{low}^n, \bar{f}_{high}^{n-1})$ considers both the image from the noised image ($\bar{f}_{high,t}^n$) and the conditioning labels $(I_{Id}, f_{low}^n, \bar{f}_{high}^{n-1})$.

The reverse process is modeled as:

$$p_\theta(\bar{f}_{high,0}^n, \dots, \bar{f}_{high,T-1}^n | \bar{f}_{high,T}^n, I_{Id}, f_{low}^n, \bar{f}_{high}^{n-1}) = \prod_{t=1}^T p_\theta(\bar{f}_{high,t-1}^n | \bar{f}_{high,t}^n, I_{Id}, f_{low}^n, \bar{f}_{high}^{n-1}),$$

with

$$\begin{aligned} p_\theta(\bar{f}_{high,t-1}^n | \bar{f}_{high,t}^n, I_{Id}, f_{low}^n, \bar{f}_{high}^{n-1}) &:= \\ &:= \mathcal{N}(\bar{f}_{high,t-1}^n; \mu_\theta(\bar{f}_{high,t}^n, t, I_{Id}, f_{low}^n, \bar{f}_{high}^{n-1}), \sigma_\theta(\bar{f}_{high,t}^n, t, I_{Id}, f_{low}^n, \bar{f}_{high}^{n-1})^2 \mathbf{I}). \end{aligned}$$

Inputs

The choice of $(I_{Id}, f_{low}^n, \bar{f}_{high}^{n-1})$ as guidance to the denoising process is based on the information contained in each one of the images. f_{low}^n , which is the frame sampled from the low-resolution video, contains information about the performed facial expression. I_{Id} , the high-resolution input image of the same person performing a neutral expression, provides the facial identity features and background information. $\bar{f}_{high}^{n-1}$, the previously generated frame, creates temporal dependencies between frame generation, ensuring that the generated facial frames maintain temporal consistency, an essential aspect of video generation. Without this temporal dependency, frames generated independently would

exhibit temporal inconsistencies, resulting in distortions such as visible noise across the video.

Initial attempts with direct injection of the $\bar{f}_{high}^{n-1}$ yielded inferior results, as the model overly relied on specific details of f_{high}^{n-1} in the training data (real frame). During the inference phase, where the generated frame $\bar{f}_{high}^{n-1}$ is used, cumulative information loss and noise introduced by prior frame generations intensify the distortions observed in the generated frames. To handle this, we introduced noise during the training phase, deliberately obscuring facial details, creating a $(\bar{f}_{high}^{n-1} + z)$ frame where only the silhouette is distinguishable. This step eliminates dependency on $\bar{f}_{high}^{n-1}$ details, retaining only information about head, hair, and components positions. The inputs I_{Id} and $\bar{f}_{high}^{n-1} + z$ are concatenated together and combined with the noised image $\bar{f}_{high,T}^n$, and f_{low}^n is encoded using a proposed expression encoder E_{exp} and injected in the embedding space of the model. The effects of f_{low}^n injection technique, $\bar{f}_{high}^{n-1}$ and the added noise z are analyzed in section 4.4.4.

The used architecture

Following the prior works on diffusion models, we employ a conditional-UNET architecture for denoising. Noteworthy features include attention blocks for emphasizing facial features, residual blocks for facilitating information flow, and temporal embedding for considering noise levels at specific timesteps. Our model directly predicts the denoised image $\bar{f}_{high,t-1}^n$ from the input, which is four RGB images (noised $\bar{f}_{high,t}^n, I_{Id}, f_{low}^n, \bar{f}_{high}^{n-1}$), and the diffusion timestep t .

In addition, we introduce an extra encoder, E_{exp} , specifically designed to extract expression embeddings from the expression frames f_{low}^n . These features are then incorporated into the UNET embedding space via a self-attention layer. This self-attention mechanism serves as a selective information focus based on the unique characteristics of facial expressions and aims to enhance the contextual understanding of relationships within the inputs. The objective of this modification is to utilize expression information to guide the denoising process. The impact of this encoder is highlighted in our ablation study (Section 4.4.4).

Since our model predicts the denoised image $\bar{f}_{high,t-1}^n$ considering expression information, training involves minimizing the Mean Squared Error (MSE) between the predicted denoised images and their ground-truth counterparts:

$$\mathcal{L}_{MSE} = |\bar{f}_{high,t-1}^n - f_{high,t-1}^n|^2 \quad (4.1)$$

This loss function captures the overall structure and enables preserving identity details, while incorporating expression-related features through the introduced encoder.

4.3.2 Training

Algorithm 2 outlines the training procedure for a Denoising Diffusion Probabilistic Model (DDPM) designed for generating high-resolution facial images from low-resolution inputs. The process involves (1) initializing a conditional denoiser (θ) and (2) iteratively refining it through a training loop. Within each iteration, (3) samples from the dataset comprising low-resolution frames, high-resolution identity images, and previous high-resolution frames are drawn. (4) Gaussian noise is introduced, and a diffusion timestep is chosen. (5,6) The high-resolution frames undergo a diffusion process to generate $f_{high,t}^n$ and $f_{high,t-1}^n$, and (7) a denoising model (θ) use $f_{high,t}^n$ to estimate the previous high-resolution frame $f_{high,t-1}^n$. (8) The training objective involves minimizing the difference between the estimated frame and the ground truth. This iterative procedure continues until convergence (9), resulting in a denoiser capable of generating high-quality facial images.

Algorithm 2: Training Algorithm for Conditional Denoiser

Input: Low-resolution frames f_{low}^n , High-resolution identity image I_{Id} ,
High-resolution frames f_{high}^n , Previous high-resolution frames f_{high}^{n-1} ; Total steps T ;
Dataset $D = \{(f_{low}^n, I_{Id}, f_{high}^n, f_{high}^{n-1})\}_{k=1}^K$;
Output: Trained conditional denoiser θ (network with UNET and E_{exp});

```

1  $\theta \leftarrow$  randomly initialized conditional denoiser;
2 for each iteration until convergence do
3   Sample  $(f_{low}^n, I_{Id}, f_{high}^n, f_{high}^{n-1}) \sim D$ ;
4   Sample noise  $\varepsilon \sim \mathcal{N}(0, I)$  and timestep  $t \sim \text{Uniform}(1, \dots, T)$ ;
5    $f_{high,t}^n \leftarrow \sqrt{\bar{\alpha}_t} f_{high}^n + \sqrt{1 - \bar{\alpha}_t} \varepsilon$ ;
6    $f_{high,t-1}^n \leftarrow \sqrt{\bar{\alpha}_{t-1}} f_{high}^n + \sqrt{1 - \bar{\alpha}_{t-1}} \varepsilon$ ;
7    $\bar{f}_{high,t-1}^n \leftarrow \theta(f_{high,t}^n, I_{Id}, f_{low}^n, f_{high}^{n-1} + z)$  where  $z \sim \mathcal{N}(0, I)$ ;
8    $\mathcal{L}_\theta \leftarrow \|\bar{f}_{high,t-1}^n - f_{high,t-1}^n\|_2$ ;
9    $\theta \leftarrow \theta - \alpha(\partial \mathcal{L}_\theta / \partial \theta)$ ;
end
```

4.3.3 Inference

Algorithm 3 describes the frame inference process that we follow in this work. Given a low-resolution frame f_{low}^n , a high-resolution identity image I_{Id} , and the previous high-resolution frame f_{high}^{n-1} , (1) the algorithm employs a conditional denoiser θ to infer the high-resolution frame $f_{high,0}^n$. (2) The procedure involves sampling from a normal distribution and (3) iteratively applying the denoiser in reverse diffusion steps ($t = T, T-1, \dots, 1$). At each step, (4) the denoiser refines the estimate of the previous high-resolution frame $\bar{f}_{high,t-1}^n$. (6) The final output is the inferred high-resolution frame $\bar{f}_{high,0}^n$, representing a denoised version of the input frame.

Algorithm 3: Frame Inference Algorithm

Input: Low-resolution frame f_{low}^n , High-resolution identity image I_{Id} , Previous high-resolution frame f_{high}^{n-1} , Total diffusion steps T ;
Output: High-resolution frame $\bar{f}_{high,0}^n$;

- 1 Load conditional denoiser θ ;
- 2 Sample $f_{high,T}^n \sim \mathcal{N}(0, I)$;
- 3 **for** $t = T, T-1, \dots, 1$ **do**
- 4 $\bar{f}_{high,t-1}^n \leftarrow \theta(f_{high,t}^n, I_{Id}, f_{low}^n, f_{high}^{n-1} + z)$;
- 5 where $z \sim \mathcal{N}(0, I)$;
- 6 **end**
- 7 **return** $\bar{f}_{high,0}^n$;

4.3.4 Video Inference

Algorithm 4 outlines the video inference process within the framework of a Denoising Diffusion Probabilistic Model (DDPM). Given a low-resolution video $v_{low} = \{(f_{low}^n)\}_{n=1}^N$, a high-resolution identity image I_{Id} , and the total diffusion steps T , **(1)** the algorithm utilizes a conditional denoiser θ to generate a high-resolution video. The iterative procedure involves **(2)** initializing the previous high-resolution frame f_{high}^{n-1} with the identity image. For each frame f_{low}^n in the low-resolution video, **(4)** the algorithm samples a high-resolution frame $f_{high,T}^n$ from a normal distribution. Subsequently, **(5,6)** the denoiser is applied in reverse diffusion steps ($t = T, T-1, \dots, 1$) to generate denoised frames $\bar{f}_{high,0}^n$. **(8)** These denoised frames are accumulated to form the generated high-resolution video. The algorithm iterates through all frames in the low-resolution video, **(9)** updating the previous high-resolution frame at each step. This process showcases the model's capability to generate high-quality, denoised videos from low-resolution inputs.

Algorithm 4: Video Inference Algorithm

Input: Low-resolution video $v_{low} = \{f_{low}^n\}_{n=1}^N$, High-resolution identity image I_{Id} , Total diffusion steps T ;
Output: Generated high-resolution video;

- 1 Load conditional denoiser θ ;
- 2 $f_{high}^{n-1} \leftarrow I_{Id}$;
- 3 **for** $n = 1, 2, \dots, N$ **do**
- 4 Sample $f_{high,T}^n \sim \mathcal{N}(0, I)$;
- 5 **for** $t = T, T-1, \dots, 1$ **do**
- 6 $\bar{f}_{high,t-1}^n \leftarrow \theta(f_{high,t}^n, I_{Id}, f_{low}^n, f_{high}^{n-1} + z)$;
- 7 where $z \sim \mathcal{N}(0, I)$;
- 8 **end**
- 9 Add $\bar{f}_{high,0}^n$ to the generated video;
- 10 $f_{high}^{n-1} \leftarrow \bar{f}_{high,0}^n$;
- 11 **end**

4.4 Experiment

To evaluate our model, we perform an extensive evaluation on the MUG database [178], covering quantitative and qualitative analyses. We detail the experimental setup, learning process, and evaluation criteria. The model is evaluated on low-resolution videos (64x64) from the MUG dataset and on low-resolution videos generated by VideoGAN [123], ImaGINator [127], FEV-GAN [209], and by MotionGAN [136]. We compare our results with state-of-the-art models: ImaGINator [123], VDM[141], LDM [32], and LFDm[8], to highlight the improvements made by our approach.

4.4.1 Dataset

To train and evaluate our model, we used the Multimedia Understanding Group (MUG) database [178]. This database comprises videos featuring 86 individuals (52 subjects available to authorized Internet users, 25 subjects accessible upon request, and the remaining 9 subjects exclusively accessible within the MUG laboratory) expressing seven distinct emotions: happiness, sadness, surprise, anger, disgust, fear, and neutral. Each video begins and ends with a neutral expression, with the apex expression displayed in the middle. The videos consist of 50 to 160 RGB frames, each with a resolution of 896×896 pixels. For our evaluation, we used the public data of 52 subjects, specifically focusing on the first half (neutral to apex expression) of the videos representing the six basic expressions.

4.4.2 Implementation Details

For experimentation, we perform a subject-independent split, allocating 75% for training and 25% for testing. Videos are standardized to N=32 frames, and each frame is resized to (192x192) to create the high-resolution videos v_{high} . OpenFace [193] is used to extract low-resolution frames v_{low} with facial expressions against a black background, each frame resized to 64x64 pixels.

The input data consists of 9 channels, including noised image $f_{high,t}^n$, Identity image I_{Id} , and the previously generated frame $\bar{f}_{high}^{n-1}$. The model incorporates 128 hidden channels, with channel multipliers following the configuration [1, 1, 2, 2, 4, 8], indicating the number of times the hidden channels are multiplied at each level. A residual block is used at each level, and attention blocks are selectively applied at the fifth level, where embedded expression information $E_{exp}(f_{low}^n)$ is injected into the network. The diffusion process progresses over 1000 timesteps, with beta values ranging from 0.0001 to 0.02 in a linear schedule.

During training, the model is optimized for 400 epochs, reaching about 2 million steps, with a learning rate of 2×10^{-5} and a batch size of 4. Exponential Moving Average (EMA)

is utilized, with a decay rate of 0.9999. The training is conducted on a single Nvidia Titan V GPU (12GB of memory) for about 140 hours.

To evaluate our model against videos generated by low-resolution techniques, we generate facial expression videos using [123, 209, 136]. For benchmarking against state-of-the-art models, we employed public codes of ImaGINator [127] and LFDM [8], with minor modifications to enable conditionality and higher resolution. Results for VDM [141] and LDM [32] were sourced from [8].

4.4.3 Evaluation Metrics

In our quantitative assessment, we employ various metrics for a comprehensive evaluation:

- **FVD** $\downarrow$ (Frechet Video Distance): measures the dissimilarity between the distributions of features extracted from generated and real videos using a pre-trained classifier.
- **PSNR** $\uparrow$ (Peak Signal-to-Noise Ratio): Measures pixel-level similarity between generated videos and their ground truth.
- **SSIM** $\uparrow$ (Structural Similarity Index Measure): Quantifies the structural similarity between real and reconstructed videos.
- **ACD** $\downarrow$ (Average Content Distance): Evaluates facial identity consistency by calculating average pairwise L2 distances between facial features of consecutive frames. However, it focuses solely on identity consistency within the video.
- **ACD-I** $\downarrow$ (Average Content Distance - Identity): An extension of ACD, focusing preservation of the input identity in the generated video. It calculates average L2 distances between the facial features of generated video frames and the input image.

4.4.4 Experimental Results

Facial Expression Enhancement

In the process of validating the efficacy of our model in enhancing facial videos, we produce videos showcasing 13 distinct individuals, from the testing set, performing the six basic facial expressions. Our evaluation involves observing the effect of our model on videos from various models, namely VideoGAN, ImaGINator, MotionGAN, and FEVGAN. We specifically examine the quality both before and after applying enhancements by our model to understand the impact of our proposed methodologies. We note that our model is trained on the original dataset and has not learned from neither testing identities nor generated data.

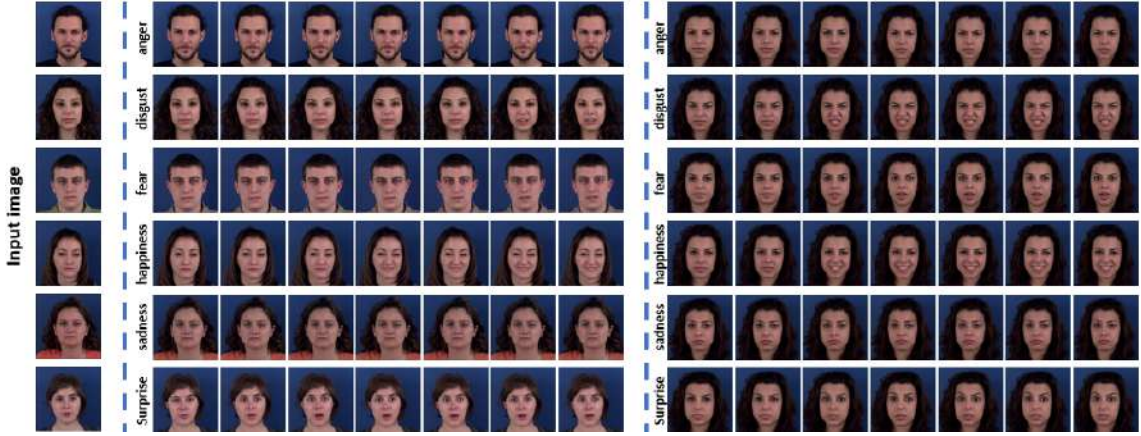


Figure 4.4: Video examples generated by our model showcasing the six basic facial expressions. The right side features generated videos of the six facial expressions performed by the same individual, while the left side presents six different individuals, with each person expressing one specific expression.

The evaluation is conducted across multiple dimensions, encompassing the quality of individual frames (PSNR and SSIM), content consistency (ACD), identity features enhancement (ACD-I), and video quality (FVD). The comprehensive quantitative and qualitative outcomes of our evaluation are presented in Fig.4.4, Fig.4.6, and Table.4.1.

Qualitative Results:

In Fig. 4.4, we present examples that demonstrate the effectiveness of our proposed model in enhancing facial expression videos. The figure showcases multiple identities expressing different facial expressions. On the left side, we present input image identities, while in the middle, we display the same identities performing one of the six basic facial expressions. Additionally, on the right side, we depict a single individual executing the six basic facial expressions. This visual representation effectively demonstrates the successful generation of expressions in high-resolution videos, highlighting the model capability to enhance facial expressions across diverse expressions and identities. Additional examples of videos generated by our model are presented in Fig.4.5.

Moreover, in Fig. 4.6, we present videos generated by low-resolution models: VideoGAN, ImaGINator, MotionGAN, and FEV-GAN, alongside their corresponding enhanced videos produced by our model. The figure visually captures the noticeable improvements in video quality, background representation, and enhanced details such as hair and clothing, contributing to an overall enhancement in visual fidelity. Beyond these enhancements, our model exhibits additional valuable properties. For instance, when VideoGAN (a) encounters challenges in identity preservation, resulting in identity loss during the generation of expressions, our enhancement (b) successfully recovers facial identity details. Furthermore, our model ensures clarity in identity representation by incorporating details such as hair and clothing. When addressing ImaGINator (c), which generates expressions that are barely noticeable, blurred, and noised, our model (d)

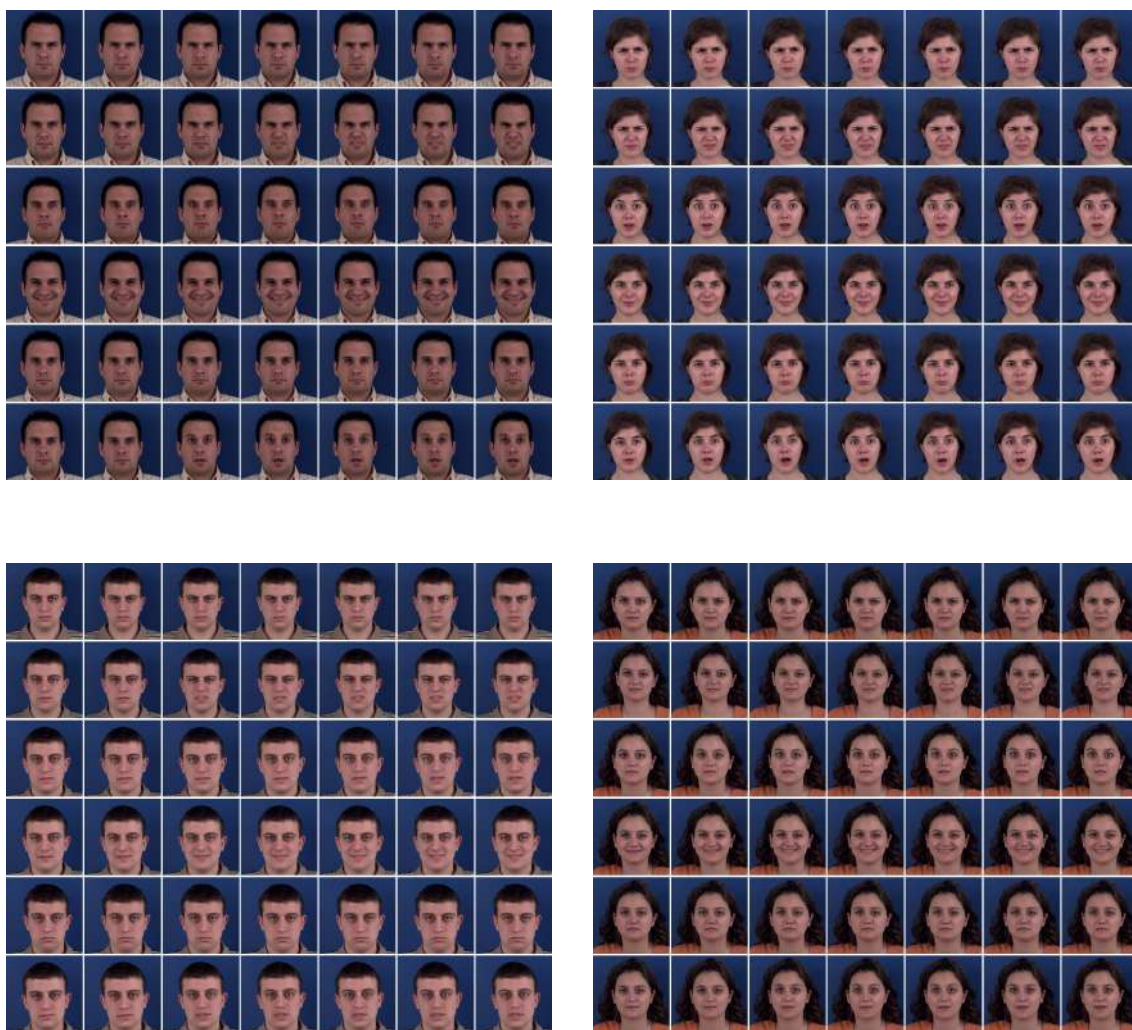


Figure 4.5: Examples of videos enhanced by our proposed model showcasing diverse identities performing the six basic facial expressions. These visual examples offer insights into the model performance and highlight its efficacy in enhancing realistic and diverse facial animations.

effectively synthesizes clear expressions while eliminating noise and blurriness. In the case of MotionGAN (e), which introduces distortions in the mouth and eyes areas, impacting the smoothness of the video, our enhancement (f) rectifies these distortions. As for FEV-GAN, which exhibits minimal distortions in low resolution, our model effectively enhances it to higher resolution.

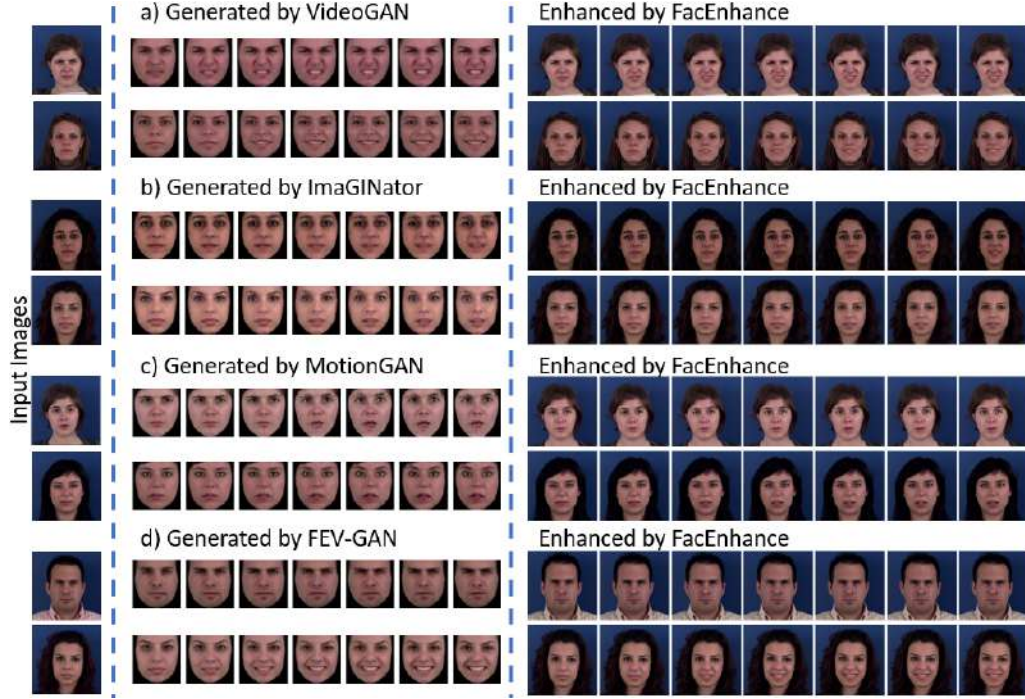


Figure 4.6: Qualitative comparison of facial expression sequences before and after enhancement using our model. On the far left of the figure, we display the input images. Adjacent to them are videos generated by VideoGAN (a), ImaGINator (b), MotionGAN (c), and FEV-GAN (d). On the right side, we present the corresponding enhanced videos by FacEnhance.

Quantitative Results:

Table 4.1 demonstrates the effectiveness of our model performance, showcasing its ability to effectively increase the resolution from (32x64x64) to (32x192x192), while improving the quality of the generated videos, as demonstrated by notable gaps in FVD, PSNR, and SSIM scores between the low-resolution generated videos and their enhanced counterparts. The model also preserves the smoothness of the expressions and the input identity as indicated in ACD and ACD-I values.

Both quantitative and qualitative analyses consistently confirm the effectiveness of our model in achieving substantial enhancements in facial expression videos. This is evident in the case of VideoGAN, which exhibits challenges in identity preservation, as reflected in ACD-I scores reaching as high as 0.73, compared to the lower scores of ImaGINator, MotionGAN, and FEV-GAN ranging from 0.2 to 0.3. However, our proposed model successfully addresses identity loss issue, achieving an ACD-I as low as 0.09. Comparable patterns are noted with ImaGINator, where the model produces high FVD 2205.26 and low

Table 4.1: Quantitative results for facial expression enhancement. The table presents results for videos generated by low-resolution models (VideoGAN, ImaGINator, MotionGAN, FEV-GAN) both before and after enhancement using our proposed model.

Model	Resolution	FVD ↓	PSNR ↑	SSIM ↑	ACD ↓	ACD-I ↓
VideoGAN	(32x64x64)	712.55	22.30	0.83	0.09	0.73
FacEnhance_{VideoGAN}	(32x192x192)	91.31	30.88	0.94	0.01	0.09
ImaGINator	(32x64x64)	2205.26	20.29	0.85	0.08	0.29
FacEnhance_{ImaGINator}	(32x192x192)	111.99	30.27	0.93	0.01	0.05
MotionGAN	(32x64x64)	654.98	23.96	0.89	0.011	0.29
FacEnhance_{MotionGAN}	(32x192x192)	167.01	31.08	0.95	0.010	0.11
FEV-GAN	(32x64x64)	452.46	24.02	0.88	0.09	0.23
FacEnhance_{FEV-GAN}	(32x192x192)	85.74	31.08	0.93	0.01	0.09

PSNR 20.09, indicating temporal and spatial difficulties and very low video quality. With the application of our model, there is a significant improvement, reflected in the decreased FVD to 111.99 and increased PSNR to 30.27. These results underscore the consistent and robust performance of our proposed model in overcoming identity preservation and quality challenges posed by existing models.

Comparison To State-of-the-art Models

In this section, we conduct a comprehensive comparison between our proposed model and state-of-the-art models for facial expression generation, including ImaGINator [127], VDM [141], LDM [32], and LFDM [8].

Qualitative Comparison

Fig. 4.7 visually compares the facial expression videos generated by our model with those produced by baseline models. The left side of the figure displays the input identity, while our generated videos (a, c, e, g) are showcased in the middle, and baseline videos (b, d, f, h) are presented on the right side. Each line in the comparison corresponds to videos generated by our model and a baseline, depicting the same identity and expression. It is important to note that the resolutions vary among the models (resolution details provided in Table 4.2). Despite these variations, both our model and the baselines exhibit robust identity preservation capabilities. Notably, ImaGINator (b), VDM (d), and LDM (f) show artifacts and blurriness in the mouth and nose area. In contrast, our proposed model consistently yields clearer expressions, improved facial details, and fewer distortions. When compared to LFDM (h), our proposed model demonstrates competitive results, highlighting its efficacy in enhancing facial expressions generated by low-resolution methods.

Quantitative Comparison

Table 4.2 presents a comprehensive quantitative comparison between our proposed

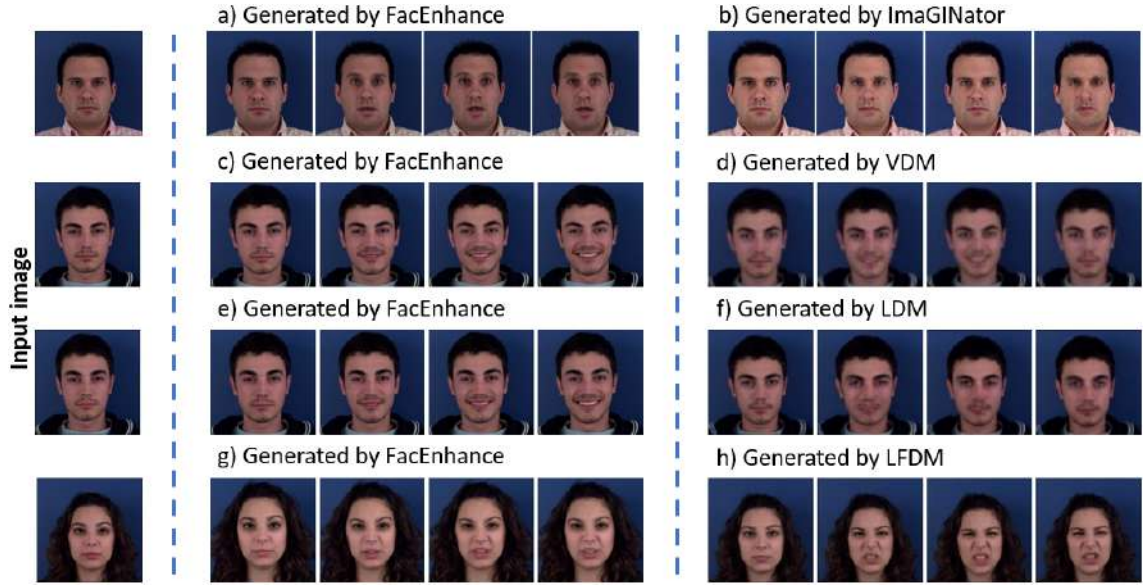


Figure 4.7: Qualitative comparison of sequences generated by the FacEnhance model and state-of-the-art models on the MUG database. The sequences of our model (a, c, e, g), ImaGINator (b), and LFDM (f) are randomly selected from the test results, and the sequences of VDM (c) and FDM (e) are taken from [8].

model and baseline methods. FacEnhance demonstrates superior reconstruction capabilities, as indicated by lower FVD scores, higher PSNR and SSIM scores. These results suggest that FacEnhance is effective in generating higher-quality videos with reduced noise and preserved overall structure throughout the entire video sequence. Even when comparing videos at higher resolutions (32x192x192) generated by FacEnhance against those produced at baseline resolutions (16x64x64 or 40x128x128), the superiority of our model remains evident. The evaluation of content consistency, as assessed by ACD, highlights the superior performance of FacEnhance compared to the baselines. This improvement is attributed to the incorporation of the previous frame $\tilde{f}_{high}^{n-1}$ in the frame generation process, enhancing the temporal context of the video and improving content consistency. In terms of identity preservation, our proposed model demonstrates a significant performance advantage over ImaGinator and LFDM, indicating its efficacy in preserving facial identities throughout the enhancement process.

Table 4.2: Quantitative comparison results of FacEnhance and baseline models.

Model	Resolution	FVD ↓	PSNR ↑	SSIM ↑	ACD ↓	ACD-I ↓
ImaGINator (WACV 2020)	(32x192x192)	415.53	22.32	0.88	0.08	0.17
VDM (ArXiv 2022)	(16x64x64)	108.02	-	-	-	-
LDM (CVPR 2022)	(40x128x128)	126.28	-	-	-	-
LFDM (CVPR 2023)	(40x128x128)	117.24	22.94	0.79	0.04	0.20
FacEnhance _{MUG}	(32x192x192)	95.38	30.88	0.94	0.01	0.07

The coherence observed in both qualitative and quantitative results emphasizes the ability of the proposed model in enhancing facial expression videos in multiple aspects.

Not only does it improve visual quality and content consistency, but it also maintains the integrity of facial identities, ensuring that generated videos remain true to the original subject throughout the sequence.

Ablation Study

In this section, our focus is on proving the significance of the principal components employed in building our model. We specifically highlight the impact of the low-resolution conditioning expression encoder E_{exp} , the use of the previous frame $\bar{f}_{high}^{n-1}$, and the addition of noise to the previous frame ($\bar{f}_{high}^{n-1} + z$). This analysis is carried out through an ablation study, involving multiple versions of our model, in which we selectively eliminate certain components to observe their effects. Three new model versions are trained for this purpose. The first version involves discarding the expression encoder E_{exp} and inputting the expression into the model concatenated with the remaining input images. In the second version, we exclude the use of the previous frame $\bar{f}_{high}^{n-1}$ in the frame generation process. In the thirist version, we train the model using the previous frame $\bar{f}_{high}^{n-1}$, but without adding noise to it ($\bar{f}_{high}^{n-1} + z$). Then, we compare the performance of these ablated versions with the full proposed model. All four networks are trained and evaluated using the same data, losses, and hyper-parameters.

Table 4.3: Performance comparison of the full model against variants that exclude the expression encoder, previous frame (t-1), and the noisy previous frame in the generation process.

Model	FVD ↓	PSNR ↑	SSIM ↑	ACD ↓	ACD-I ↓
W/o exp encoder	110.13	30.97	0.93	0.01	0.08
W/o previous frame (t-1)	126.60	31.55	0.94	0.01	0.06
W/o noisy (t-1)	721.03	22.35	0.72	0.01	0.10
Full FacEnhance	95.38	30.88	0.94	0.01	0.07

The results of our ablation study are presented in Table 4.3. Comparing our full model to the model without the expression encoder E_{exp} reveals closely matched PSNR, SSIM, ACD, and ACD-I values, suggesting comparable pixel-wise and structure-wise quality performance, as well as proficient handling of facial identity. However, the FVD metric indicates that the model without the expression encoder exhibits inferior video results, signifying a decrease in dynamic and temporal visual quality. This underscores the significant role played by the expression encoder, responsible for injecting the expression embedding into the model embedding space.

Regarding the incorporation of the previous frame in the generation process, training the model without $\bar{f}_{high}^{n-1}$ yields frames with good quality and effective facial identity handling, as suggested by the PSNR and SSIM values, but diminished temporal dynamics and overall video distribution, as proved by the FVD. Adding $\bar{f}_{high}^{n-1}$ without introducing noise results in even poorer visual and dynamic quality, as the model tends to overfit and overly depend on specific details of $\bar{f}_{high}^{n-1}$, leading to frames with accumulated noise and

distortions from the previous ones. However, with the addition of noise ($\bar{f}_{high}^{n-1} + z$), the image loses detailed features but retains the silhouette and component positions. This helps avoiding overfitting on $\bar{f}_{high}^{n-1}$, making detail generation more reliant on other images (I_{Id} and f_{low}^n). The observed disparities between the full model and the ablated versions underscore the superior performance of the full model in capturing the overall video and motion structure, affirming the significance of the components employed in the proposed model.

4.4.5 Discussion & Limitations

Our proposed model showcases robust capabilities in enhancing video content, effectively elevating resolution, improving overall quality, and incorporating complex details, including clothes, hair style, and background. One of the model’s notable strengths is its ability to preserve motion characteristics from low-resolution videos, effectively translating them into higher-quality outputs. In direct comparison with video diffusion models, our model demonstrates heightened efficiency and outperforms alternative approaches, such as LFDM, particularly in terms of content consistency.



Figure 4.8: Examples of flawed frames from videos enhanced by our model.

Despite the notable successes of our model, it is essential to address inherent limitations and challenges. Occasional failures and distortions may manifest, particularly when the input low-resolution video comprises blurry or distorted images. The challenges become evident in scenarios where the visual information in the low-resolution input is inherently ambiguous, or when the input quality deviates significantly from the training data. Figure 4.8 showcases instances of such occasional failures, highlighting the need for further refinement. Moreover, while our model demonstrates efficiency in comparison to methods such as VDM, it is crucial to recognize that the computational demands of our model remain substantial. The complexity of the underlying neural network architecture and the

high-dimensional facial data may pose constraints on real-time applicability.

As we explore future improvements, our proposed model shows promise in achieving even higher resolutions, with validation results at 512x512 resolution illustrated in Fig.4.9. However, increasing the resolution introduces exponential complexity, particularly in terms of computational load and memory requirements. This escalation presents significant challenges in optimization, generalization, and overfitting mitigation, which become more pronounced at higher resolutions. Additionally, the associated increase in training time poses a considerable obstacle.



Figure 4.9: Examples of our validation results at 512x512.

4.5 Conclusion & Perspectives

In this chapter, we introduced a novel framework for enhancing dynamic 2D facial expression videos, transforming low-resolution (64x64) facial videos into higher-resolution (192x192) outputs with improved visual quality and added background details. By leveraging low-resolution expression generation models as input for a diffusion-based enhancement process, our method demonstrated the ability to significantly improve video quality while maintaining the original motion characteristics.

Our experimental results highlight the effectiveness of the approach, proving that diffusion models can successfully elevate the visual quality of facial expressions generated by lower-resolution methods. Despite the promising results, the framework faces several limitations. One key challenge lies in its inability to generate the temporal evolution of facial expressions, requiring reliance on external models for motion generation. This reliance limits the model ability to fully capture the dynamic nature of facial expressions over time. Additionally, the computational demands of the diffusion process prevent real-time application, making it more suitable for offline processing.

Future work will focus on addressing these limitations by exploring more efficient diffusion techniques, such as Latent Diffusion Models (LDMs), to reduce computational complexity and move towards real-time capabilities. Additionally, we aim to integrate diffusion models for motion generation to capture the stochasticity of temporal evolution, leading to more diverse and realistic facial expression sequences.

Chapter 5

Latent Motion Diffusion Models for Facial Expression Synthesis

5.1 Introduction

In recent years, conditional image-to-video (cI2V) generation methods have experienced significant advancements, transforming how we generate lifelike and dynamic video content from static images. These methods exploit the potential of neural networks to synthesize realistic and versatile video content. Given a single image denoted as I_{in} and a specified condition y , the process of conditional image-to-video (cI2V) generation aims at synthesising a video $\hat{v}$ containing N frames $\hat{f}_n$, represented as $\hat{v} = \{\hat{f}_1, \hat{f}_2, \dots, \hat{f}_N\}$, resembling the target condition while retaining the identity and details of the input image.

One particularly interesting application of these methods is facial expression generation, a field that extends its applications to healthcare, education, automotive, gaming, and customer experience. While facial expression recognition has been widely researched with remarkable results over the years [193, 201, 202], the task of facial expression generation remains more complex and has been comparatively less studied. This is due to the intricate nature of human facial expressions, where even subtle changes can convey significantly different emotions. Furthermore, human sensitivity to facial dynamics means that generating convincing expressions requires high precision in modeling the nuances of facial structure, background consistency, and smooth temporal transitions, all while preserving the identity of the subject.

Several strategies have been introduced to address this challenge. A subset of research [123, 127, 209] has leveraged spatio-temporal (3D) convolutions to synthesize entire videos in a single time-step. These approaches exploit the simultaneous processing of spatial and temporal information, leading to better spatio-temporal consistency. However, they suffer from considerable computational complexity, memory requirements, and time constraints. As a result, these methods encounter challenges in capturing long-term dependencies and are constrained to generating videos at relatively low resolutions, typically limited to 32x64x64.

Alternative approaches decouple spatial content and temporal dynamics, resulting in improved synthesis of spatial details and motion coherence [128, 136, 129]. These

approaches aim to generate representations of motion dynamics (e.g., latent vectors, landmarks, motion flow), which are then integrated into to each frame independently. This strategy eliminates the need for computationally expensive spatio-temporal (3D) convolutions, but it can introduce issues related to content consistency across frames.

More recently, diffusion models (DMs) have emerged as a promising framework for video generation. DMs, initially designed for high-quality image synthesis, have been adapted to video either by generating video frames directly [141, 217] or by enhancing frame quality [139, 140]. However, this comes at the cost of a significant increase in computational complexity and inference time due to the combination of the iterative nature of diffusion models and the high resolution required for videos, particularly during the inference phase when generating videos. Alternatively, other approaches [8, 143, 144] employ diffusion models to learn the embedding space of the data and synthesis content-motion representations. However, diffusion-based methods introduce stochasticity throughout the entire latent space, affecting not only temporal dynamics but also deterministic elements of videos, such as identity information and background details. Therefore, this stochasticity introduces variability in identity features, potentially leading to the loss of input information.

In this context, we introduce a novel model, referred to as **Latent Motion Diffusion Model (LMDM)**, for conditional image-to-video generation. Our model focuses on motion features within the embedding space, incorporating stochasticity to create more varied and smoother motions while preserving the deterministic nature of subjects and background features throughout the video generation process. To achieve this, we decouple the embedding space into content and motion spaces using two distinct encoders. One encoder processes an input image containing a subject in the initial pose, and the other handles an image containing keypoints (landmarks or skeletons) and optical flow, serving as a driving image representing the motion. After learning both embedding spaces, diffusion is applied exclusively to manipulate the motion space and generate new motions. Fig.5.1 illustrates the proposed model.

We conduct quantitative and qualitative evaluations of our model on the MUG facial expression database [178] and Oulu-Casia [179] to prove the effectiveness of our model. Our evaluation includes a comparison with state-of-the-art models: ImaGINator [127], LFDM [8] and FacEnhance [140]. The results of our experiments underscore the efficacy of our approach in generating a diversity of clear smooth facial expressions and its competitiveness with the state-of-the-art models. We attempt to extend the model to human action using NATOPS dataset [218], results in this domain were more limited, indicating potential areas for future improvement.

To recap, our main contributions are as follows:

- We introduce a Conditional Image-to-Video Generation based on Diffusion Models

(DM). Our proposed LMDM separates video into content and temporal motion, allowing us to explicitly model the temporal dynamics. For this purpose, we use a latent diffusion-based approach to capture the temporal motions and align them with the initial frame.

- We investigate the impact of injecting diffusion models stochasticity into the motion space only, compared to applying it to both content and motion spaces, demonstrating improved results in maintaining subject identity and background stability.
- We extensively evaluate our model on the MUG facial expression, Oulu-Casia, and NATOPS datasets, comparing it to the recent state-of-the-art methods and showcasing our superior performance in generating diverse and lifelike facial expressions.

The remainder of this paper is organized as follows. In Section 5.2, we provide a review of the related work. Section 5.3 details our proposed Latent Motion Diffusion Model (LMDM). In Section 5.4, we present the experimental setup, followed by an in-depth analysis and discussion of the results, comparing our approach with existing state-of-the-art models. Finally, in Section 5.5, we summarize the findings of this study and suggest potential directions for future research.

5.2 Related Work

In Section 2.3, we provide an extensive review of the current state-of-the-art techniques in facial expression generation. Building upon it, this section focuses specifically on diffusion-based methods that either directly influenced our work or align closely with the objectives of our research. By examining these approaches in detail, we aim to position our work within the research literature.

Video Diffusion Model (VDM) [141] extends diffusion models to video generation. This is achieved by utilizing a denoiser network composed of 3D spatio-temporal convolutions, enabling the model to capture both spatial and temporal correlations inherent in video data. Latent Flow Diffusion Model (LFDM) [8] adopts a different approach by leveraging diffusion processes to generate sequences of optical flow maps. These flow maps are then employed to warp the latent representations of the input images, effectively guiding the content in alignment with the desired motion represented by the generated flow. FacEnhance [140] is a diffusion-based approach specifically designed to enhance low-resolution facial expression videos. It employs conditional denoising, taking as input a background-free low-resolution video of facial expressions and a single high-resolution neutral facial image of the same individual. The model generates a high-resolution video that combines the dynamic expressions from the low-resolution input with the background and identity details from the high-resolution image.

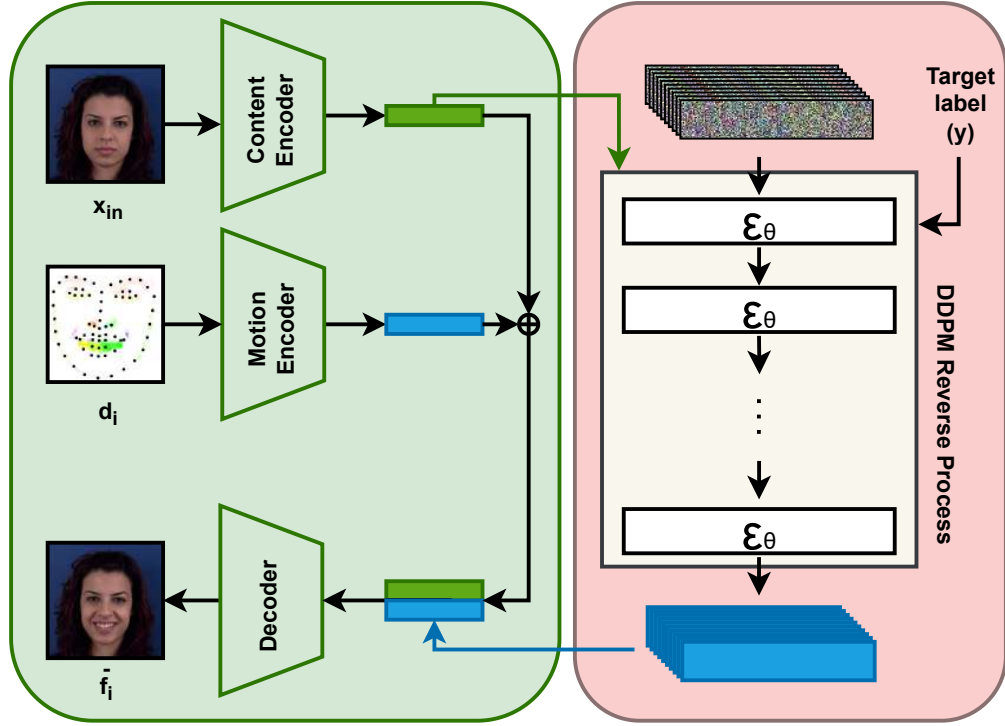


Figure 5.1: Flowchart of the LMDM framework for facial expression generation. The model consists of two stages: spatial generation (autoencoder on the left) and motion generation (diffusion on the right).

5.3 Proposed Approach: LMDM

Our proposed model tackles the challenge of conditional image-to-video generation. It takes as input a single image x_{in} (e.g., a person’s face or body) and a target motion label y (e.g., an facial expression or action class label), and generates a video $\bar{v} = [\bar{f}_1, \bar{f}_2, \dots, \bar{f}_N]$ containing the input content performing the desired motion. The objective is to develop a model, we refer to as Latent Motion Diffusion Model (LMDM), formulated as $\text{LMDM}(x_{in}, y) \rightarrow \bar{v}$.

The LMDM framework decomposes the task of video generation into two key steps: first, the learning of content and motion latent spaces, and second, the generation of motion within the learned motion latent space (Fig.5.1). To implement this approach, we first start with an unsupervised learning stage to train a conditional autoencoder (FAE). FAE is designed to take an image in an initial pose x_{in} (first frame of the video), and a driving signal dr_i (for another frame in the video) represented by an optical flow map, to reconstruct the corresponding frame $\bar{f}_i$. During the training phase, the autoencoder learns to separate the content and motion into two distinct latent spaces. Once these latent spaces are established, we introduce a diffusion model that focuses solely on learning the motion latent space. By leveraging the motion embeddings created by the autoencoder, the diffusion model is trained to produce new motions conditioned on specific expression classes y . In the inference phase, the motion encoder and driving images are replaced with

the embeddings generated by the diffusion model, allowing the system to generate realistic videos based on the input image and the target motion label.

5.3.1 Flow Conditioned Autoencoder (FAE)

In the initial stage of our approach, we employ a flow conditioned autoencoder (FAE) designed to learn the embedding spaces for both content and motion features. The FAE takes as input the initial image x_{in} and the i_{th} driving frame information dr_i , with the objective of generating the i_{th} frame $\bar{f}_i$ that resembles the target frame f_i . This process is formally represented as $\text{FAE}(x_{in}, dr_i) \rightarrow \bar{f}_i$.

To accomplish this, our autoencoder consists of three key components, as illustrated in Fig.5.2. First, a content encoder (E_c) learns the embedding z_{in} from the input reference image x_{in} , extracting the identity and spatial features. This enables subject identity and content details preservation in the generated frames. In addition, we use a motion encoder (E_m) to capture the motion embedding z_i from the driving information dr_i . Finally, a decoder network reconstructs the target image $\bar{f}_i$ by combining the embeddings $[z_{in}, z_i]$ from both the expression encoder and the identity encoder.

During training, we select two frames from the same video: the reference frame x_{in} and a randomly selected target frame f_i , along with its corresponding driving information dr_i . All images are RGB with dimensions $H_x \times W_x \times 3$. The encoders E_c and E_m process the reference frame x_{in} and the driving image dr_i to generate latent maps z_{in} and z_i , respectively, each with dimensions $H_z \times W_z \times C_z$, where C_z represents the number of channels.

The FAE is trained in an unsupervised manner using a reconstruction loss applied to unlabeled video frames. The training objective is to minimize the difference between the generated frame $\bar{f}_i$ and the actual target frame f_i , using the following loss function:

$$\mathcal{L}_{FAE} = \mathcal{L}_{perc} + \lambda_1 * \mathcal{L}_{rec}.$$

In this formulation, $\mathcal{L}_{perc}$ is the perceptual loss, implemented using a pretrained VGG network. This loss function compares high-level feature representations between the reconstructed and ground truth images, ensuring that the FAE not only reconstructs pixel-level details but also preserves the perceptual quality of the driving frame.

$$\mathcal{L}_{perc} = \|E_{VGG}(\bar{f}_i) - E_{VGG}(f_i)\|^2.$$

Additionally, $\mathcal{L}_{AE}$ is the L_2 loss measuring the pixel-wise difference between the generated frame $\bar{f}_i$ and the target frame f_i .

$$\mathcal{L}_{rec} = \|\bar{f}_i - f_i\|^2,$$

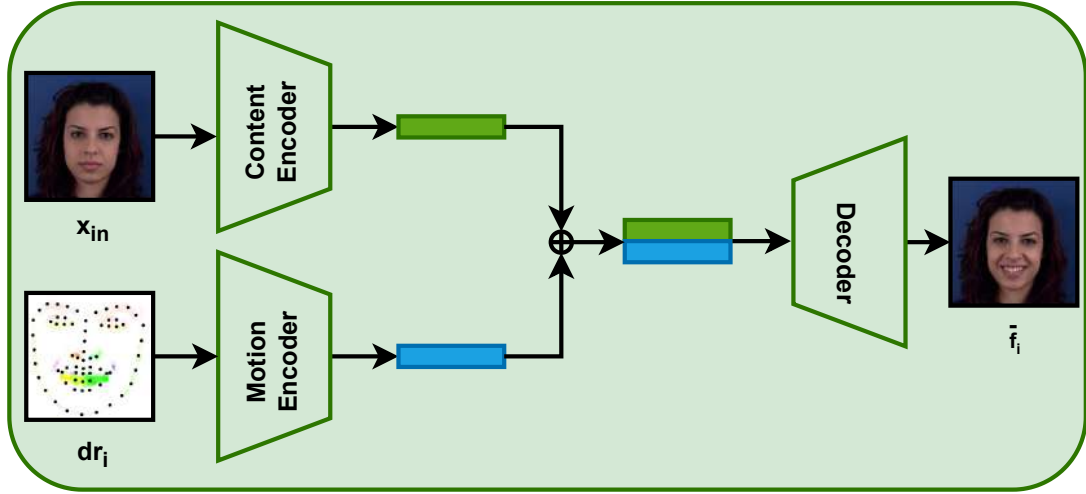


Figure 5.2: Flowchart of FAE.

By combining both perceptual and reconstruction losses, the FAE is trained to generate high-quality frames that are perceptually similar to the target frames while also minimizing pixel-wise differences. The coefficient (λ_1) balances the importance of the two loss components during training, fine-tuning the focus between perceptual quality and accurate pixel-level reconstruction.

5.3.2 Proposed Flow Diffusion Model

In the second stage of our training pipeline, we introduce a diffusion model (DM) based on a 3D-UNet architecture. This model is designed to learn the generation of motions within the motion embedding space, which was previously learned by the motion encoder E_m . The diffusion model training follows a conditional framework, where it is guided by the reference image embedding z_{in} and by the target motion label y . These inputs are crucial as they allow the model to generate motions that fit the reference image (identity/body) and align with the desired motion category. The goal is to model the conditional distribution p_θ , which enables the generation of motions that respect the given conditions.

During the diffusion process, noise is added incrementally and unconditionally to the sequence of motion embeddings $q(x_t|x_{t-1})$, where $x_0 = z_{i=1,...,N} = E_m([dr_1, dr_2, ..., dr_N])$. Here, $z_{i=1,...,N}$ represents the sequence of motion embeddings corresponding to a sequence of frames $[f_1, f_2, ..., f_N]$. The objective of the denoising model $p_\theta(x_{t-1}|x_t, z_{in}, y)$ is to predict the noise $\mathcal{E}$ that was added to the embeddings, taking into account the current noised embedding x_t , the content embedding z_{in} , and the target expression label y .

The denoising process is iteratively performed across the sequence, following the probabilistic formulation:

$$p_\theta(x_0, \dots, x_{T-1}|x_T, z_{in}, y) = \prod_{t=1}^T p_\theta(x_{t-1}|x_t, z_{in}, y),$$

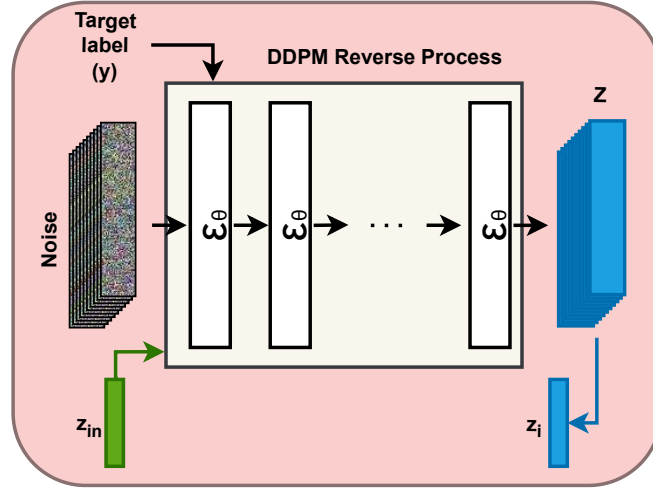


Figure 5.3: Flowchart of latent diffusion.

where each step of the denoising process is defined by:

$$p_{\theta}(x_{t-1}|x_t, z_{in}, y) = \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t, z_{in}, y, t); \sigma_{\theta}(x_t, z_{in}, y, t)^2 \mathbf{I}).$$

In this equation, μ_{θ} and σ_{θ} are the predicted mean and variance, respectively, and $\mathcal{N}$ denotes the normal distribution.

The diffusion model is trained to minimize the loss function:

$$\mathcal{L}_{DM} = \mathcal{L}_{diff} + \lambda_2 \cdot \mathcal{L}_{rec},$$

where λ_2 is a hyperparameter that balances the importance of the diffusion loss $\mathcal{L}_{diff}$ and the reconstruction loss $\mathcal{L}_{rec}$.

- **Diffusion Loss $\mathcal{L}_{diff}$:** measures the difference between the actual noise ε introduced during the forward diffusion process and the noise $\varepsilon_{\theta}(x_t, t)$ predicted by the model during the reverse denoising process. It is computed as:

$$\mathcal{L}_{diff} = \mathbb{E} [\|\varepsilon - \varepsilon_{\theta}(x_t, t)\|^2].$$

This loss ensures that the model accurately learns the noise distribution, which is critical for successful expression motion generation.

- **Reconstruction Loss $\mathcal{L}_{rec}$:** evaluates the quality of the generated videos by measuring the pixel-wise L_2 distance between the real video v and the reconstructed video $\bar{v}$ after the *Dec* has processed the output embeddings. It is expressed as:

$$\mathcal{L}_{rec} = \|v - \bar{v}\|^2.$$

The reconstruction loss ensures that the generated sequence closely matches the ground truth sequence in terms of visual quality.

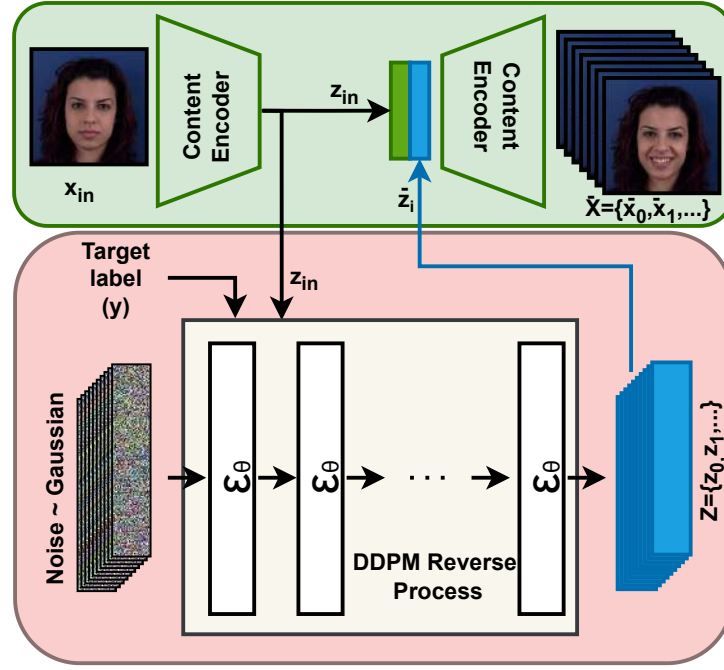


Figure 5.4: Flowchart of inference.

This two-part loss function guides the training of the diffusion model, ensuring that the generated expression motions are both temporally coherent and visually accurate, while also maintaining consistency with the subject’s identity and the desired expression.

5.3.3 Inference

In the inference phase, the driving images and motion encoder are removed, and instead, the embeddings generated by the diffusion model are directly used in the latent space of the FAE. The inference process starts with an input image x_{in} and a target expression label y . The first step involves extracting the content embedding z_{in} from the reference image x_{in} using the content encoder E_c . This embedding captures the identity and spatial features of the subject, ensuring that the generated frames maintain content consistency throughout the video.

With z_{in} obtained, the next phase is to generate a latent representation that encodes the target motion. This is achieved through the diffusion model, which takes into account both the content embedding z_{in} and the target motion label y . The diffusion model operates by sampling an initial noise vector from a normal distribution and then iteratively denoising it through a series of reverse diffusion steps. The denoiser, represented by the network θ , progressively refines the noised vector through these steps, working backward from an initial state $t = T$ to the final state $t = 1$, resulting in the final motion embedding $\bar{Z} = \{\bar{z}_{1..N}\}$ that captures the dynamic evolution of the motion across multiple frames.

The final stage involves generating the frames of the output video. Frame embeddings

$\bar{z}_i$ are then combined with the content embedding z_{in} and passed through the decoder *Dec* of FAE, which reconstructs the visual content of each frame. This process results in a sequence of frames $[\bar{f}_1, \dots, \bar{f}_N]$ that compose the final video $\bar{v}$. Fig.5.4 shows an overview of the inference process.

5.4 Experiment

To evaluate our model, we perform an extensive evaluation on the MUG [178], Oulu-Casia [179], and NATOPS [218] databases. Our evaluations cover quantitative and qualitative analyses. We detail the experimental setup, training process, and evaluation metrics. Our results are benchmarked against state-of-the-art models, including ImaGINator [127], LFDM [8], and FacEnhance [140].

5.4.1 Datasets

To evaluate our proposed model, we conduct experiments on the following datasets:

the Multimedia Understanding Group database (MUG) [178]: This database comprises 52 subjects expressing seven distinct emotions: "happiness," "sadness," "surprise," "anger," "disgust," "fear," and "neutral". Notably, each video starts and ends with a neutral expression, with the apex expression displayed in the middle. The videos consist of 50 to 160 RGB frames. For our evaluation, we split the dataset by randomly selecting 32 individuals for training, resulting in 612 videos, and the remaining 18 individuals for testing, resulting in 317 videos.

Oulu-Casia [179]: comprises 480 videos featuring 80 individuals, aged between 23 and 58, demonstrating six distinct facial expressions: "happiness," "sadness," "surprise," "anger," "disgust," and "fear." Each video sequence contains between 9 and 72 RGB frames, with a resolution of 320×240 . The sequences begin with a neutral expression and progress to the apex of the corresponding facial expression. The database is available for use by authorized researchers. We split the dataset to training and test sets, with 80:20 ratio.

The NATOPS Aircraft Handling Signal dataset [218]: includes 9600 videos where 20 subjects perform 24 distinct body-and-hand gestures used in communication with U.S. Navy pilots. This dataset includes common signals such as spreading wings and stopping. We split the dataset evenly, with 10 subjects and 4,800 videos for training, and the other 10 subjects and 4,800 videos for testing.

Data Preprocessing

All videos were resized to a resolution of 128×128 pixels to ensure consistency across our models. To extract optical flow, we used the "calcOpticalFlowFarneback" method

[219] in the OpenCV framework. For the keypoints extraction, we have employed [220] for facial landmarks and [221] for skeleton points. Since most videos in the used datasets are relatively short, we randomly sampled 40 frames from each training video rather than uniformly sampling them. The sampled frames were then ordered chronologically to create diverse, fixed-length video clips, which were used for training the second stage of our diffusion model.

5.4.2 Implementation Details

In our implementation, the Flow Autoencoder (FAE) processes two input images, each with dimensions $128 \times 128 \times 3$. The corresponding embeddings, z_i and z_{in} , are of size $32 \times 32 \times 3$. These embeddings are concatenated to form the decoder input with a size of $32 \times 32 \times 6$. The FAE is trained for 150 epochs with a learning rate of 2×10^{-4} and a batch size of 16. Training is performed on a single NVIDIA Titan V 12GB GPU for approximately 20 hours.

The Flow Diffusion Model (FDM) operates on motion embeddings of 40 frames, each $32 \times 32 \times 3$, resulting in size of $40 \times 32 \times 32 \times 3$. FDM training takes about 120K steps, also with a learning rate of 2×10^{-4} but with a smaller batch size of 4. To ensure training stability, we utilize an Exponential Moving Average (EMA) with a decay rate of 0.9999. The FDM training is also performed on a single NVIDIA Titan V 12GB GPU and takes approximately 60 hours.

For benchmarking, we employed the public implementations of ImaGINator [127], LFDM [8], and FacEnhance [140].

5.4.3 Evaluation Metrics

For our quantitative analysis, we utilize several key metrics to ensure a thorough evaluation:

FVD $\downarrow$ (Frechet Video Distance): FVD measures the difference between the distributions of features extracted from generated and real videos using a pre-trained classifier, capturing both temporal and spatial inconsistencies.

ACD $\downarrow$ (Average Content Distance): This metric assesses the consistency of content within a video by calculating the average pairwise L_2 distances between facial features across consecutive frames, emphasizing identity preservation within the video.

ACD-I $\downarrow$: ACD-I extends ACD by specifically focusing on the preservation of the input identity in the generated video. It computes the average L_2 distances between the facial features of the generated video frames and those of the input image.

5.4.4 Experimental Results

Qualitative Results

Fig.5.5 presents examples of videos generated by our proposed model, each of which displays one of the six basic facial expressions for different identities. The figure illustrates the results from our model trained on the MUG dataset (left) and the Oulu-CASIA dataset (right). Notably, the data used during both the training and testing phases is subject-independent. The resulting videos show continuous, smooth, and natural transitions between facial expressions, demonstrating that the expressions can be effectively controlled based on the input labels. The generated videos maintain the same details from the input image, such identity, beards, hair style, and clothes. Our model performs slightly better on the MUG dataset compared to the Oulu-CASIA dataset. This is due to the fact that Oulu-Casia is smaller, yet it shows more variety in skin colors, lighting conditions, and accessories (glasses) on the face, which increases the difficulty of pattern learning for the model. These results confirm the effectiveness of our model in generating realistic facial expressions.

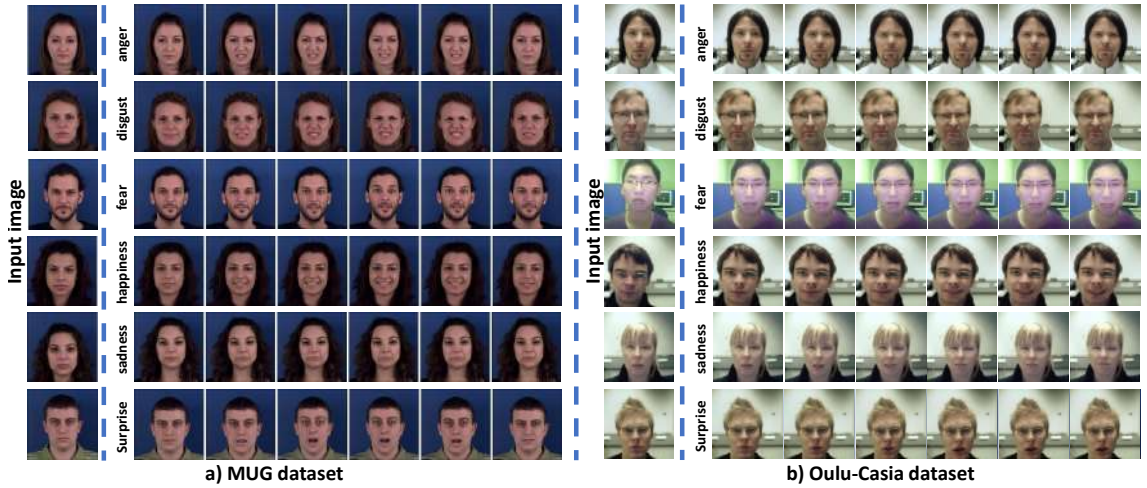


Figure 5.5: Examples of videos generated by our model showcasing the six basic facial expressions, given the input identity and the target expression, on MUG and Oulu-Casia datasets.

Figure 5.6 showcases a comparison between the results of our model and state-of-the-art approaches on the MUG database. It presents sequences generated by our model (on the left side) and by state-of-the-art models (on the right side) of the same person performing the same expression (in each line). In comparison with ImaGINator (a, b), our model shows smoother and more natural facial expressions, with significantly fewer artifacts. Alternatively, ImaGINator displays limited expression evolution, often generating a single expression level throughout the video, with noticeable distortions. When compared to LFDm (c, d), our model demonstrates competitive performance, specifically in terms of expression naturalness and overall visual quality. In (e, f), FacEnhance displays better

visual quality. However, this model does not focus on learning or generating expression motion, but rather it enhances the resolution and quality of existing facial expression sequences, either captured or generated by other models. Despite FacEnhance higher visual performance, it lacks the ability to synthesize expression dynamics from scratch. This analysis underscores the competitiveness of our proposed approach, particularly in its ability to generate realistic, smooth, and artifact-free facial expressions compared to existing state-of-the-art models.

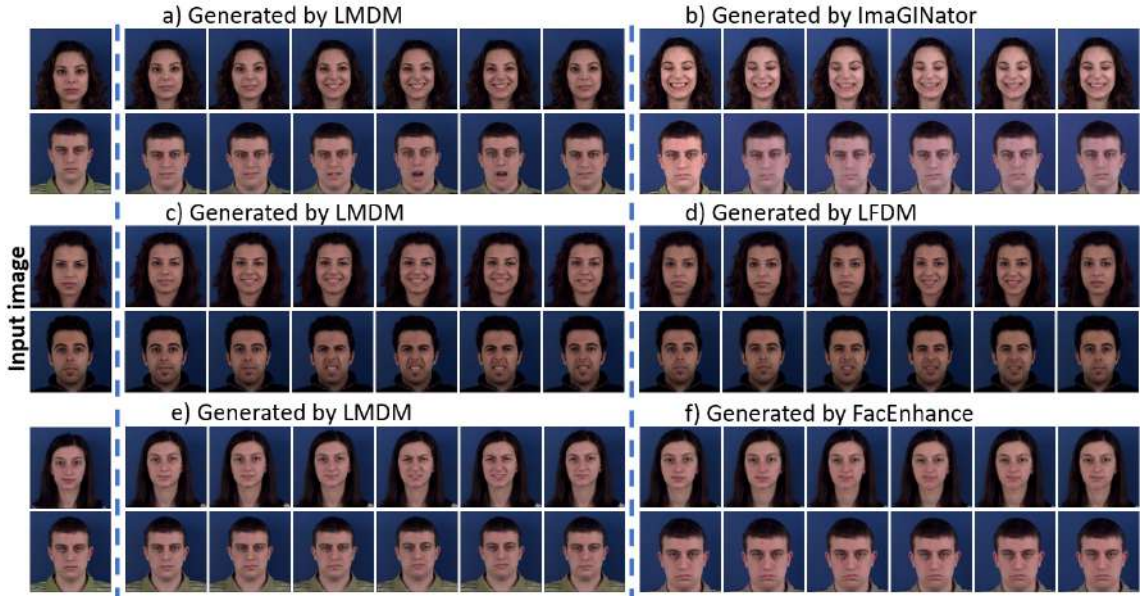


Figure 5.6: Qualitative comparison of sequences generated by the LMDM model and the baselines on the MUG database.

Quantitative Results

In our quantitative analysis, we generate over 300 videos using 18 different test subjects, each displaying the six basic facial expressions. To evaluate these generated videos, we compute three key metrics: FVD, ACD, and ACD-I. Table 5.1 presents a comparison summary of our proposed model (LMDM) against the ground truth and the baseline models.

Table 5.1: Quantitative comparison results of LMDM and baseline models using FVD, ACD, and ACD-I metrics.

Model	Resolution	FVD ↓	ACD ↓	ACD-I ↓
ImaGINator (WACV 2020)	(32x128x128)	349.44	0.13	0.26
LFDM (CVPR 2023)	(40x128x128)	354.83	0.03	0.22
FacEnhance (Arxiv 2024)	(32x192x192)	95.38	0.01	0.14
Our model	(40x128x128)	269.03	0.04	0.25
Ground Truth	(40x128x128)	0.0	0.02	0.17

The results show that compared to the ground truth, LMDM achieves close values, indicating its ability to generate realistic and natural facial expressions. Compared to ImaGINator, LMDM generates videos with greater similarity to real data and outperforms ImaGINator in terms of content consistency and identity preservation, as reflected by lower FVD, ACD, and ACD-I values. In comparison to LFDM, while LMDM demonstrates competitive performance in ACD and ACD-I, shows much better capabilities in generating videos that closely resemble real data, as proved by the the lower FVD score. As for FacEnhance, Although it achieves the best visual quality and the lowest FVD, it is specifically designed to enhance the resolution and quality of facial images rather than to learn or generate the motion of facial expressions. These results highlight the robust performance of LMDM in producing realistic and consistent facial expression sequences.

Variance Test

In this study, we highlight the model’s capability to generate diverse outputs even when conditioned on the same expression label, enhancing both the naturalness of the generated motions. By avoiding repetitive expressions across different faces and instances, the model adds variability to the outputs. Figure 5.7 showcases examples where the same input face and expression label lead to different output videos, each displaying variations of the same expression. These variations include differences in temporal evolution and expression intensity, indicating that the model captures a broader spectrum of emotional expressions.

Additionally, the different temporal evolutions in the generated videos demonstrate the model’s ability to simulate the dynamic nature of facial expressions over time. Some sequences show a gradual intensification of the expression, while others display quicker transitions. This variability is critical for producing realistic and refined outputs, as it mirrors how emotions are expressed differently by various individuals and even by the same individual in different contexts. By capturing these details, the model enhances the lifelike quality of the animations, contributing to more engaging and believable interactions in virtual environments.

5.4.5 Attempt for Extension to Human Action Generation

In this section, we discuss our ongoing efforts to extend the proposed model to human action generation. To achieve this, we use the NATOPS dataset [218]. This dataset includes complex sequences of human actions of more classes compared to facial expressions, which present additional challenges compared to facial expressions. The goal is to assess the generalizability of the model when applied to this new domain.

We initially applied the model with the same parameters used for facial expression generation. However, the preliminary results were significantly insufficient. Notably, key body parts that were expected to move disappeared entirely from the generated outputs

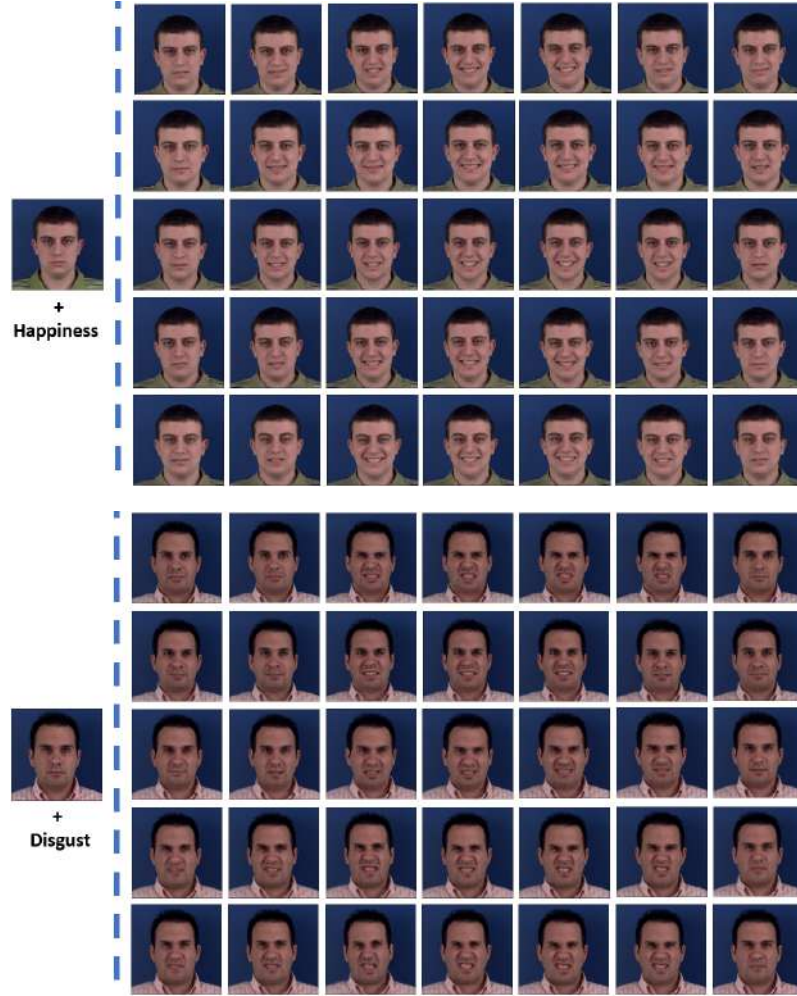


Figure 5.7: Examples of different generated videos by our model using the same input image and expression label.

(Fig.5.8.a). Despite attempts at tuning hyperparameters, such as adjusting the learning rate, no substantial improvement was observed. In some cases, the quality of the results further deteriorated.

Upon investigation, we identified two potential reasons for the this poor performance:

- **Task Complexity:** The complexity of human action generation could be exceeding that of facial expression generation. The current model architecture may be insufficiently expressive to handle this task, particularly given our current computational limitations. These restrictions prevent us from scaling the model to the desired level of complexity.
- **Model Robustness:** The Flow Conditioned Autoencoder (FAE) appears to lack robustness, particularly in handling errors in the latent space. This issue manifests more visibly in human action generation than in facial expression generation. Specifically, in facial expressions, the spatial relationships between components (e.g., eyes, mouth) remain consistent, whereas in human actions, misalignments in the

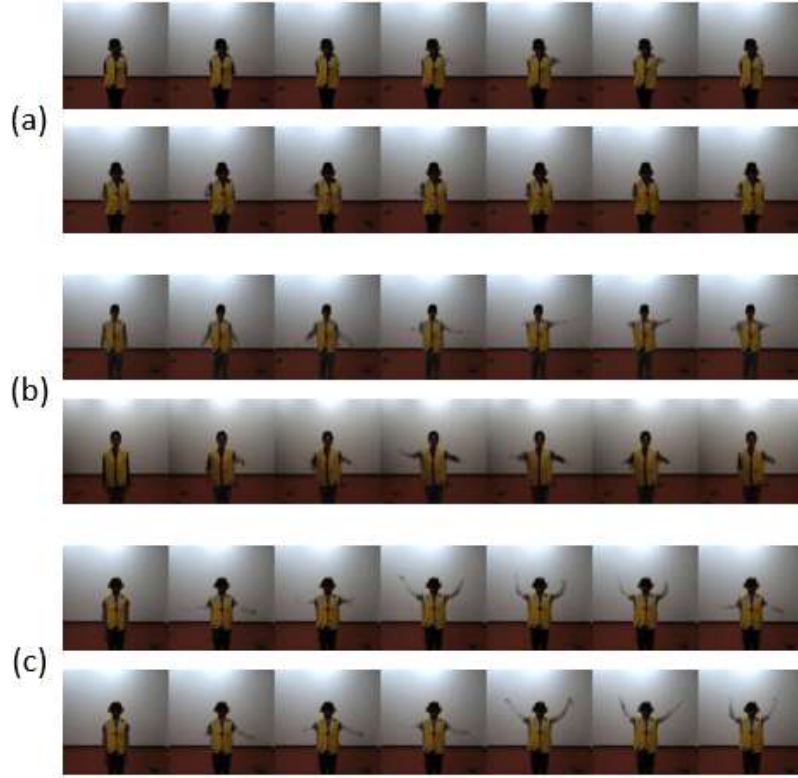


Figure 5.8: Examples of sequences generated using LMDM, during our work on human action generation.

latent space result in the dislocation of entire body parts. This inconsistency highlights the model’s sensitivity when tasked with generating full-body movements.

To address the robustness issue, we introduced noise during the training process to make the FAE more resilient to small errors in the latent space. While this modification led to minimal improvements, allowing the model to begin generating some motion, the moving components—such as arms—remained barely visible (Fig.5.8.b). Although this adjustment did not fully resolve the problem, it indicated that increasing the model’s robustness may be a viable path forward.

We also experimented with removing the skip connections from the network architecture. This modification reduced the quality of the generated frames, as expected, but it simultaneously decreased the model’s dependency on the input image. Interestingly, this helped the model generate more coherent actions, although the visibility of moving components remained insufficient (Fig.5.8.c).

Our current work is focused on improving the model further through several strategies. First, we plan to replace the autoencoder with a Variational Autoencoder (VAE). Unlike standard autoencoders, VAEs introduce a probabilistic structure to the latent space. By enforcing a Gaussian distribution through the addition of a regularization term (typically the Kullback-Leibler divergence), the latent space becomes continuous. This characteristic

allows the VAE to generate more realistic data by sampling from this regularized latent space, potentially addressing the robustness issues we encountered with the FAE.

Furthermore, we aim to increase the complexity of the diffusion models employed in the training process. By enhancing the model’s capacity to handle high-dimensional action sequences, we anticipate that it will become better equipped to generate coherent, complex human actions. This extension of the model’s capability represents a critical next step in our research trajectory.

This work remains in progress, and we continue to explore these and other approaches to improve the model effectiveness and robustness in human action generation.

5.4.6 Discussion & Limitations

The quantitative and qualitative evaluation of our model demonstrates competitive performance when compared to state-of-the-art methods. By operating in the latent space, our model shows increased efficiency relative to approaches that process pixel-level data directly. While LFDM also operates in the latent space, our approach simplifies the process by applying a straightforward diffusion mechanism within the motion space, which has shown promising potential, in contrast to LFDM, which requires multiple steps such as generating flow sequences, occlusion maps, and warping embeddings. Our method reduces complexity while still achieving competitive results in initial tests. Moreover, the proposed model applies diffusion specifically in the motion embedding space, introducing stochasticity into the motion while preserving the discriminative nature of identity features. This leads to more diverse generated motions while maintaining the same input identity.

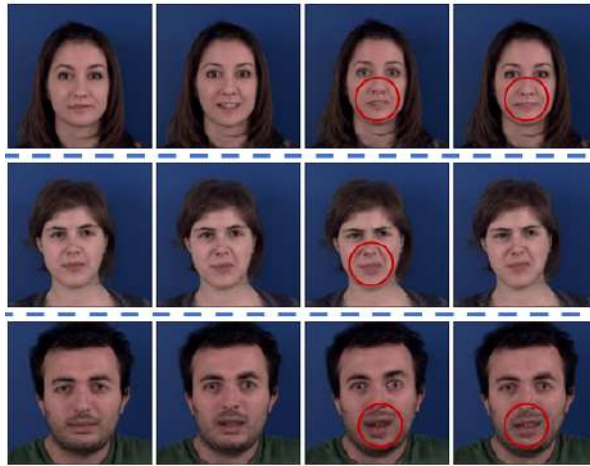


Figure 5.9: Examples of video generation with noticeable flaws and inaccuracies.

Despite these promising outcomes, our LMDM model struggles with several limitations. Both components of the model must demonstrate high performance and robustness, otherwise, errors in either part can result in inaccuracies or failures in the generation

process. Another limitation is that, **on occasion**, the generated actions do not align well with the given face, resulting in unnatural movements. Fig.5.9 displays some instances of flawed video generation that highlight these issues. For instance, the model occasionally produces distorted frames, with noticeable defects around the mouth or eyes. To address this, we plan to enhance the AE by incorporating adversarial training, which could improve the model’s ability to accurately learn and represent facial structures. Furthermore, applying our model to more complex tasks, such as generating human actions with a larger number of classes than facial expressions, will require increasing the model’s complexity. This, in turn, demands more resources and significantly complicates optimization.

5.5 Conclusion & Perspectives

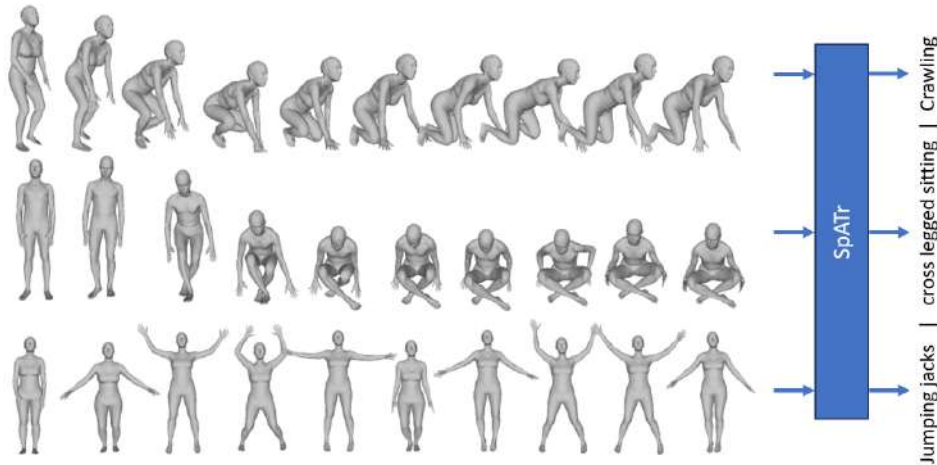
In this work, we introduced a novel Latent-Motion Diffusion Model (LMDM) for dynamic facial expression generation, using a single neutral image and a target facial motion category as input. By injecting diffusion-based stochasticity into the motion embedding space, our model generates diverse expressions while preserving the fine details of the input identity. This is achieved by decoupling the identity and motion embedding spaces, then applying diffusion solely to learn the motion latent space.

We evaluated LMDM on the MUG and Oulu-Casia databases, using both qualitative and quantitative assessments. The results highlight competitiveness of the proposed model with state-of-the-art methods, showcasing high visual quality and significant variance in the generated expressions. We also explored the initial steps toward extending this approach to human action generation. While the proposed model shows potential, it is important to note that the model exhibits certain shortcomings. These limitations include occasional artifacts and errors in the generated videos, difficulties in handling complex tasks such as human action generation, and high computational demands despite efforts to improve efficiency.

Future work will focus on improving the autoencoder’s robustness and ability to better capture and represent facial structures, addressing the observed limitations. Furthermore, we aim to continue our attempt to expand the model beyond facial expressions to cover human action generation. Given the increasing importance of the 3D domain in computer vision and its wide range of applications, we also plan to extend our work into the 3D space, opening up new avenues for advanced motion synthesis (Chapter 7).

Chapter 6

MoCap 3D Human Action Recognition Based on Spiral Auto-encoder and Transformer Network



6.1 Introduction

The field of 3D Human Action Recognition, including facial expression recognition (3D FER) and body action recognition (3D HAR), has rapidly gained attention due to advancements in 3D technologies. These technologies hold significant potential across various applications due to the robustness of 3D data against variations in shadowing conditions, lighting, and pose, as well as their capacity to offer precise geometric representations of the human body. Despite these advantages, 3D FER/HAR poses numerous challenges, including: **1)** the complexity of actions, which involves variations in motion, speed, timing, and spatial configuration; **2)** the large variability in action classes, which complicates model generalization; **3)** the difficulty of learning effective and discriminative representations that capture both spatial and temporal dynamics in 3D data.

Several strategies have been developed to address these challenges. One approach focuses on encoding the spatial and/or temporal information of the 3D video using hand-designed feature descriptors, then using the extracted codes to train machine learning classifiers, such as HMMs, SVMs, or Random Forests, to label the expressions. The most popular examples of this group are based on descriptors that compute: distances between extracted keypoints [68, 69, 70], distances between segments of level curves [71, 222], LBP-TOP from consecutive frames [223], Free-Form Deformations (FFD) [73, 74], local occupancy patterns and Fourier temporal pyramid [224], Deformation Vector Field (DVF) [225], Dense Scalar Field (DSF) [76, 226], Nebula features [78], covariance matrices [227], three-dimensional discrete cosine transform (3D-DCT)[79], histograms of 3D gradient (HOG3D) [80]. Feature descriptor-based methods have demonstrated promising outcomes when applied to limited datasets. However, they suffer from several limitations. Intense feature selection is required to determine the most relevant features for the task at hand. Additionally, scalability poses a challenge when confronted with variations in data size, distribution, or classes, often demanding human intervention to adapt the feature descriptors. These methods also tend to perform well only in subject-dependent scenarios or when recognizing a small set of expressions, typically constrained to brief sequences of 6 to 12 frames.

To overcome these limitations, another group of methods, projection-based methods, adopts 2D deep learning to tackle 3D FER/HAR. In this approach, the 3D face scenes are projected onto 2D to create RGB views [82], depth maps [228, 229], geometrical images [83, 85, 230], fusion of texture and geometric information [84], and other forms of input [86, 87]. Although effective, these methods suffer from projection-related drawbacks, including the time-consumption and geometric information loss during projection from 3D to 2D domain. Keypoints-based (Skeleton/Facial Landmarks) methods [231, 232, 233, 234, 235, 236] represent human actions using keypoints positions and rotations extracted from depth or motion sensors. However, these methods suffer from the loss of geometrical information. Point cloud-based methods [237, 238, 239, 240, 241], on the other hand, utilize 3D point cloud data to model the human face or body as a collection of points, thereby capturing spatial and motion information for action recognition. Point cloud-based methods face challenges in regularizing the irregular structure of the human body, spatio-temporal information loss, difficulty in recognizing small and close movements, and most importantly, the substantial complexity and cost associated with these models.

Recognizing the potential of motion capture (MoCap) fixed-topology data recognition in various applications, such as human-computer interaction, sports analysis, animation, gaming, we emphasize the importance of developing models specific to fixed-topology MoCap data. Unlike depth data, MoCap data lacks many of the challenges associated with geometric information, spatio-temporal complexities, and irregular mesh structures. Therefore, it does not necessitate resource-heavy models but rather demands a tailored approach to its unique characteristics. To the best of our knowledge, STMT [242] is the only state-of-the-art method that addresses action recognition on 3D MoCap by directly

using 3D meshes, without converting them to keypoints, 2D projections, or point clouds. STMT uses a transformer that employs two types of attention mechanisms to capture both spatial and temporal context. However, using transformers for both spatial and temporal information can be computationally expensive and may require large training datasets, which are not always available.

In response, we propose a lightweight 3D FER/HAR model specifically designed for MoCap fixed-topology data. This model eliminates unnecessary complexity and minimizes resource consumption while ensuring efficient and accurate action recognition. While our primary emphasis in the proposed model is on fixed-template meshes within Motion Capture (MoCap) data, it is noteworthy that the model retains the capability to handle non-fixed data after fitting them into a template, employing methods like FLAME [39], SMPL [243], or SCAPE [244].

Our proposed model, namely SpATr (Spiral Auto-encoder + Transformer Network), is a fully learned model that directly utilizes mesh data to model the dynamic changes for 3D action recognition. We decouple spatial and temporal information by first extracting spatial features from each frame using a spiral convolutional auto-encoder [7]. The sequence of extracted embeddings is then used to train a transformer network [33] to model the temporal context between frames (inter-frame attention) and classify the 4D motion. The spiral auto-encoder produces task-free spatial embeddings, rich in geometrical information extracted from face/body meshes. These flexible embeddings are suitable for diverse classification tasks such as facial expression recognition, action recognition, identity recognition, anomaly detection, and pose estimation. Their task-free nature eliminates the need for specific feature engineering across scenarios, ensuring resource savings and efficiency. The transformer architecture demonstrates remarkable effectiveness in parallel processing, handling sequences concurrently. It captures long-range dependencies and efficiently manages long sequences with minimal memory increments, easily reaching thousands of frames. Additionally, it produces context-aware temporal representations, enhancing its effectiveness in capturing dynamic temporal patterns. Fig. 6.1 illustrates the model overview.

In summary, our contributions include the following aspects:

1. We propose a fully learned model for 3D FER/HAR directly from fixed-topology 3D meshes.
2. We investigate the impact of decoupling the 3D FER/HAR process into two stages: a spiral auto-encoder for spatial feature extraction and a transformer for temporal sequence modeling.
3. Our model is efficient, lightweight, easily optimized, and scalable for larger sequences, in contrast to most state-of-the-art models.
4. The model can be utilized for various tasks, including facial expression recognition,

human action recognition, face recognition, and sex recognition, without the need to retraining the feature extractor. This capability maximizes resource efficiency and enables different predictions using minimal resources.

5. We extensively evaluate the proposed model on two benchmark datasets for facial expression recognition (MUG [178] and BU-4DFE [186]) and three benchmarks for human action recognition (Babel [245], MoVi [246], and BMLrub [247] from the AMASS archive [248]), comparing our results with the recent 3D HAR state-of-the-art approaches.

The remainder of this chapter is organized in the following manner. Section 6.2 presents the related work for FER and HAR. Section 6.3 describes the proposed method and motivations for using spiral convolutions and transformers. The experimental results analysis on facial expression recognition is presented in section 6.4, and on human action recognition is in section 6.5. Discussion of the advantages and limitations of the model is in section 6.6. And finally, the conclusion of this work is given in section 6.7.

6.2 Related Work

In this work, we evaluate our proposed model across two distinct tasks: 3D facial expression recognition (FER) and 3D human action recognition (HAR).

6.2.1 3D Facial Expression Recognition

A detailed discussion of the related work in the field of 3D facial expression recognition is presented in Section 2.3.

6.2.2 3D Human Action Recognition

Numerous methods have been proposed to address 3D HAR. These methods involve projection-based approaches, that project 3D information (depth, normal, curvature maps, etc.) to the 2D domain. They then combine this information with 2D views of the object and feed it into classification deep neural networks, typically convolutional networks [249, 228, 230]. However, such methods suffer from geometrical information loss during the 2D projection process. Other methods focus on computing hand-crafted 3D/4D features directly from the static/dynamic 3D body scans. Then, they utilize dimensional reduction techniques (e.g., PCA, LDA, NCA) and classic Machine Learning algorithms (e.g., HMMs, SVMs, multi-class Random Forest) for recognition. The most popular examples of this group of methods include those based on local occupancy patterns and Fourier temporal pyramid [224], covariance matrices [227], rotations and translations between body parts/joints [222]. Although hand-designed feature descriptors have shown

impressive results and offer compact meaningful insights, they often require intense feature selection algorithms to determine the most relevant features for each task. Moreover, their scalability is limited when dealing with different and larger datasets, increased numbers of action classes, and varying data distributions. Additionally, these methods are primarily designed for short scenes.

Alternatively, some researchers have tackled these limitations by utilizing deep learning architectures like RNNs [250, 251, 252], temporal CNNs [253, 254, 255, 256], and GCNs [257, 258, 232] to model spatial and temporal representations of skeleton sequences. However, these approaches do not fully leverage the available data, as they rely on skeleton representations that lack crucial information, such as topology information, surface motion, and body shape information.

To overcome these limitations, geometric deep learning approaches have emerged, focusing on non-Euclidean data, such as point clouds, graphs, and manifold data (meshes). For instance, 3DV-PointNet++ [259] employ temporal rank pooling [260] on a set of depth frames to generate a 3D dynamic voxel (3DV) that represents both spatial structure and motion patterns of the sequence. A set of points is then sampled from the 3DV and input into a PointNet++ [261] for 3D HAR. However, forcing a regular voxelized structure on the 3D human body leads to the loss of smoothness (information loss). Additionally, transforming the 3D sequence into a single static point cloud increases computational costs, causes spatio-temporal information loss, and increases the difficulty of recognizing small/close motions. In [262], the authors address the difficulty of computing accurate point trajectories in point cloud videos, especially colorless point clouds. The proposed model avoids point tracking by employing a point 4D convolution that extracts local patches from each point cloud frame, and a transformer that uses self-attention on the patches to capture position, appearance, and motion information. The same authors also propose PSTNet [241], where spatial and temporal information is disentangled to model point cloud sequences. PSTNet applies point-based spatial convolution first to extract spatial encoding, followed by temporal convolution to capture temporal information. In addition, due to the difficulty of performing convolution across frames, the point tube technique is proposed to preserve the spatio-temporal local structure. In SequentialPointNet [240], *Li et al.* propose dividing the main modeling operations into frame-level units executed in parallel. Each of the frames is flattened into a hyperpoint that contains the spatial information. A Hyperpoint-Mixer module is then used to merge the spatial intra-frame and temporal inter-frame features of all hyperpoints to obtain the spatio-temporal representation for HAR. Point cloud methods fail to exploit the connections between vertices, which can offer valuable information. The primary challenge of these methods is the large memory consumption that imposes downsampling the object into a minimal number of points and the sequence into a limited number of frames. STMT, proposed in [242], directly employs 3D MoCap mesh data for 3D human action recognition (HAR), without sampling to skeleton, point cloud, or 2D projection. STMT utilizes a hierarchical transformer architecture featuring two types of attention

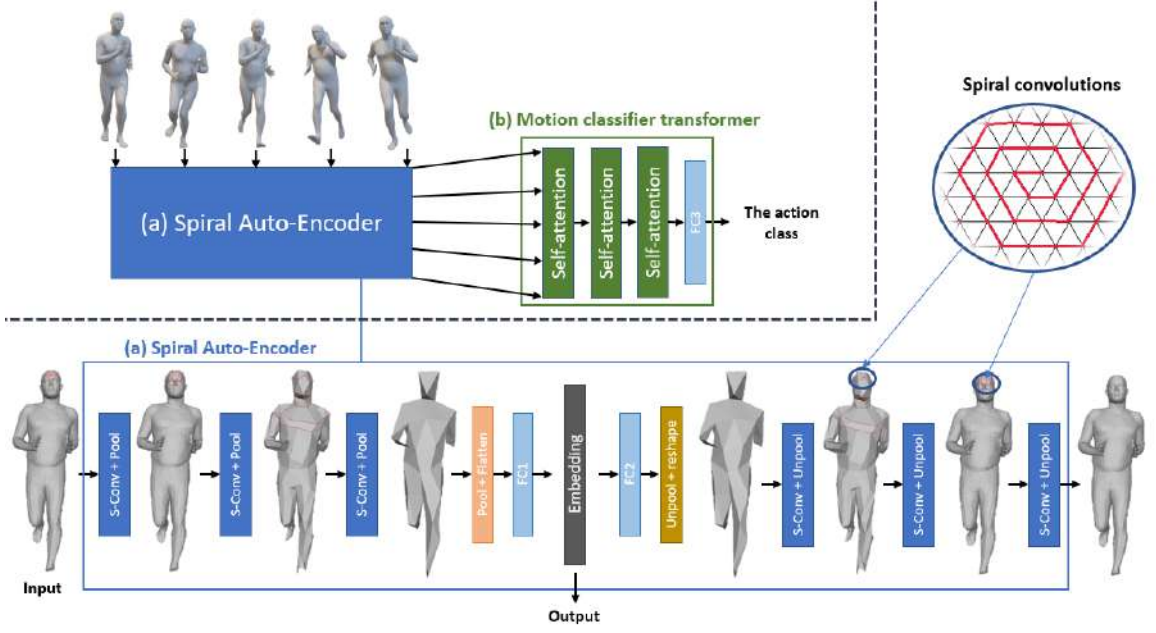


Figure 6.1: Overview of the proposed model for 3D human action recognition. Firstly, each mesh from the action sequence is mapped through the spiral auto-encoder, which extracts its embedding. Next, the sequence of embeddings is utilized by the temporal transformer, which employs self-attention to capture temporal dependencies, enabling the prediction of the action class.

mechanisms. The first, intra-frame off-set attention, focuses on spatial information within individual frames. The second, inter-frame self-attention, targets temporal interactions between frames. This dual attention enables grasping the context between any two vertex patches in the spatio-temporal domain.

6.3 Proposed Approach

In this section, we present the overall design of our proposed model, SpATr, which is aimed at recognizing actions from sequences of 3D motion capture (MoCap) meshes. Our model takes a sequence of MoCap meshes as input and predicts the corresponding action class. Specifically, we formulate our goal as learning a function $\text{SpATr}(S_{eq}) \rightarrow \bar{Y}$, where $S_{eq} = [M_1, M_2, \dots, M_L]$ is the input sequence of length L , containing meshes M_i of the same topology, and $\bar{Y}$ represents the predicted label.

The model consists of two key components: a spiral auto-encoder (SpAE) and a temporal transformer classifier (T). The SpAE takes in the meshes M_i of the sequence S_{eq} individually and uses spiral convolutions [7], to produce task-free spatial embeddings, $\text{SpAE}(M_i) \rightarrow E_i$. The transformer T utilizes self-attention on the sequence of extracted embeddings to learn modeling the temporal evolution of the sequence and recognizing the action, $T(\text{SpAE}(M_1), \text{SpAE}(M_2), \dots, \text{SpAE}(M_L)) = T(E_1, E_2, \dots, E_L) \rightarrow \bar{Y}$. The overview of our approach is illustrated in Fig. 6.1.

6.3.1 Spiral Auto-Encoder

The primary objective of the spiral auto-encoder (SpAE) lies in finding a mapping denoted as $\text{SpAE} : M_i \in (R^{N \times 3}) \rightarrow E_i \in (R^C)$, such that $M_i \approx \bar{M}_i$. Conventional techniques for acquiring shape embeddings often utilize linear functions such as PCA or LDA, which causes difficulties in capturing extreme and non-linear deformations. This necessitates the use of deep learning techniques tailored specifically for 3D data. However, developing convolution-like techniques on graphs and manifolds is challenging due to the absence of a comprehensive global coordinate system capable of effectively representing points and their ordering. Earlier intrinsic mesh techniques, such as Geodesic CNN [40] and Anisotropic CNN [41], achieve convolution on mesh data through local systems of coordinates centered around individual vertices. Nonetheless, they lack global context, demand high computational complexity and parameter count, can present optimization challenges, and often require manual feature engineering and pre-computation of local systems of coordinates.

We suggest using spiral convolutions [7], spiral convolutions are detailed in section 1.3.6, as a building block of a fully differentiable encoder that learns hierarchical representations E_i of given meshes M_i , and a decoder that learns to reconstruct the meshes $\bar{M}_i$ from the representations E_i . Spiral convolutions are advantageous in taking into account the connectivity of the mesh graph to design global (across the entire mesh) and local (within patches) ordering of the vertices. This enables local processing while capturing the global context for each shape. In addition, spiral filters benefit from weight sharing, which reduces the number of parameters, resulting in a lightweight and easy-to-optimize network.

The encoder takes a single mesh M_i of N vertices as input. M_i undergoes four layers of spiral convolution, each followed by downsampling. The mesh is then encoded into a C -dimensional latent space using a fully connected layer (FC1) to obtain the embedding E_i . This embedding is used as input to the decoder, where it passes through a fully connected layer (FC2) and four spiral convolutional layers, each followed by upsampling. The decoder aims to reverse the effect of the encoder and reconstruct the initial mesh $\bar{M}_i$.

The network is trained to minimize the L1 distance between the input and the predicted meshes:

$$\text{SpAE}^* = \arg \min_{\text{SpAE}} \mathcal{L}_{\text{SpAE}}(M_i, \bar{M}_i) = \arg \min_{\text{SpAE}} [||M_i - \bar{M}_i ||_1]. \quad (6.1)$$

Reconstructing a mesh similar to the input ensures the extraction of features representing the whole object, including shape, expression, pose, and appearance.

6.3.2 Temporal Transformer

Inspired by the success of transformers [33] in natural language processing [263, 264, 265] and computer vision tasks [266, 267, 268], We leverage self-attention to learn the temporal context of meshes sequences. The self-attention mechanism offers the advantage of parallel training, as each token in the input sequence can be processed independently, and it proves effective in modeling long-range dependencies. More details about Transformers are discussed in section 1.3.5.

Given $E = [E_1, E_2, \dots, E_L]$, the embedding sequence that encodes spatial information of the action mesh sequence, we employ self-attention (SA) to learn the temporal context between frames. More precisely, each embedding is used as a token to the transformer. These tokens are then linearly transformed to generate (Q, K, V) the queries, keys, and values, respectively.

$$Q = W_q \cdot E, \quad K = W_k \cdot E, \quad V = W_v \cdot E, \quad (6.2)$$

where $W_q, W_k \in \mathbb{R}^{C_k \times C}$, and $W_v \in \mathbb{R}^{C_v \times C}$ are the weights of the linear transformation, with C_k and C_v are the dimensions of key and value, respectively. Next, the dot product of the query with all the keys, followed by the softmax function are applied to obtain the attention weights,

$$attention(Q, K) = softmax(\frac{Q^T \cdot K}{\sqrt{C_K}}). \quad (6.3)$$

These weights represent affinities between different frames, i.e. the temporal context. Finally, *self-attention* is computed as the weighted sum of the values, where the weights are the learned $attention(Q, K)$.

$$F_{out} = self-attention(Q, K, V) = V \cdot attention(Q, K) \quad (6.4)$$

To enhance the transformer learning ability, we employ multi-head self-attention, which performs h self-attentions with independent W_q , W_k , and W_v , instead of a single self-attention. The outputs of the self-attention heads are concatenated to construct the output features of a self-attention layer. This multi-head self-attention layer is employed three times in our model. Subsequently, we utilize a fully connected layer (FC3) for action classification.

The transformer is trained to minimize the multi-class cross-entropy loss function,

$$T^* = arg \min_T \mathcal{L}_T(\mathbf{y}, \bar{\mathbf{y}}) = arg \min_T [-\sum_{i=1}^{n_c} y_i \log(\bar{y}_i)], \quad (6.5)$$

where $\mathbf{y}$ represents the true label of the input sequence, $\bar{\mathbf{y}}$ represents the predicted probability distribution over the classes, and n_c is the number of classes.

6.3.3 Training Algorithm

Algorithm 5 outlines the training procedure for the Spiral Auto-Encoder (SpAE). The training dataset (M), comprising $n1$ individual meshes, is input into the algorithm. SpAE takes input meshes, individually, extracts their embeddings using the SpAE encoder (4), decodes the embeddings to reconstruct the original meshes (5), and calculates the reconstruction loss (6). The model parameters are updated using gradient descent to minimize the reconstruction loss (7). This process is repeated until convergence.

Algorithm 5: Spiral auto-encoder SpAE training algorithm.

Input: Input meshes $M = [M_1, M_2, \dots, M_{n1}]$;
Output: Output embeddings $E = [E_1, E_2, \dots, E_{n1}]$;
1 SpAE $\leftarrow$ initialized with learnable parameters sampled from a Gaussian distribution;
2 **for** *number of epochs* **do**
3 **for** *the number of iterations in epoch* **do**
4 $E_i \leftarrow \text{Encoder}_{\text{SpAE}}(M_i)$;
5 $\bar{M}_i \leftarrow \text{Decoder}_{\text{SpAE}}(E_i)$;
6 $\mathcal{L}_{\text{rec}} \leftarrow || M_i - \bar{M}_i ||_1$;
7 SpAE $\leftarrow$ SpAE $- \alpha(\partial \mathcal{L}_{\text{rec}} / \partial \text{SpAE})$;
 end
 end

Algorithm 6 describes the training of the temporal transformer T . The training dataset ($E(S_{eq}), Y$), consisting of $n2$ embedding sequences and their corresponding action labels, is fed to the algorithm. Each embedding of the sequence serves as a single token for the Transformer T (4). The input embedding sequences are fed to the transformer T to obtain predicted labels $\bar{Y}$ (5). We then calculate the cross-entropy loss between the predicted $\bar{Y}_i$ and the target Y_i labels (6). The loss is backpropagated, and the parameters of Transformer T are updated using gradient descent (7). These steps are repeated until convergence.

6.4 Experiment: 3D FER

6.4.1 Datasets

The evaluation of our proposed model is performed on two facial expression databases:

The MUG Facial Expression database [178]: (Multimedia Understanding Group Facial Expression database) contains videos of 86 people displaying the seven basic facial expressions: "happiness", "sadness", "surprise", "anger", "disgust", "fear" and "neutral". Each video captures both the onset and the offset phases of these expressions, starting and ending with neutral facial expressions, with the apex expression displayed in the

Algorithm 6: Temporal transformer T training algorithm.

Input: Input Embedding Sequences
 $E(S_{eq}) = [E(S_{eq-1}), E(S_{eq-2}), \dots, E(S_{eq-n2})] =$
 $[[E_1, E_2, \dots, E_l]_1, [E_1, E_2, \dots, E_l]_2, \dots, [E_1, E_2, \dots, E_l]_{n2}];$
Output: The sequences actions labels $\bar{Y} = [\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_{n2}];$

```

1  $T \leftarrow$  initialized with learnable parameters sampled from a Gaussian
   distribution;
2 for number of epochs do
3   for the number of iterations in epoch do
4      $T_{Tokens} \leftarrow [E_1, E_2, \dots, E_l]_i;$ 
5      $\bar{Y}_i \leftarrow T(T_{Tokens});$ 
6      $\mathcal{L}_{CrossEnt} \leftarrow -\sum_{i=1}^C Y_i \log(\bar{Y}_i);$ 
7      $T \leftarrow T - \alpha(\partial \mathcal{L}_{CrossEnt} / \partial T);$ 
   end
end

```

middle. The videos consist of 50 to 160 RGB frames of 896×896 resolution. In this work, we assess the model performance across different configurations, wherein subsets of $\{16, 32, 64, 96\}$ frames, sampled from the onset phase of expressions, are utilized for evaluation.

BU-4DFE [186]: (Binghamton University 4D Facial Expression) comprises 606 sequences of 101 individuals, 58 female and 43 male subjects, with a variety of ethnic/racial ancestries. These individuals perform six facial expressions: "happiness", "sadness", "surprise", "anger", "disgust", and "fear". Scans in the database consist of approximately 35,000 vertices. In our study, we sample 20 frames uniformly from each sequence.

6.4.2 Implementation Details

To prepare the data for training, we perform several pre-processing steps to ensure the compatibility of the data with our model. For the BU-4DFE database, we start by extracting facial landmarks, that are then used in combination with the original scans by FLAME [39] in order to generate new data with consistent vertex count $N=5023$ and the same topology. As for MUG, we employ EMOCA [159] to reconstruct 3D faces from the 2D database. Fig.6.2 illustrates examples of mesh registration using FLAME and EMOCA. Finally, we split the dataset into training and test sets, with 80:20 ratio.

The spiral auto-encoder used in this work uses the following configuration. The encoder consists of four spiral layers of filter sizes $[64, 64, 64, 128]$, each followed by downsampling. An FC transforms the generated features into a vector with size 1024. This vector is then used by a decoder that mirrors the encoder architecture, utilizing spiral filters of sizes $[128, 64, 64, 64]$, each followed by upsampling. ELU (Exponential Linear Unit) is used for all activation functions. To optimize Eq 6.1, the network is trained for 300 epochs using gradient descent and the Adam optimizer. The learning rate is set to 10^{-3} , with a decay rate of 0.99, and a weight decay of 5×10^{-5} to enhance generalization performance. The

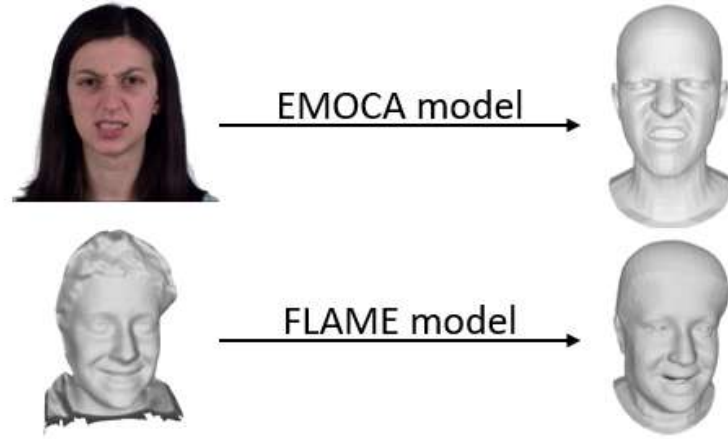


Figure 6.2: Examples of mesh registration.

implementation is based on PyTorch Framework and follows the approach outlined in [7]. The auto-encoder training on both datasets takes approximately 18 hours, utilizing Nvidia GTX 1650 (4GB of memory).

For the proposed transformer, it takes as input a sequence of L feature vectors, where $L=20$ for BU-4DFE and $L \in \{16, 32, 64, 96\}$ for MUG, each vector being of size 1024. Each vector of the sequence is used as a single transformer token. The transformer is built of 3 multi-head self-attention layers with $[2, 3, 4]$ heads, respectively. Gradient descent and Adam optimizer with a learning rate of 10^{-6} and a decay rate of 0.7 are used to train the transformer for 100 epochs, in order to optimize Eq 6.5. PyTorch Framework is once again used for implementation. The training on both datasets takes approximately 1 hour on an Nvidia GTX 1650 (4GB of memory).

6.4.3 Experimental Results

For the MUG database, the registration using FLAME results in natural faces displaying smooth transitions between facial expressions. This high-quality registration process significantly impacts classification outcomes, enabling our model to achieve a notable overall recognition accuracy of 91.07%. These outcomes are visually represented in Fig.6.3 (a), showcasing our model performance across various expression categories. We observe that certain expressions achieve remarkable recognition scores. For instance, the "happiness" and "disgust" expressions achieve high recognition rates of 100% and 97%, respectively. In contrast, the "anger" expression exhibits a relatively lower recognition rate of 85%. This inconsistency is likely due to the more distinctive, universally recognizable characteristics of "happiness" and "disgust," compared to the subtler and more refined features associated with "anger". Conversely, applying FLAME registration to BU-4DFE dataset yields poorer results, with notable errors, artifacts, and noise. Fig.6.4 presents examples that highlight these flaws. The poor registration quality translates

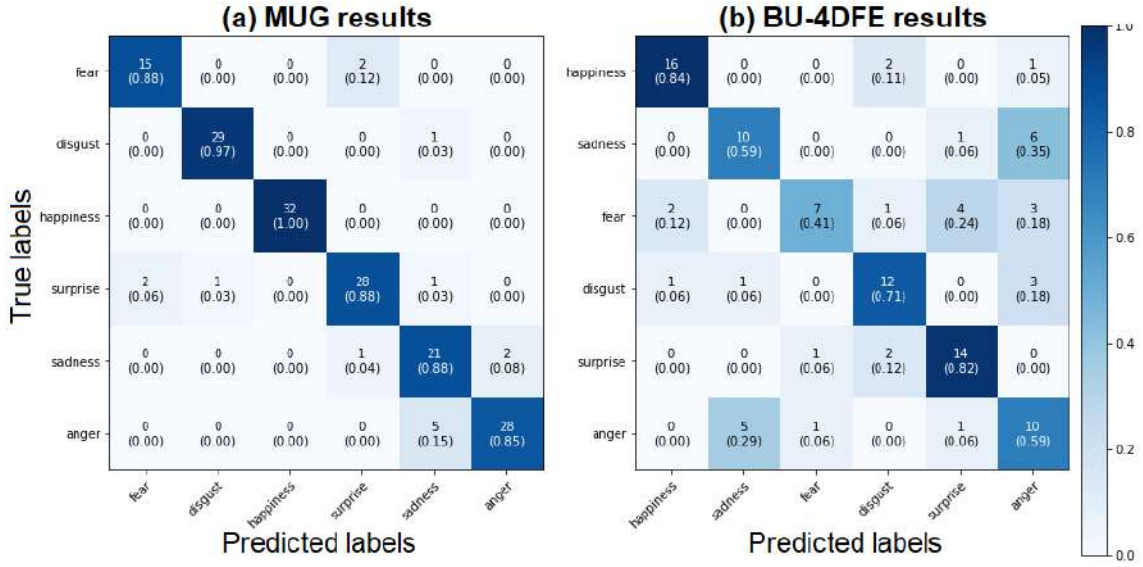


Figure 6.3: confusion matrices showing the classification results of our proposed model on the test set of (a) MUG and (b) BU-4DFE databases. We note that the model is trained on 24 frames.

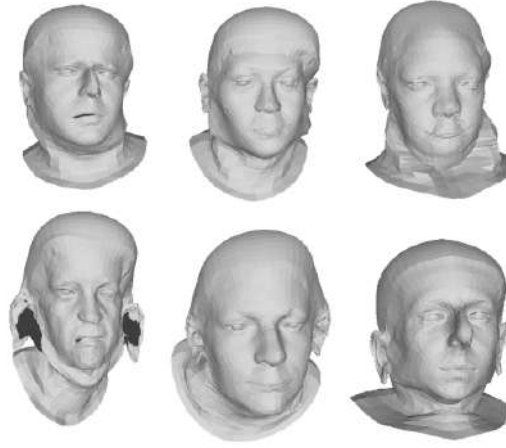


Figure 6.4: Examples of failed FLAME mesh registration from BU-4DFE database.

into a much lower overall classification accuracy of 67.64%, as displayed in Fig.6.3 (b). This underperformance on BU-4DFE underscores the importance of robust registration processes, as flaws in this step can significantly degrade the model accuracy.

Next, We demonstrate the efficiency and scalability of our proposed model by training it on sequences with varying frame counts from the MUG dataset. The results, outlined in table 6.1, represent the accuracy and memory usage across various frame counts. The results reveal that our SpATr exhibits efficient memory usage, starting from a modest $1804 + 115\text{MB}$ (SpAE+T) allocation for 16 frames and gradually extending to $1804 + 220\text{MB}$ for 128 frames, all while maintaining the same level of recognition rate. This highlights the memory-efficiency of our model, proving its adaptability to larger sequences with only marginal increments in memory consumption.

Table 6.1: FACIAL EXPRESSION RECOGNITION ACCURACY (%) USING THE SAME FEATURES ON MUG DATASET.

Frame count	16	32	64	128
#Param SpAE+T (M)	5.85+6.60	5.85+6.61	5.85+6.65	5.85+6.71
Mem SpAE+T (MB)	1804+115	1804+130	1804+158	1804+220
Accuracy (%)	91.15	91.67	90.10	90.63

Additionally, we demonstrate the adaptability of the features extracted by the spiral auto-encoder, underscoring their potential to be utilized across diverse tasks without necessitating retraining efforts. By exploiting these task-free features, we employ a range of classifiers to make predictions related to facial expressions recognition, facial recognition, and sex prediction. The summarized results are outlined in Table 6.2. Our results include an accuracy of 91.07% for the classification of facial expressions, 98.21% for the recognition of facial identities, and 88.33% for sex prediction. These outcomes validate the spiral auto-encoder’s capacity to generate versatile features, resulting in improved resource utilization and significant time savings.

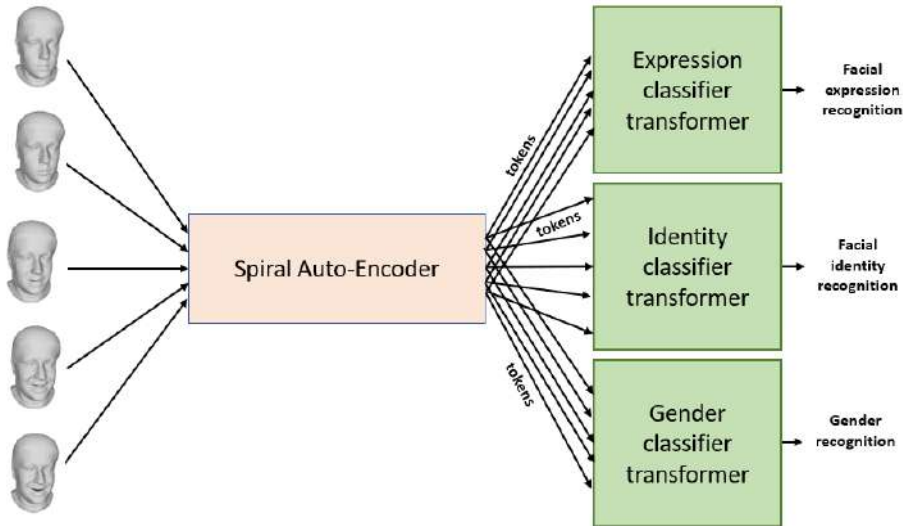


Figure 6.5: Pipeline of using the proposed model for multitasking: facial expression recognition, facial recognition, and sex prediction.

Table 6.2: Facial expression/Identity/sex recognition accuracy (%) using the same features on MUG dataset.

Facial expression recognition	facial identity recognition	Sex recognition
91.07%	98.21%	88.33%

In summary, the model demonstrates impressive memory and resource efficiency across varying frame count, making it capable of handling large sequences, which is a feature that sets it apart from many state-of-the-art models, which often focus on shorter sequences. The versatility of the spiral auto-encoder in extracting features for multiple tasks without

retraining also enhances the model’s overall utility. However, the model’s performance remains highly dependent on the quality of pre-processing and registration. Errors in the FLAME registration, the spiral auto-encoder, or the transformer stages can propagate through the pipeline, potentially leading to classification inconsistencies. Future work should focus on improving the robustness of the pre-processing steps and mitigating error accumulation throughout the model’s architecture.

6.5 Experiment: 3D HAR

6.5.1 Datasets

The evaluation of our method is performed on three MoCap datasets, namely Babel [245], MoVi [246], and BMLrub [247], from the AMASS Archive [248]. AMASS provides high-quality 3D human MoCap data aligned using the SMPL model [243].

Babel [245]: stands out as one of the largest 3D MoCap databases, combining data derived from 15 datasets from the AMASS archive. Babel includes 43.5 hours of recordings that encapsulate over 250 distinct action categories, performed by more than 346 subjects. In our study, we select the most dominant 60-class subset, resulting in 21,653 sequences with single-class labels. These sequences exhibit a diverse range of frames/objects, varying from a few to thousands of frames, with each object composed of 6,980 vertices.

BMLrub [247]: contains sequences of 111 individuals, comprising 50 males and 61 females, performing twenty actions, in a total of 3061 actions performed over a duration of 522.69 minutes. Each sequence consists of 200 to 2500 human-shaped objects (frames) of 6980 vertices. In this work, we remove two actions that are represented by less than 3 samples.

MoVi [246]: contains sequences of 86 individuals, comprising 29 males and 57 females, performing a collection of 20 everyday actions and sports movements, along with one self-chosen movement. The dataset includes a total of 1864 actions performed over a duration of 174.39 minutes. Each sequence consists of an average of 150 to 2000 human-shaped objects (frames) of 6980 vertices. We remove the "random" self-chosen movement for each subject in our analysis.

6.5.2 Implementation Details

Prior to training the model, we perform a subject-independent data split into training and testing sets, with an 80:20 ratio, for both MoVi and BLMrub datasets. As for Babel, we use the exact same split in [242], to the individual data instances level the level for fair comparison. In this split, the data is split into training, validation, and testing sets, with a 70:15:15 ratio.

For the spiral auto-encoder SpAE, The encoder consists of four spiral convolutional layers, with filter sizes of [16, 32, 64, 128], followed by downsampling. Subsequently, a fully connected layer transforms the codes into an embedding vector of size 1024. The decoder is built using five spiral convolutional layers, with filter sizes of [128, 64, 32, 32, 16], followed by upsampling. All activation functions are Exponential Linear Units (ELUs) [269]. In order to optimize Eq 6.1, the network is trained for 300 epochs using gradient descent and Adam optimizer [195], with a learning rate of 10^{-3} and a decay rate of 0.99. Weight decay 5×10^{-5} is applied to improve the generalization performance of the network. The implementation of the network is carried out with PyTorch Framework, based on the implementation of [7]. The training process takes approximately 10 hours on both MoVi and BMLrub datasets, while it takes approximately 26 hours on Babel dataset, using an Nvidia Titan V GPU (12GB of memory).

The temporal transformer T takes as input a sequence of L embedding vectors of size 1024, then uses each one of them as a single token. It consists of 3 multi-head self-attention layers with [2, 2, 2] heads. In order to optimize Eq 6.5, we train the transformer for 100 epochs with gradient descent and Adam optimizer, with a learning rate of 10^{-4} and a decay rate of 0.7. The implementation of the network is conducted with PyTorch Framework. Using an Nvidia GTX 1650 (4GB of memory), the training takes approximately 1 hour on MoVi and BMLrub datasets and about 3 hours on Babel.

We established baseline performance by training P4Transformer [262], PSTNet [241], and SequentialPointNet [240] using public code provided by the authors, with minor additions involving point cloud frame extraction from mesh frames in the dataset. For STMT, 2s-AGG-FL, 2s-AGCN-CE, and CTR-GCN, we directly utilized results from [242], as we replicated their evaluation protocol and data partitioning with precision, down to the instance level. This approach ensures a robust basis for comprehensive comparisons, aligning our model performance with the benchmarks in a consistent and reliable manner.

6.5.3 Experimental Results

After training our model, we predict the action classes of scenes from the testing subsets. Fig. 6.7 presents examples of both successful and failed prediction attempts performed by our proposed model. Notably, we achieve a mean accuracy of 70.36%, 95.42%, and 86.99% for Babel, MoVi, and BMLrub respectively. Fig. 6.6 displays the performance across action classes visualized as confusion matrices. The figure indicates that our model provides advantageous results on MoVi and BMLrub datasets. However, the Babel dataset poses a challenge due to its inherent characteristics, resulting in comparatively lower accuracy. This challenge is attributed to the considerable class imbalance in Babel. The training set encompasses class instances ranging from as few as 6 to a substantial 3000 instances, creating a highly skewed distribution (long-tailed dataset), which makes it difficult for the model to achieve high recognition rates. Fig.6.8 displays the class distribution of Babel.

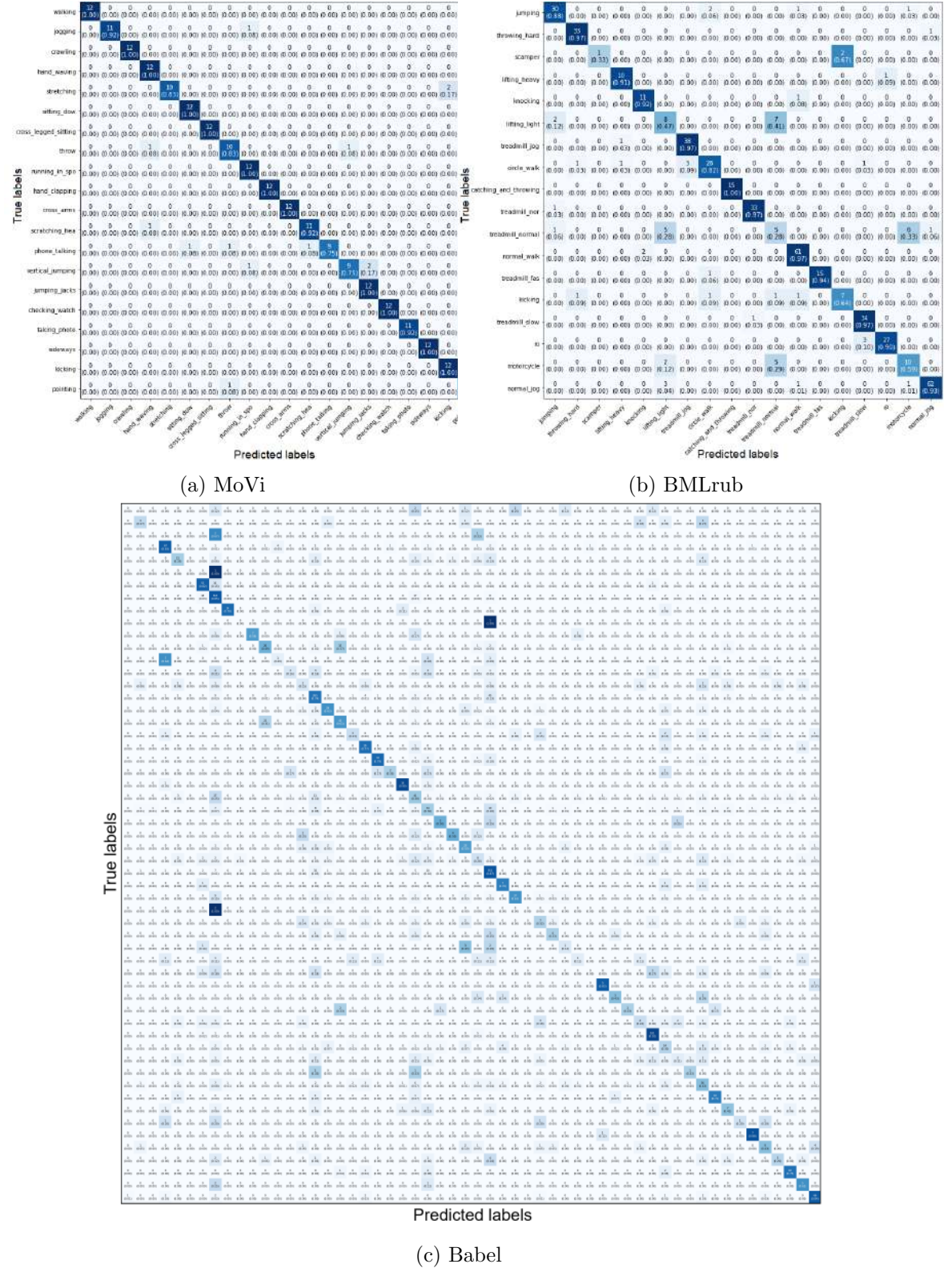


Figure 6.6: confusion matrices showing the classification results of our proposed model on the test set of (a) MoVi, (b) BMLrub, and (c) Babel databases. We note that the model is trained on 24 frames.

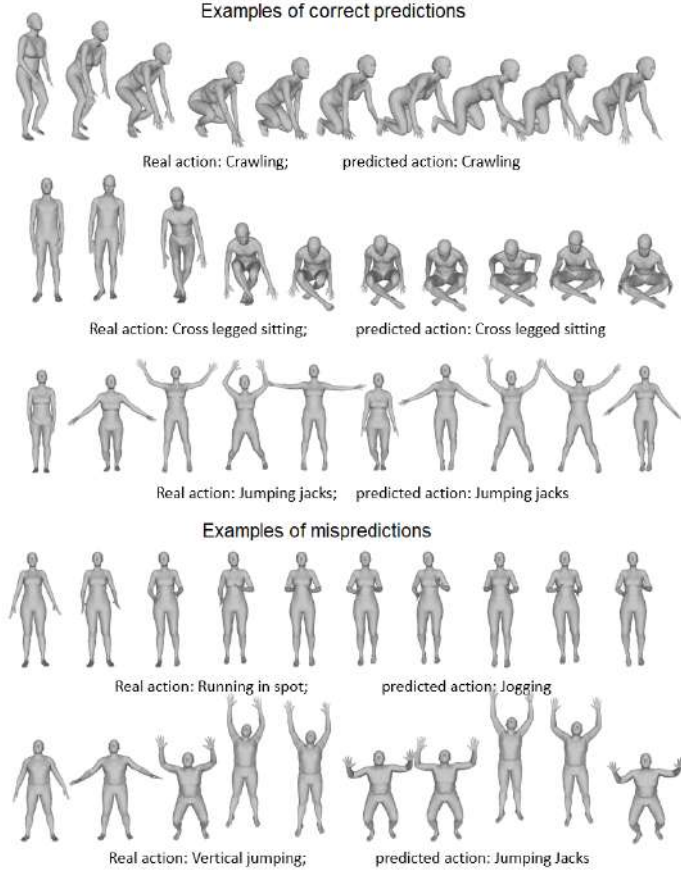


Figure 6.7: Examples of correct and faulty predictions.

Next, we compare our approach with the state-of-the-art in terms of max achieved accuracy, using Babel, MoVi, and BMLrub datasets. The baselines encompass mesh, point cloud-based, and skeleton-based methods. The accuracy results of our model and the baselines across all datasets are presented in table 6.3. Notably, our method demonstrates competitive performance compared to the baselines. It achieves the highest accuracy on Babel and BMLrub datasets and the second-highest accuracy on the MoVi dataset.

Table 6.3: Action recognition accuracy (%) results on MoVi, BMLrub, and Babel datasets.

Method	Input	MoVi	BMLrub	Babel	
		Top-1(%)	Top-1(%)	Top-1(%)	Top-5(%)
2s-AGCN-FL (CVPR'19)	3D Skeleton	-	-	49.62	79.12
2s-AGCN-CE (CVPR'19)	3D Skeleton	-	-	63.57	86.77
CTR-GCN (ICCV'21)	3D Skeleton	-	-	67.30	88.50
MS-G3D (CVPR'20)	3D Skeleton	-	-	67.43	87.99
P4Transformer (CVPR'21)	Point cloud	91.25	81.24	63.54	86.55
PSTNet (ICLR'21)	Point cloud	88.75	80.20	61.94	84.11
SequentialPointNet (IEEE TII'22)	Point cloud	98.84	82.45	62.92	84.58
STMT (CVPR'23)	Mesh	-	-	67.65	88.68
SpATr (Ours)	Mesh	95.42	86.99	70.36	89.12

Table 6.4: Action recognition accuracy (%) and memory usage/batch on MoVi dataset.

Method	Input	nb Frames	nb points	Memory (MB) / batch size	Accuracy (%)
P4Transformer	Point cloud	12	2048	1184/4	89.16
		24		2298/4	91.25
		48		3224/2	90.00
PSTNet	Point cloud	12	2048	2129/2	88.33
		24		3206/2	88.75
		48		3151/1	88.91
SequentialPointNet	Point cloud	12	2048	4352/8	98.46
		24		11806/8	98.84
		48		40105/8	99.22
Our model	Mesh	12	6890	480+157/8	93.75
		24		480+167/8	94.16
		48		480+187/8	95.42
		96		480+230/8	95.00
		192		480+328/8	94.58
		384		480+554/8	94.17
		960		480+1480/8	94.58

The findings of this analysis suggest that our model shows competitiveness with state-of-the-art methods in terms of recognition rate, while exhibiting significantly better performance in terms of memory efficiency and handling large sequences.

We note that the batch size varies between setups due to the differences in the maximum capacity of the GPUs used to train the models (our model, P4Transformer, and PSTNet were trained on NVIDIA GeForce GTX 1650 4GB, while SequentialPointNet was trained on NVIDIA Quadro RTX 8000 48GB).

6.5.4 Ablation Study

In order to demonstrate the significance of self-attention in capturing temporal context and to compare it to other techniques we conduct an ablation study on the MoVi dataset. We train different versions of our model, where the temporal transformer is replaced by other networks, to observe their impacts in terms of recognition accuracy and memory efficiency. Specifically, we train three versions of the proposed model: one with a Multi-Layer Perceptron (MLP) with three hidden layers, another using a Long Short-Term Memory (LSTM) network with three LSTM layers, and the third with a Convolutional Neural Network (CNN) with three convolutional layers. All networks were trained multiple times using sequences of different lengths {24,48,96,192,960}. We maintain the same spiral auto-encoder, parameters, losses, and number of epochs for training all models. The accuracy and memory consumption for the trained models is outlined in table 6.5, enabling a comparative analysis with the complete model.

Table 6.5: Ablation study results.

Model nb frames	with MLP		with LSTM		with CNN		Our model	
	Acc (%)	Mem (MB)	Acc (%)	Mem (MB)	Acc (%)	Mem (MB)	Acc (%)	Mem (MB)
24	94.58	585	95.42	25	88.75	12	94.14	167
48	95.00	1192	94.58	29	91.25	69	95.42	187
96	92.92	2315	93.75	40	89.58	93	95.00	230
192	Out Of Memory ($\geq$ 4GB)		90.00	59	92.50	184	93.75	328
960	Out Of Memory ($\geq$ 4GB)		74.17	212	87.04	488	94.58	1480

The results reveal the following:

- In comparison to the MLP-based model, our proposed model demonstrates competitive accuracy while being more memory efficient. This is attributed to the dense connectivity and larger parameter count of MLP networks.
- Although the LSTM-based model exhibits lower memory consumption and comparable performance to our model for short sequences, it falls short when applied to larger sequences due to difficulties in capturing long-term dependencies.
- While the CNN-based model demonstrates better memory efficiency than our model, its accuracy significantly falls behind. This arises from the limited capability of CNNs in capturing global dependencies and long-range interactions in the data.

Therefore, the transformer-based model emerges as the most favorable choice, showcasing superior performance in terms of accuracy and the ability to capture long-range dependencies, while still maintaining reasonable memory efficiency compared to the other networks considered in this study.

6.5.5 Multi-Tasking Evaluation

We highlight the adaptability of the task-free features derived from the spiral auto-encoder, emphasizing their utility across diverse tasks without the need to retraining. Utilizing features extracted from the MoVi dataset, we conduct different tests predicting subject actions, identities, and sexes. We use the same settings as the previous tests for 24 frames. We modify the data splits for each task. For action predictions, the splits remain unchanged. For identity recognition, we randomly select 20 identities, and we split the data into 80/20% for training and testing. For sex prediction, the dataset is split into 80/20% subject-dependent segments. Regarding model adjustments, we retain the same feature extractor with identical weights, ensuring consistent embeddings. These embeddings serve as inputs for training three temporal transformers tailored to the specific requirements of each task. Although the transformer architecture remains constant, the last layer is modified to accommodate the distinct number of classes for each task. Fig.6.9 illustrates the pipeline followed for this experiment.

The results show a notable (95.42%) accuracy for action classification, perfect accuracy (100%) for identity recognition, and also (100%) for sex prediction. These results highlight the spiral auto-encoder capability to generate versatile features, leading to optimized resource allocation efficiency and significant time savings.

6.5.6 Parameters Analysis

Frames Number: In Fig. 6.10, we summarize the impact of different numbers of frames on the performance of the transformer model. We retrain and evaluate the model using

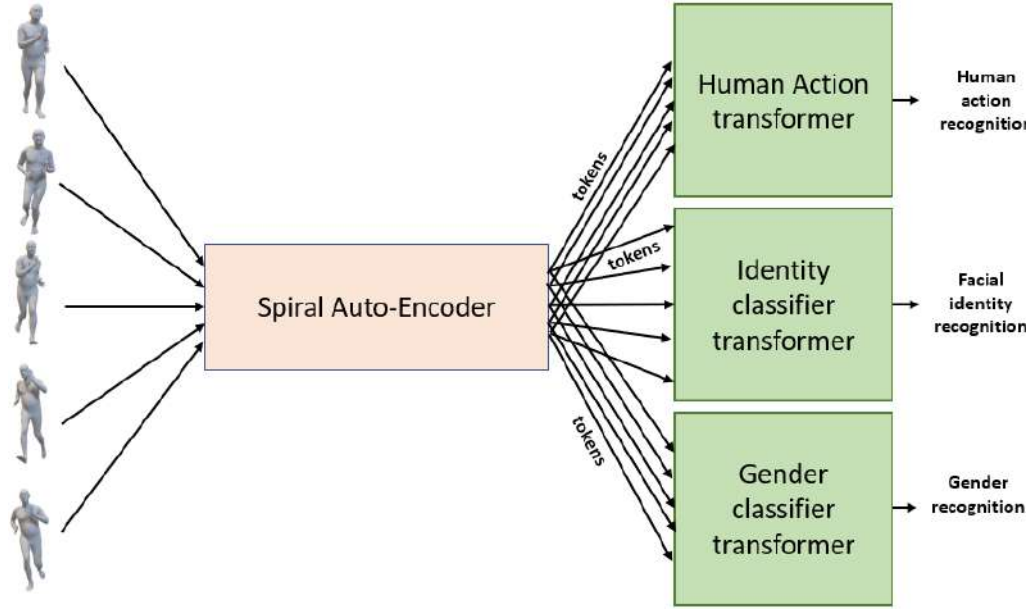


Figure 6.9: Pipeline of using the proposed model for multitasking: action recognition, identity recognition, and sex recognition.

various numbers of frames and recorded the corresponding accuracies. Initially, as the number of frames increased from '12' to '48', the accuracy steadily improved, reaching its peak at '48' frames with an accuracy of 95.42%. Beyond this point, with further increases in the number of frames, the accuracy exhibits a slight fluctuation but remains consistently high, indicating the model scalability and robustness in handling longer sequences.

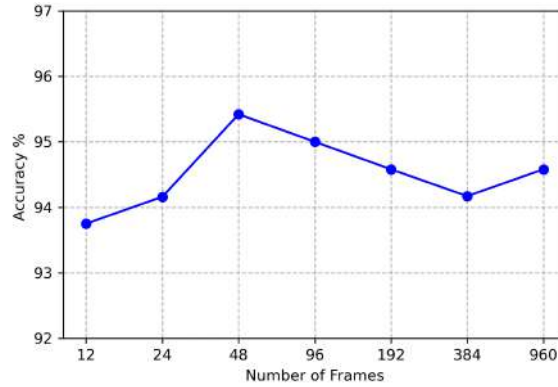


Figure 6.10: Evolution of model accuracy based on the number of frames.

Attention head Number: The performance of the transformer model is largely influenced by the number of attention heads used. We investigate this connection by examining the effect of varying the number of multi-head self-attention heads on the temporal transformer architecture. Fig. 6.10 reports the evolution of the model accuracy depending on the number of heads. The figure demonstrates an initial enhancement in accuracy as the number of heads is augmented from [1, 1, 1] to [2, 2, 2]. However, beyond this point, the accuracy reaches a plateau and diminishes with further increases

in attention heads. This behavior could be explained by the positive correlation between the number of attention heads and the model capacity to capture complex patterns and dependencies. As a result, the model is more prone to overfitting in scenarios where the complexity and scale of the data are insufficient.

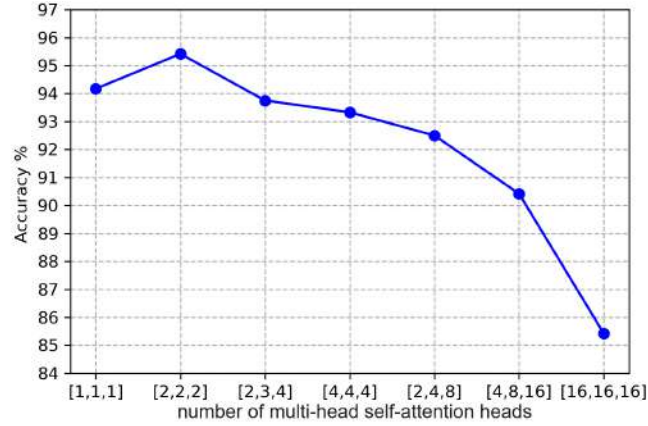


Figure 6.11: Evolution of model accuracy based on the number of multi-head self-attention heads.

6.6 Discussion

Our experiments prove that the proposed model shows competitive results to state-of-the-art. This is due to the use of a combination of **1)** spiral convolutions that effectively capture the geometric information directly from mesh structures, and **2)** self-attention, employed in the transformer, to model the temporal evolution of the sequence.

In addition, the proposed model proved to be very light compared to the point cloud-based methods. when trained on sequences of the same length, our model shows more memory efficiency, using only 2% - 15% of the memory used by the baselines. This, combined with the use of transformers that can handle large sequences, enables our model to process videos of more 1000+ frames without any memory issues. Moreover, the proposed model offers more resource efficiency by extracting task-free features that can be used for multiple purposes without the need to retrain the feature extractor.

However, while the SpATr model presents impressive performance and resource efficiency, several limitations persist that highlight areas for future improvement. A primary limitation is the dependency on meshes with fixed topology, thereby limiting its applicability to Motion Capture (MoCap) data. This contrasts with point cloud-based methodologies, which show the ability to handle irregularities present in 3D scans. While registration techniques can be employed to align irregular scans with fixed templates, these methods often require considerable time and computational cost. To address this challenge, future research could explore enhancing the model flexibility to manage varying

topologies, potentially through the utilization of 3D convolutions suitable for irregular scans.

Moreover, decoupling the action recognition in our model introduces the possibility of errors propagation through both sub-networks, consequently influencing overall accuracy. A potential mitigation strategy involves exploring alternative end-to-end architectures.

Additionally, while transformers exhibit promise in FER/HAR by effectively managing long-range dependencies, their demanding data requirements present a significant challenge, particularly within the MoCap data domain. Strategies of data augmentation could be explored to enhance the model performance. Addressing these challenges stands to contribute to refining and enhancing the effectiveness of the SpATr model in various applications.

6.7 Conclusion & Perspectives

In this chapter, we introduce a novel fully learned 3D MoCap Facial Expression Recognition (3D FER) and Human Action Recognition (3D HAR) model, that strategically separates the task into spatial and temporal modeling components. Our approach addresses the resource consumption challenges of state-of-the-art methods by leveraging lightweight spiral convolutions for 3D mesh spatial encoding and a compact temporal transformer for capturing temporal evolution.

Comprehensive evaluations on multiple datasets, including MUG and BU-4DFE for facial expressions and Babel, MoVi, and BMLrub for human actions, showcase the accuracy and memory usage of our model in comparison to state-of-the-art approaches across various settings, examining the trade-off between accuracy and memory efficiency. The results show that our method achieves competitive performance in recognition accuracy and significant improvement in memory efficiency.

Despite its strengths, SpATr faces certain limitations, primarily its dependence on fixed-topology meshes and the potential introduction of errors from decoupling action recognition. This underscores avenues for future research, emphasizing the need for model refinement and exploration of alternative architectures.

Chapter 7

Latent-Motion 3D Diffusion Model for 4D Facial Expression Generation

This chapter presents initial steps in an ongoing project. The theoretical framework is well-established, but the experimental results discussed here are preliminary, as the experiments are still in the early stages.

7.1 Introduction

Facial expression generation has recently attracted attention in the research community. Integrating the ability to synthesize facial expressions into machines enhances natural and smooth human-computer interactions. However, this task presents several inherent challenges: **1)** learning the spatial distribution of facial features and structures, including variations in facial anatomy and background elements, **2)** capturing the natural and smooth progression of facial expressions in effective temporal representations, **3)** preserving the identity of the input subject throughout the expression sequence to maintain consistency.

With the advancement of generative models such as GANs [28] and DMs [270], several studies have addressed the task of 2D facial expression generation, producing promising and visually convincing results [128, 127, 136, 209, 8]. However, the demand for 3D facial expression generation is growing due to its vast potential applications. These include character animation in movies, video games, virtual environments, and advancements in virtual reality (VR) and augmented reality (AR) technologies. With the rapid growth of the metaverse, the creation of realistic 3D avatars has become particularly important for social interaction, immersive experiences, and telepresence.

Despite its importance, 3D facial expression generation is more challenging and less explored compared to its 2D counterpart. This is due to additional challenges inherent in the 3D domain. One major challenge is the lack of 3D data suitable for facial expression generation. Acquiring high-quality 3D facial data is resource-intensive compared to 2D image data, limiting the amount of training data available for models. Furthermore, handling 3D data introduces higher dimensionality and complexity, given the unstructured nature of mesh data. Processing this data demands significant computational resources,

which can be a limiting factor.

To address these challenges, researchers have proposed various methods for 3D facial expression generation. Some approaches manipulate the input mesh to produce another static mesh that display the target pose or expression. While these methods can generate realistic static expressions, they do not account for the dynamic nature of facial expressions, which are naturally temporal.

To incorporate the dynamic aspect, other methods train models to regress 3D facial geometry from 2D video sequences, often processing frames individually [161, 162, 271, 159]. These approaches aim to reconstruct 3D facial expressions by leveraging temporal information from videos. However, they typically transfer expressions from 2D without explicitly modeling the temporal evolution of facial shapes, potentially missing out on the detailed variations of expression dynamics.

Another group of studies attempt capturing facial motions by exploiting external information such as audio features and speech signals [163, 164, 165, 166, 167]. While these methods incorporate additional modalities to inform facial movement, they often rely on external data sources and do not explicitly model the temporal dynamics inherent to facial expressions. Therefore, they may depend heavily on the quality and availability of external information, limiting their applicability.

To address this limitation, other works have directly tackled the 3D facial expression generation problem by generating sequences of 3D facial landmarks representing the expressions. These works have exploited sequential deep learning approaches, such as Long Short-Term Memory networks (LSTMs) [168], Variational Autoencoders (VAEs) [169], Wasserstein Generative Adversarial Networks (WGANs) for manifold learning [172], or Diffusion Models (DMs) [173], to generate 3D facial landmarks. The generated landmark sequences are then used to manipulate the input mesh using techniques like spiral autoencoders [7]. However, this multi-stage process can introduce errors and may limit the expressiveness of the generated sequences.

In this work, we introduce a novel latent diffusion-based model for 3D facial expression generation, referred to as the Latent-Motion 3D Diffusion Model (LM3DM). Unlike models that first generate landmarks, then integrate them as conditions into the latent space, and use them for the mesh generation, our proposed approach operates directly within the latent space of the mesh generator. By working in this abstract, high-level representation, the model gains more control over the complexity and structure of the embedding space, leading to more control on versatility and flexibility in the generated expressions.

Moreover, our model specifically focuses on the motion features within the embedding space. By applying diffusion processes to the motion space, it incorporates stochasticity, resulting in more varied and smoother motions while preserving the consistent identity of the subject throughout the mesh sequences. This is achieved by decoupling the embedding space into content and motion components using two distinct encoders. One encoder

processes an input mesh of the subject in an initial pose, capturing the static features, while the other handles vectors of landmarks that serve as driving information for motion. After learning these embedding spaces, diffusion is applied exclusively in the motion space to learn expression patterns and generate new expressions. Diffusion models have shown great promise in generating high-quality data by modeling the distribution of complex datasets through iterative refinement. By applying diffusion directly in the motion embedding space, our model can generate realistic and diverse facial motion sequences.

To the best of our knowledge, this is the first work to explore the use of diffusion models directly within the mesh generator embedding space for dynamic 3D facial expression generation. While 3D landmarks are utilized during the training phase to construct the motion embedding space, the inference phase bypasses the need for these intermediate landmark representations. Instead, the model directly generates the motion embedding, optimizing the process. Fig.7.3 provides a detailed overview of the proposed model architecture.

In summary, our work offers the following key contributions:

1. We introduce a Latent-Motion 3D Diffusion Model (LM3DM) for dynamic 3D facial expression generation. The model decouples the task into content and temporal motion, allowing us to explicitly model the temporal dynamics. For this purpose, we use a latent diffusion-based approach to capture the temporal motions and align them with the initial mesh.
2. To the best of our knowledge, our proposed model is the first to directly operate on the mesh generator latent space to learn 3D facial expression dynamics.
3. We present preliminary results of our model on the MUG facial expression dataset [178] and COMA dataset [153]. The results highlight the effectiveness of the proposed model in generating 3D facial expressions and demonstrate the large variance in the generated expressions.

The remainder of this chapter is organized as follows. In Section 7.2, we review related work in 3D facial expression generation. In section 7.3, we provide a detailed description of our proposed LM3DM model. Section 7.4 presents experimental results demonstrating the effectiveness of our approach. Finally, Section 7.5 concludes the chapter with a summary and a discussion of the challenges and potential directions for future research.

7.2 Related Work

In Section 2.3, we present an detailed review of the current state-of-the-art techniques in 3D facial expression generation. Building on it, this section focuses specifically on methods that learn both the spatial and temporal aspects of dynamic facial expression generation.

By examining these approaches in detail, we aim to position our work within the research literature.

Potamias et al. [168] proposed modeling facial motion using a temporal encoder based on a Long Short-Term Memory (LSTM) network, which generates a per-frame latent code starting from a per-frame expression labels. These codes are then fed into a mesh decoder, which generates per-vertex displacements that are applied to a neutral 3D face, resulting in expressive meshes. Despite promising results, this approach is deterministic and for a given label the same displacements are generated without variation. *Zou et al.* [169] introduced a Transformer-based Variational Auto-Encoder, exploiting the advantages of transformers, known for their efficiency at processing long sequences, and VAE, which are capable of generating diverse outputs while providing structured latent representations of data. The proposed model uses a Transformer-based Variational Auto-Encoder to generate landmark sequences conditioned on a desired expression. These sequences are then used to guide landmark-based mesh deformation, resulting in animated 3D faces. *Otberdout et al.* [170, 171, 172] presented a novel approach that uses a conditional manifold-valued Wasserstein GAN to model facial expression dynamics as curves on a hypersphere. The the temporal evolution and mesh deformation are handled separately by two different networks. A manifold-valued GAN (Motion3DGAN) learns the expression dynamics as sequences of points (representing each landmark set) on a hypersphere, while a Sparse2Dense mesh Decoder (S2D-Dec) generates dense 3D faces guided by the landmarks motion for each frame of the sequence. *Zou et al.* [173] exploits the advantages of diffusion models for generating more diversity and high-quality outputs. Using Diffusion Probabilistic Models (DDPMs), the model generates sequences of landmarks that represent facial expressions. These landmarks are then used to drive the generation of 3D meshes, achieving significant success by leveraging the high variance and quality output capabilities of diffusion models.

7.3 Proposed Approach

Our proposed model addresses the challenge of conditional one-to-many generation, specifically for 4D facial expression generation from a single input mesh. The model takes as input an initial face mesh m_{in} and a target expression label y , and produces a sequence of meshes $\bar{S}_{eq} = [\bar{m}_1, \bar{m}_2, \dots, \bar{m}_N]$, in which the input face performs the desired expression. The proposed model, referred to as Latent Motion 3D Diffusion Model (LM3DM), is formulated as $\text{LM3DM}(m_{in}, y) \rightarrow \bar{S}_{eq}$.

The LM3DM framework decomposes the task of 4D facial expression generation into two key components. First, an autoencoder learns the content and motion latent spaces. Second, a diffusion model learns generating motion within the learned motion latent space. Fig.7.3 illustrates the overview of the proposed model.

The process starts with an unsupervised learning stage, where a conditional mesh Spiral autoencoder (SAE) is trained. The SAE is designed to take an initial input mesh

m_{in} (representing the first frame of the sequence) and a driving signal L_i (landmarks corresponding to another frame in the sequence) to reconstruct the target mesh frame $\bar{m}_i$. During training, the autoencoder learns to decompose the content (static features of the face) from the motion (dynamic facial expressions), establishing two different latent spaces: one for content and another for motion. Once the content and motion latent spaces are learned, diffusion model is used to operate exclusively within the motion latent space. Using the motion embeddings learned by the autoencoder, the diffusion model is trained to generate new motion sequences conditioned on specific expression labels y . By focusing only on the motion latent space, the diffusion process captures the variability and dynamics of facial expressions, ensuring that the generated sequences are both diverse and realistic.

During inference, the motion encoder and driving signal are replaced by the motion embeddings generated by the diffusion model. This allows the system to generate a realistic sequence of facial expressions directly from the initial input mesh m_{in} and the desired target expression y , without the need for intermediate landmark representations.

7.3.1 Spiral Auto-Encoder

In the first phase of our approach, we employ a conditional spiral autoencoder (SAE) designed to learn the embedding spaces for both content and motion features. The SAE takes as input the reference mesh m_{in} and the i_{th} driving frame information L_i , with the objective of generating the i_{th} frame $\bar{m}_i$ that resembles the target frame m_i . This process is formally represented as $\text{SAE}(m_{in}, L_i) \rightarrow \bar{m}_i$.

To accomplish this, our autoencoder consists of three key components, as illustrated in Fig.7.1. First, a content encoder (E_c) learns the embedding z_{in} from the input reference mesh m_{in} of N_v vertices, extracting the identity and spatial features. This enables subject identity and content details preservation in the generated frames. In addition, we use a motion encoder (E_m) to capture the motion embedding z_i from the driving information L_i . Finally, a decoder network reconstructs the target mesh $\bar{m}_i$ by combining the embeddings $[z_{in}, z_i]$ from both the expression encoder and the identity encoder.

During training, we select two frames from the same sequence: the reference frame m_{in} and a randomly selected target frame m_i , along with its corresponding driving information L_i . m_{in} and m_i are meshes with dimensions $N_v \times 3$. The encoders E_c and E_m process the reference frame x_{in} and the driving information L_i to generate latent maps z_{in} and z_i , respectively, each with dimensions C_z , where C_z represents the embedding size.

Since we are using fixed topology mesh data, we suggest using spiral convolutions [7], detailed in section 1.3.6, as a building block of a fully differentiable Spiral encoder that learns hierarchical representations z_{in} , and a decoder that learns to reconstruct the meshes $\bar{m}_i$ from the representations $[z_{in}, z_i]$. Spiral convolutions are advantageous in taking into account the connectivity of the mesh graph to design global (across the entire

mesh) and local (within patches) ordering of the vertices. This enables local processing while capturing the global context for each shape. In addition, spiral filters benefit from weight sharing, which reduces the number of parameters, resulting in a lightweight and easy-to-optimize network.

The SAE is trained in an unsupervised manner using a reconstruction loss applied to unlabeled video frames. The training objective is to minimize the difference between the generated mesh $\bar{m}_i$ and the actual target mesh m_i , using the following loss function:

$$\text{SpAE}^* = \arg \min_{\text{SAE}} \mathcal{L}_{\text{SAE}}(m_i, \bar{m}_i) = \arg \min_{\text{SAE}} [\|m_i - \bar{m}_i\|_1]. \quad (7.1)$$

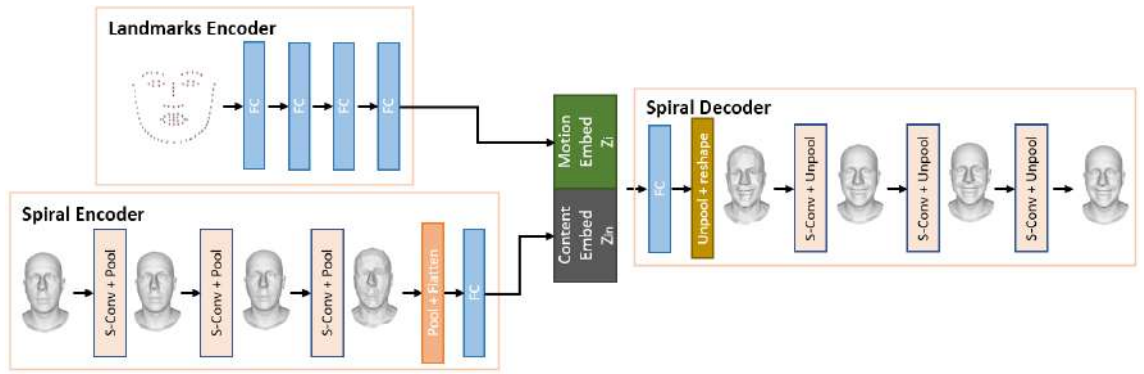


Figure 7.1: Our conditional Spiral autoencoder overview.

7.3.2 Proposed Flow Diffusion Model

In the second phase of our training, we introduce a diffusion model based on a 2D-UNet architecture, one dimension for the vector embedding and another for the temporal evolution. This model is designed to learn the generation of motions within the motion embedding space, which has previously been learned by the motion encoder E_m . The diffusion model training follows a conditional framework, where it is guided by the reference image embedding z_{in} and by the target motion label y . These inputs are important as they allow the model to generate motions that fit the reference mesh (identity/body) and align with the desired motion category. The goal is to model the conditional distribution p_θ , which enables the generation of motions corresponding to the given conditions.

During the diffusion process, noise is added incrementally and unconditionally to the sequence of motion embeddings $q(x_t|x_{t-1})$, where $x_0 = z_{i=1,\dots,N} = E_m([m_1, m_2, \dots, m_N])$. Here, $z_{i=1,\dots,N}$ represents the sequence of motion embeddings corresponding to a sequence of frames $[m_1, m_2, \dots, m_N]$. The objective of the denoising model $p_\theta(x_{t-1}|x_t, z_{in}, y)$ is to predict the noise $\mathcal{E}$ that was added to the embeddings, taking into account the current noised embedding x_t , the content embedding z_{in} , and the target expression label y .

The denoising process is iteratively performed across the sequence, following the

probabilistic formulation:

$$p_{\theta}(x_0, \dots, x_{T-1} | x_T, z_{in}, y) = \prod_{t=1}^T p_{\theta}(x_{t-1} | x_t, z_{in}, y),$$

where each step of the denoising process is defined by:

$$p_{\theta}(x_{t-1} | x_t, z_{in}, y) = \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t, z_{in}, y, t); \sigma_{\theta}(x_t, y, t)^2 \mathbf{I}).$$

In this equation, μ_{θ} and σ_{θ} are the predicted mean and variance, respectively, and $\mathcal{N}$ denotes the normal distribution. These parameters are conditioned on the noised embedding x_t , the identity embedding z_{in} , the expression label y , and the time step t .

The diffusion model is trained to minimize the loss function:

$$\mathcal{L}_{DM} = \mathbb{E} [\|\varepsilon - \varepsilon_{\theta}(x_t, t)\|^2].$$

This loss measures the difference between the real noise ε introduced during the forward diffusion process and the noise $\varepsilon_{\theta}(x_t, t)$ predicted by the model during the reverse denoising process. This loss ensures that the model accurately learns the noise distribution, which is critical for successful expression motion generation.

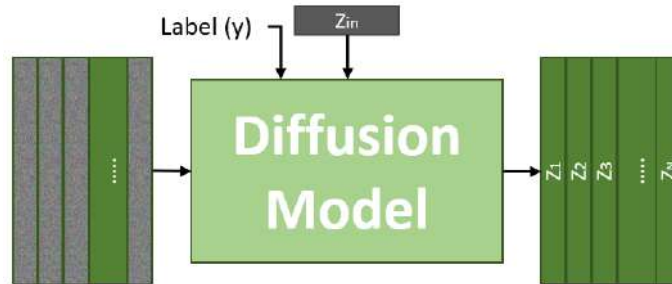


Figure 7.2: Motion latent diffusion overview.

7.3.3 Inference

In the inference phase, the driving signal and motion encoder are removed, and instead, the embeddings generated by the diffusion model are directly used in the latent space of the SAE.

The inference process begins with an input mesh m_{in} and a target expression label y . The first step involves extracting the content embedding z_{in} from the reference mesh m_{in} using the content encoder E_c . This embedding captures the identity and spatial features of the subject, ensuring that the generated frames maintain content consistency throughout the sequence.

With z_{in} obtained, the next phase is to generate a latent representation that encodes the target motion using the diffusion model, which takes into account both the content

embedding z_{in} and the target motion label y . The diffusion model operates by sampling an initial noise vector sequence from a normal distribution and then iteratively denoising it through a series of reverse diffusion steps. The denoiser, represented by the network θ , progressively refines the noised vector sequence through these steps, working backward from an initial state $t = T$ to the final state $t = 1$, resulting in the final motion embedding $\{\bar{z}_{1...N}\}$ that captures the dynamic evolution of the motion.

The final stage involves generating the meshes of the output sequences. Each frame embeddings $\bar{z}_i$ is combined with the content embedding z_{in} and passed through the decoder of SAE, which reconstructs the visual content of each mesh (frame). This process results in a sequence of meshes $[\bar{m}_1, \dots, \bar{m}_N]$ that compose the final sequence $\bar{S}_{eq}$. Fig.7.3 shows an overview of the inference process.

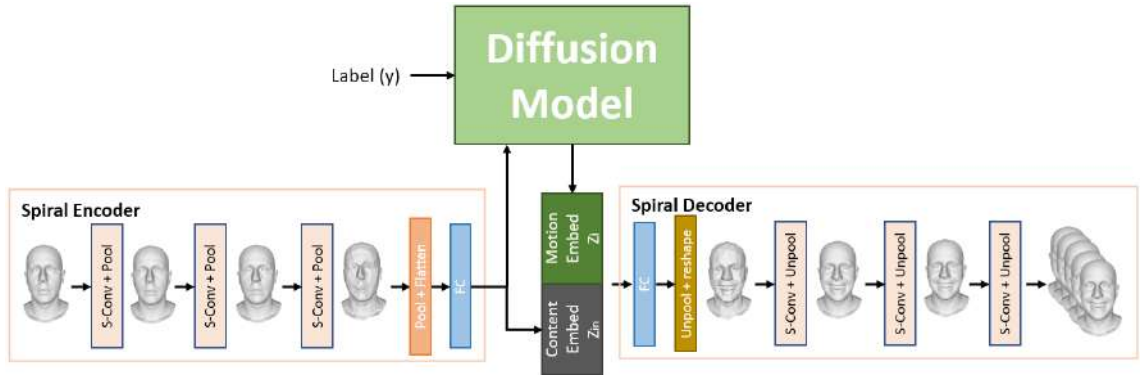


Figure 7.3: Overview of the LM3DM in the inference phase.

7.4 Experiment

In this section, we evaluate our proposed model on the MUG and COMA databases. We detail the experimental setup, training process, and evaluation metrics. And, we shows preliminary results to prove the concept of our method.

7.4.1 Datasets

We evaluate our proposed model, we use conducts experiments on the following datasets:

the Multimedia Understanding Group database (MUG) [178]: This database comprises 52 subjects expressing seven distinct emotions. Notably, each video starts and ends with a neutral expression, with the apex expression displayed in the middle. The videos consist of 50 to 160 RGB frames. For our evaluation, we split the dataset by randomly selecting 32 individuals for training, resulting in 612 videos, and the remaining 18 individuals for testing, resulting in 317 videos.

CoMA dataset [153]: consists of 12 subjects, each one performing 12 extreme and asymmetric expressions. Each expression comes as a sequence of 140 meshes on average,

with $N = 5023$ vertices. For our evaluation, we split the dataset by randomly selecting 9 individuals for training, and the remaining 3 individuals for testing.

7.4.2 Implementation Details

In both dataset, we randomly sample 40 frames from the sequences in each iterations. All vertex-value are normalized to mean zero and unit variance. The Spiral autoencoder is trained for 250 epochs, using Adam Optimizer [195] with a learning rate of 0.0002 and a momentum of 0.95. The diffusion model is trained for 100k steps, using Adam Optimizer [195] with a learning rate of 0.0002 and a momentum of 0.95. The implementation of the model is conducted using the Pytorch Framework, building upon the implementation from [8]. Training SAE on the MUG dataset took approximately 10 hours, while training DM took about 6 hours, both on an Nvidia GeForce RTX 4080 with 16 GB of memory.

7.4.3 Experimental Results

Fig.7.4 presents examples of sequences generated by our proposed model trained on MUG dataset, each of which displays one of the six basic facial expressions for different identities. The resulting sequences show continuous, smooth, and natural transitions between facial expressions, demonstrating that the expressions can be effectively controlled based on the input labels. The generated videos maintain the same details from the input image, such identity, beards, hair style, and clothes. This proves that our model is effective for the six basic facial expressions generation.

Fig.7.5 presents examples of sequences generated by our proposed model trained on COMA dataset, each of which displays one of twelve extreme and non-symmetric facial expressions for different identities. These results prove that LM3DM can handle more complex tasks than generating the six basic facial expressions, being able to handle twelve classes of more expressive and extreme expressions.

In addition, we perform an experiment to prove the variance of our model in generating the same expression, avoiding mode collapse. We generate multiple sequences of the same subject performing the same expression. This experiment is conducted on MUG and COMA datasets. Fig.7.6 showcases examples of the generated sequences. The generated sequences display the same subject and expression but different output sequences. These sequences are different in terms of temporal evolution and expression intensity, indicating that the model learns variance in generating a broader range of facial expressions. Additionally, the differing temporal evolutions in the generated videos demonstrate the model’s ability to simulate the dynamic nature of facial expressions over time. Some sequences show a gradual intensification of the expression, while others depict quicker transitions. This variability is critical for producing realistic and refined outputs, as it mimics how emotions are expressed differently by various individuals and even by the same individual in different contexts.

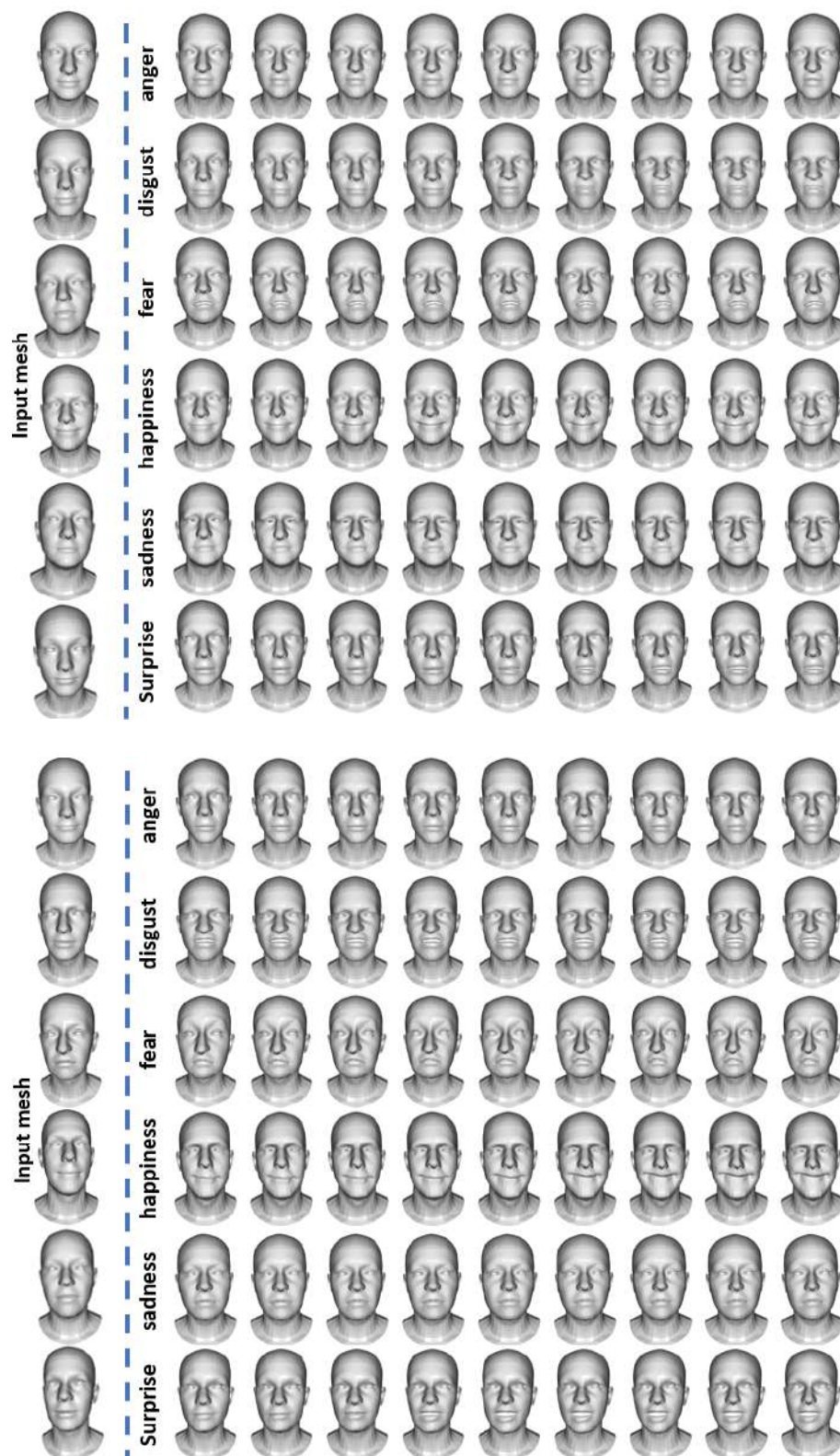


Figure 7.4: Examples of generated sequences for the same subject performing the same expression, in each two lines.

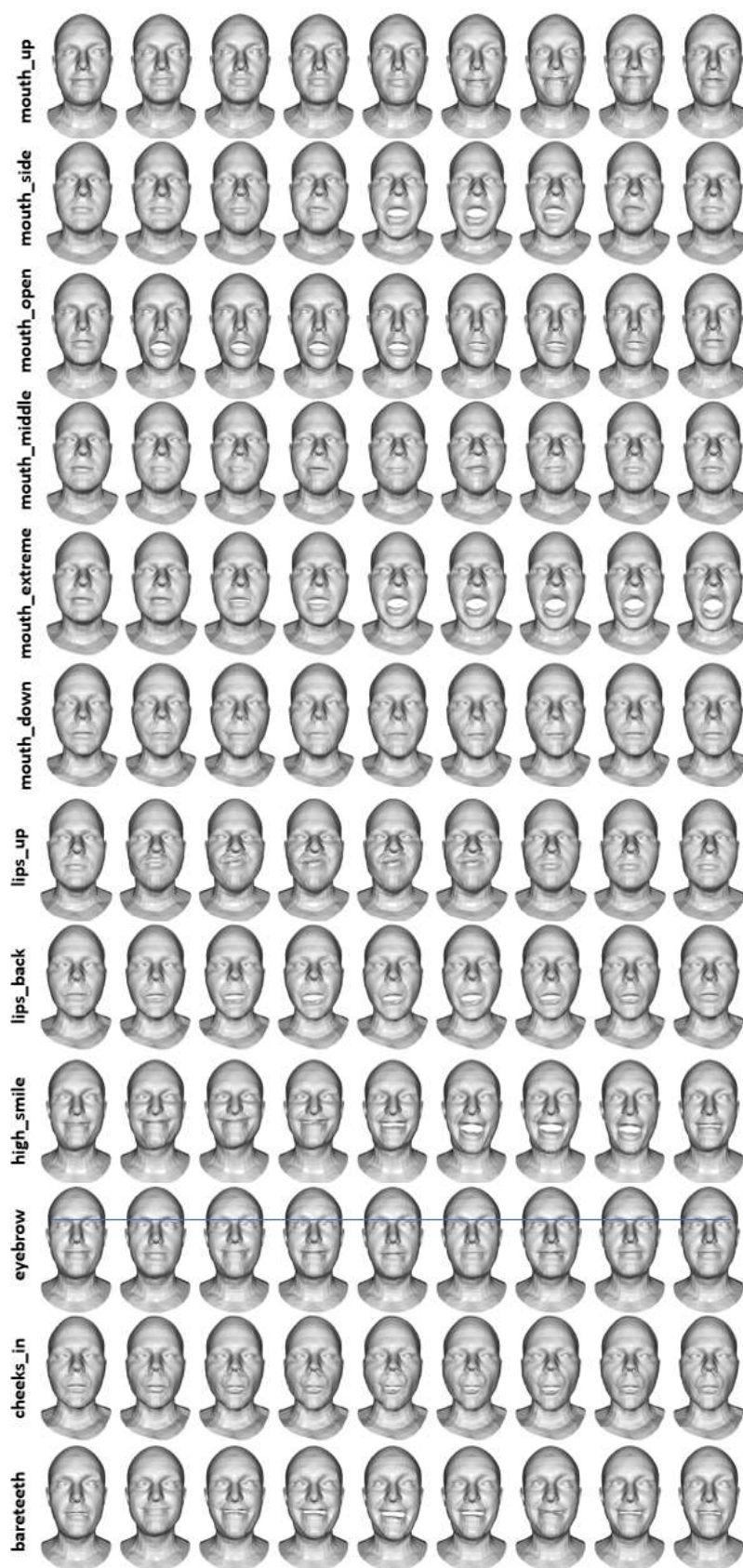


Figure 7.5: Examples generated by LM3DM trained on COMA.

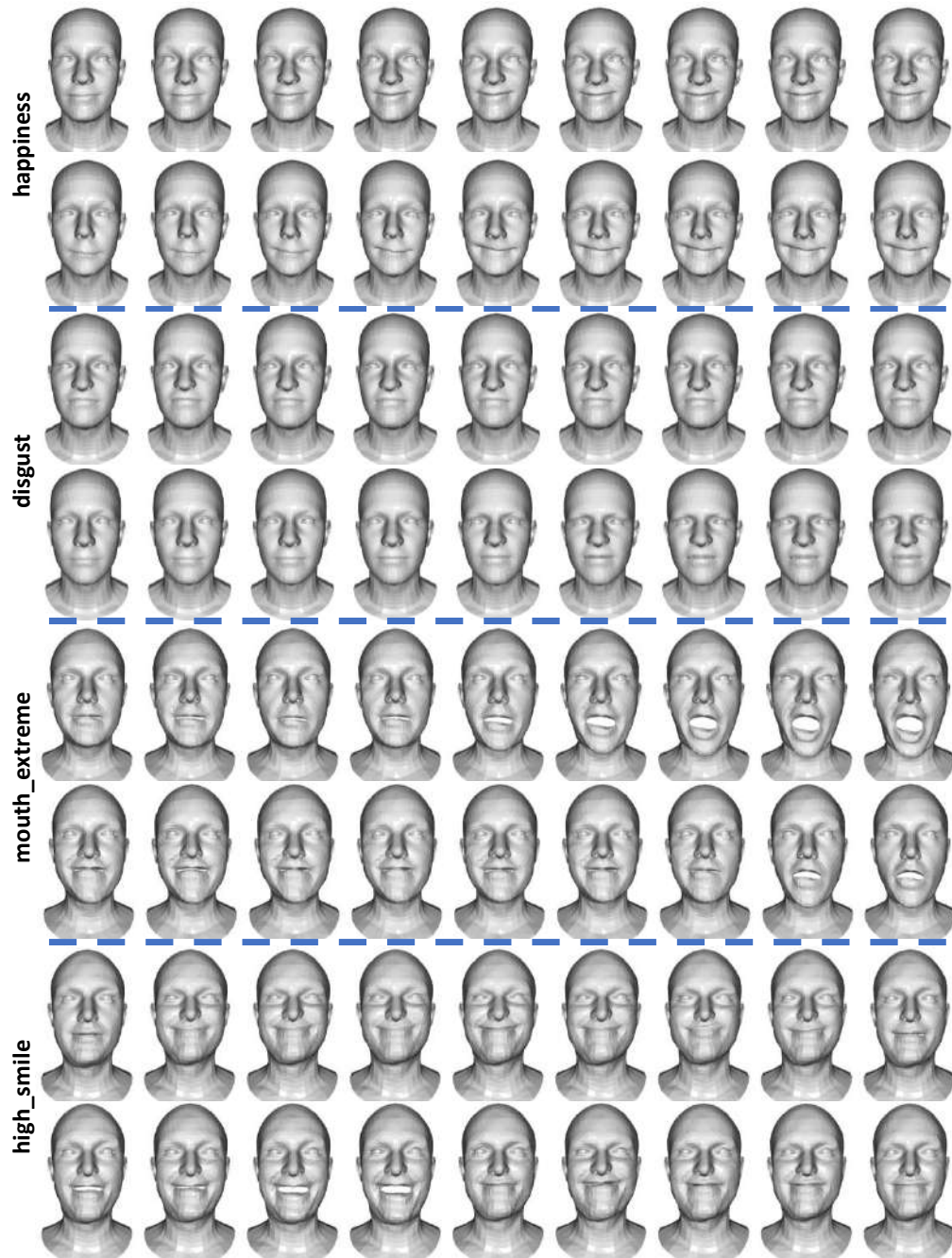


Figure 7.6: Examples of the same expression generated by LM3DM.

7.5 Conclusion & Perspectives

In this work, we introduced a novel Latent-Motion 3D Diffusion Model (LM3DM) for dynamic facial expression generation, using a single input neutral mesh and a target facial motion category as input. The proposed model directly works on the latent space of the mesh generator, specifically on the motion space. By injecting diffusion stochasticity into the motion embedding space, our model generates diverse expressions while preserving the fine details of the input identity. This is achieved by decoupling the identity and motion embedding spaces, then applying diffusion solely to learn the motion latent space.

LM3DM is evaluated on the MUG and COMA databases. The preliminary results highlight the effectiveness of the proposed model in generating facial expressions with variant, smooth, and natural motions. While the proposed model shows potential, it is important to note that our model is in its developmental stage and exhibits certain shortcomings. These limitations include some artifacts and errors in the generated sequences and occasionally very low intensity expressions. Future work will focus on conducting quantitative evaluations and comparison with state-of-the-art, to assess the value of our work. Furthermore, we aim to extend the model beyond facial expressions to encompass human action generation.

Conclusions and Perspectives

Conclusion

Facial expression generation has become an intriguing and challenging task in computer vision, gaining a lot of attention in recent years, specifically with the appearance of large models such as I2VGen-XL [272], VideoComposer [273], and VideoCrafter1 [274]. Although these models have achieved impressive results, they come with important trade-offs. They are characterized by their depth reaching billions of parameters, which increases complexity. This latter requires high-end computational resources and a considerable amount of video data, often amounting to millions of samples. Most of this data remains unlabeled due to the time-consuming nature of the labeling process, which leads to the creation of generalized models. These models are prone to hallucinations and struggle to perform well on specific tasks, as their training data rarely covers all potential application areas. As a result, customizing these models for specific tasks in private or smaller-scale environments becomes challenging due to the lack of access to large-scale datasets and the computational power necessary for both training and inference.

This highlights the importance and value of smaller models like VideoGAN [123], MoCoGAN [128], ImaGINnator [127], and LFDM [8]. These models trained and deployed on low to mid-range machines and smaller datasets, of only hundreds of videos. As a result, they are efficient and easily customizable for specific tasks. However, these models come with their own set of limitations, such as low resolution and quality, as well as limited complexity. Their simplified architectures often result in models that struggle to capture the complex representations needed for more complex tasks, which can lead to reduced performance in real-world applications.

In this work, we focus on enhancing low-end models to improve their performance in facial expression generation. We started by introducing FEV-GAN (Face Expression Video GAN), which generates videos of basic facial expressions from a single neutral image while preserving identity. This model addresses low quality and identity preservation issues in Spatio-temporal Convolutional GANs by integrating identity and spatial features from VGG-FACE, resulting in improved video quality compared to state-of-the-art methods. However, the generated videos remain low resolution. To tackle this limitation, we introduced FacEnhance, that utilizes diffusion to augment the resolution and quality of generated videos while incorporating background details. This method strikes a balance between computational efficiency and desirable image quality, demonstrating significant enhancements in low-resolution outputs from [123, 127, 136, 209] to higher resolution and

quality. Next, we propose the Latent Motion Diffusion Model (LMDM) for both facial expression and human action generation. LMDM leverages latent diffusion techniques to learn complex latent spaces and introduce controlled stochasticity to diversify generated content. By separating the embedding space into content and motion domains and applying diffusion exclusively to the motion latent space, LMDM ensures varied motions while preserving content details.

Recognizing the growing demand for 3D facial expression generation, particularly within the expanding metaverse, we further extend our work into the 3D domain. We start with SpATr (Spiral Autoencoder and Transformer Network) for facial expression and human action recognition, effectively decoupling spatial and temporal features, through lightweight spiral convolutions for spatial information and self-attention for temporal dependencies. Building on this recognition framework, we present LM3DM (Latent-Motion 3D Diffusion Model), which integrates a spiral autoencoder designed for fixed-topology mesh sequences to learn effective data embeddings. LM3DM then directly applies a diffusion model within the latent space of the mesh generator, capturing complex expression dynamics. Fig.7.7 illustrates the evolution of our works in facial expression generation, starting from a comparison with VGAN and ImaGINator, and advancing through our proposed models organized chronologically.

While these models offer computational efficiency and require less memory, they sacrifice video quality. The limited capacity and complexity of these architectures restrict their ability to capture high variability and complex dynamics in video data, making it challenging to learn long-term dependencies and often resulting in shorter generated videos. Additionally, despite efforts to increase the resolution of generated videos, both quality and resolution remain limited, and pushing beyond these constraints leads to substantial computational requirements. Furthermore, there is a notable lack of diversity in generated samples. We attempted to address this issue by employing diffusion models to enhance temporal evolution, but further improvements in this area are still necessary. Small models also face challenges in their training procedures, as they tend to be less stable, more sensitive to noise, and less robust compared to larger models, primarily due to the limited data available for training. Addressing these limitations will be crucial for advancing the effectiveness of low-end models in video generation tasks.

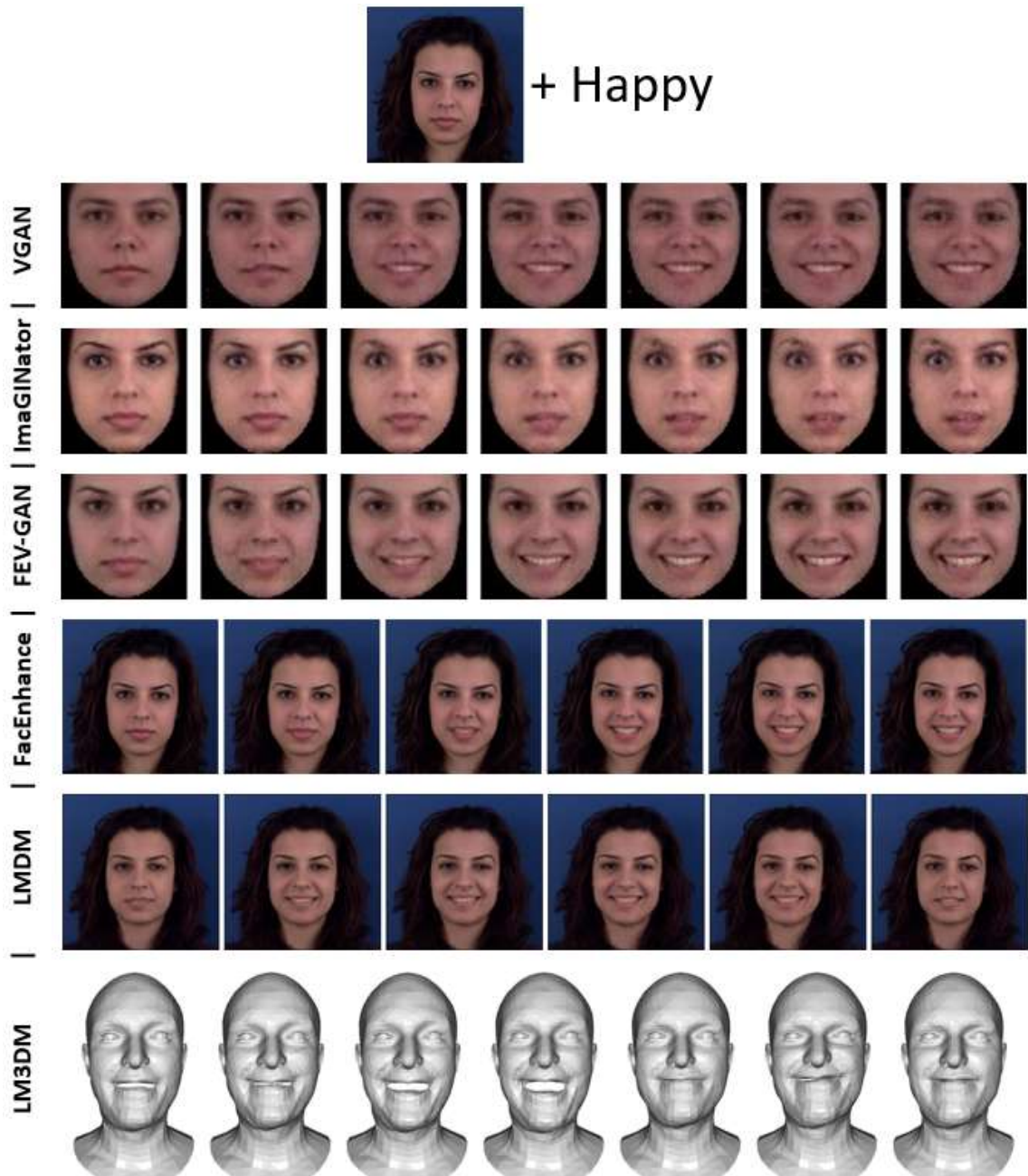


Figure 7.7: Comparison of the facial expression generation contributions proposed in this thesis, and baselines.

Perspectives and Future Works

The research and experiments carried out in this thesis provide a foundation for many more directions to be explored and open up several promising avenues for future work.

- Regarding 2D facial expression synthesis, this thesis focused on frontal facial expression generation. However, the motion generated by these approaches can be extended to non-frontal facial expressions. Future work could explore adapting these models to changing poses and conditions, which would enable the generation of annotated, realistic expressions that mimic those observed in wild data. This approach could help remedy the challenge of lack of training data in facial expression recognition systems.

- Future work may also focus on enhancing the efficiency of facial expression generation systems, meaning to improve resolution and model complexity without compromising computational efficiency. One potential direction involves exploring model quantization techniques, such as [275, 276, 277], to reduce the computational and memory requirement while maintaining decent performance. Another promising approach is leveraging prompt learning [278, 279], where large video models are hosted in the cloud, and a local model learns to generate suitable prompts for the target expressions. Additionally, knowledge distillation techniques [280, 281] could be explored to train lightweight models that retain the knowledge and capabilities of larger, more complex models, therefore improving both efficiency and performance.

- In this thesis, we introduced LM3DM, an extension of LMDM from 2D videos to 3D mesh sequences. This work provides a starting point for future research in 3D facial expression generation, a highly challenging task. One of the primary challenges is the unstructured nature of mesh data, which requires handling higher dimensionality and complexity, therefore requiring significant computational resources. Additionally, acquiring high-quality 3D facial data is resource-intensive compared to 2D image data, limiting the availability of training datasets. To address this, inspiring works such as spiralConv [7], SpiralNet++ [282], and spectral [283], have shown progress in managing mesh structures, leading to methods, such as [168, 169, 172, 173], for 3D facial expression generation. Despite these advancements, the field is not well explored, with many open research questions, making it a fertile area for future studies.

- Another promising direction for future work includes extending our proposed methods to human action generation, a topic that attracts much attention [284]. This extension is based on the common representation methods between the human body and face. They can be represented in different forms, such as pixel-wise representations, 3D meshes/point-clouds/voxels, or geometric keypoints (landmarks for the face and skeleton

for the body). We have already started exploring this in our models, LMDM and SpATr, and there is potential to apply approaches such as FacEnhance and LM3DM to 2D/3D human action generation.

- In the context of our work in the 3D domain, both for recognition and synthesis, we focused on fixed-topology data, which requires mesh registration and limits the applicability of the proposed methods. An important future direction is to explore unregistered 3D scan generation, as it offers significant advantages, including the availability of more training data, flexibility with diverse datasets without the need for costly registration, easier generalization to unseen faces, and the potential to extend to non-human faces (e.g., animals, video game characters). While there has been progress in unregistered 3D facial expression recognition, 3D facial expression generation remains unexplored. To the best of our knowledge, Nocentini et al. [285, 286] proposed a model that explores generating unregistered scans, but their work focuses on generating talking faces based on audio snippets, without integrating facial expressions or their temporal sequence evolution. Their method generates single-frame meshes guided by corresponding audio snippets, but it does not address the dynamic nature of facial expressions.

Inspired by this gap, we aim to focus on developing 3D unregistered facial expression generation models. A potential approach would be to generate motion in a motion embedding space, similar to our method in LM3DM, and then use these generated embeddings to guide the manipulation of unregistered scans, following the approach in Nocentini et al.’s work. This would open new possibilities for more flexible and generalized 3D facial expression generation models.

References

- [1] Marian Bartlett, Gwen Littlewort, Esra Vural, Kang Lee, Mujdat Cetin, Aytul Ercil, and Javier Movellan. Data mining spontaneous facial behavior with automatic expression coding. In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction: COST Action 2102 International Conference, Patras, Greece, October 29-31, 2007. Revised Papers*, pages 1–20. Springer, 2008.
- [2] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.
- [3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- [4] Michail Tarasiou, Stylianos Moschoglou, Jiankang Deng, and Stefanos Zafeiriou. Improving face generation quality and prompt following with synthetic captions. *arXiv preprint arXiv:2405.10864*, 2024.
- [5] Giuseppe Lisanti and Nico Giambi. Conditioning diffusion models via attributes and semantic masks for face generation. *Computer Vision and Image Understanding*, 244: 104026, 2024.
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [7] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7213–7222, 2019.
- [8] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023.
- [9] Charles Darwin. *The Expression of the Emotions in Man and Animals*. John Murray, 1872.
- [10] Albert Mehrabian and Morton Wiener. Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, 6(1):109–114, 1967. doi: 10.1037/h0024532.
- [11] Albert Mehrabian and Susan R. Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, 31(3):248–252, Jun 1967. doi: 10.1037/h0024648.
- [12] Paul Watzlawick, Janet Beavin Bavelas, and Don D. Jackson. Some tentative axioms of communication. In Paul Watzlawick, Janet Beavin Bavelas, and Don D. Jackson, editors, *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes*, pages 48–71. W. W. Norton & Co, New York, 1967. Accessed from <https://pdfs.semanticscholar.org/e998/92445f82>.

- [13] Morton Wiener, Shirley Devoe, Stephen Rubinow, and Jesse Geller. Nonverbal behavior and nonverbal communication. *Psychological Review*, 79(3):185–214, 1972. doi: 10.1037/h0032710. URL <https://doi.org/10.1037/h0032710>.
- [14] Maja Pantic. Facial expression recognition. In *Encyclopedia of biometrics*, pages 400–406. Springer, 2009.
- [15] Paul Ekman and Dacher Keltner. Universal facial expressions of emotion. *California mental health research digest*, 8(4):151–158, 1970.
- [16] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.
- [17] Emily B Prince, Katherine B Martin, Daniel S Messinger, and M Allen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1, 2015.
- [18] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221, 2007.
- [19] Alfonso Semeraro, Salvatore Vilella, and Giancarlo Ruffo. Pyplutchik: Visualising and comparing emotion-annotated corpora. *Plos one*, 16(9):e0256503, 2021.
- [20] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.
- [21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*, volume 1. MIT press, 2016.
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [23] AG Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [24] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.
- [25] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.
- [26] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
- [27] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.
- [28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [29] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.
- [30] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

- [31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [34] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [35] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [36] Junyoung Chung. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [37] Mario Botsch, Leif Kobbelt, Mark Pauly, Pierre Alliez, and Bruno Lévy. *Polygon mesh processing*. CRC press, 2010.
- [38] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. pages 851–866, 2023.
- [39] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6): 194–1, 2017.
- [40] Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015.
- [41] Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.
- [42] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5115–5124, 2017.
- [43] Fu-Song Hsu, Wei-Yang Lin, and Tzu-Wei Tsai. Facial expression recognition using bag of distances. *Multimedia tools and applications*, 73:309–326, 2014.
- [44] T Ahonen. Face recognition with local binary patterns. ECCV, 2004.
- [45] Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE transactions on image processing*, 19(6):1657–1663, 2010.
- [46] Wei-Lun Chao, Jian-Jiun Ding, and Jun-Zuo Liu. Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection. *Signal Processing*, 117:1–10, 2015.
- [47] Hamit Soyel and Hasan Demirel. Improved sift matching for pose robust facial expression recognition. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 585–590. IEEE, 2011.

- [48] Yong Shi, Zhao Lv, Ning Bi, and Chao Zhang. An improved sift algorithm for robust emotion recognition under various face poses and illuminations. *Neural Computing and Applications*, 32:9267–9281, 2020.
- [49] Pierluigi Carcagnì, Marco Del Coco, Marco Leo, and Cosimo Distanto. Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus*, 4(1):645, 2015.
- [50] Priyanka N Maraskolhe and AS Bhalchandra. Analysis of facial expression recognition using histogram of oriented gradient (hog). In *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1007–1011. IEEE, 2019.
- [51] Kranthi Kumar Talluri, Marc-André Fiedler, and Ayoub Al-Hamadi. Deep 3d convolutional neural network for facial micro-expression analysis from video images. *Applied Sciences*, 12(21):11078, 2022.
- [52] Sai Prasanna Teja Reddy, Surya Teja Karri, Shiv Ram Dubey, and Snehasis Mukherjee. Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.
- [53] Behzad Hasani and Mohammad H Mahoor. Facial expression recognition using enhanced deep 3d convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 30–40, 2017.
- [54] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 445–450, 2016.
- [55] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, Yuan Zong, and Ning Sun. Multi-clue fusion for emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 458–463, 2016.
- [56] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 467–474, 2015.
- [57] Andrey V Savchenko and Anna P Sidorova. Emotiefnet and temporal convolutional networks in video-based facial expression recognition and action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4849–4859, 2024.
- [58] Fuyan Ma, Bin Sun, and Shutao Li. Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [59] Yuanyuan Liu, Wenbin Wang, Chuanxu Feng, Haoyu Zhang, Zhe Chen, and Yibing Zhan. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138:109368, 2023.
- [60] Daizong Liu, Hongting Zhang, and Pan Zhou. Video-based facial expression recognition using graph convolutional networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 607–614. IEEE, 2021.
- [61] Elham S Salama, Reda A El-Khoribi, Mahmoud E Shoman, and Mohamed A Wahby Shalaby. A 3d-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition. *Egyptian Informatics Journal*, 22(2):167–176, 2021.

- [62] Ying Tan, Zhe Sun, Feng Duan, Jordi Solé-Casals, and Cesar F Caiafa. A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomedical Signal Processing and Control*, 70:103029, 2021.
- [63] Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15445–15459, 2023.
- [64] Wei Zhang, Youmei Zhang, Lin Ma, Jingwei Guan, and Shijie Gong. Multimodal learning for facial expression recognition. *Pattern Recognition*, 48(10):3191–3202, 2015.
- [65] Sima Taheri, Pavan Turaga, and Rama Chellappa. Towards view-invariant expression analysis using analytic shape manifolds. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 306–313. IEEE, 2011.
- [66] Anis Kacem, Mohamed Daoudi, Boulbaba Ben Amor, and Juan Carlos Alvarez-Paiva. A novel space-time representation on the positive semidefinite cone for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3180–3189, 2017.
- [67] Naima Otberdout, Anis Kacem, Mohamed Daoudi, Lahoucine Ballihi, and Stefano Berretti. Automatic analysis of facial expressions based on deep covariance trajectories. *IEEE transactions on neural networks and learning systems*, 31(10):3892–3905, 2019.
- [68] Yunfu Sun, Michael Reale, and Lijun Yin. Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition. In *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8. IEEE, 2008.
- [69] Yunfu Sun and Lijun Yin. Facial expression recognition based on 3d dynamic range model sequences. In *European Conference on Computer Vision*, pages 58–71. Springer, 2008.
- [70] Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. Automatic facial expression recognition in real-time from dynamic sequences of 3d face scans. *The Visual Computer*, 29:1333–1350, 2013.
- [71] Viet Le, Hao Tang, and Thomas S Huang. Expression recognition from 3d dynamic faces using robust spatio-temporal shape features. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 414–421. IEEE, 2011.
- [72] Tianhong Fang, Xi Zhao, Shishir K Shah, and Ioannis A Kakadiaris. 4d facial expression recognition. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 1594–1601. IEEE, 2011.
- [73] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Daniel Rueckert. A dynamic approach to the recognition of 3d facial expressions and their temporal models. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 406–413. IEEE, 2011.
- [74] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Daniel Rueckert. Recognition of 3d facial expression dynamics. *Image and Vision Computing*, 30(10):762–773, 2012.

- [75] Hassen Drira, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava, and Stefano Berretti. 3d dynamic expression recognition based on a novel deformation vector field and random forest. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1104–1107. IEEE, 2012.
- [76] Mohamed Daoudi, Hassen Drira, Boulbaba Ben Amor, and Stefano Berretti. A dynamic geometry-based approach for 4d facial expressions recognition. In *European Workshop on Visual Information Processing (EUVIP)*, pages 280–284. IEEE, 2013.
- [77] Boulbaba Ben Amor, Hassen Drira, Stefano Berretti, Mohamed Daoudi, and Anuj Srivastava. 4-d facial expression recognition by learning geometric deformations. *IEEE transactions on cybernetics*, 44(12):2443–2457, 2014.
- [78] Michael Reale, Xing Zhang, and Lijun Yin. Nebula feature: A space-time feature for posed and spontaneous 4d facial behavior analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013.
- [79] Mingda Xue, Ajmal Mian, Wen Liu, and Liang Li. Automatic 4d facial expression recognition using dct features. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 199–206. IEEE, 2015.
- [80] Mingda Xue, Ajmal Mian, Xiaojun Duan, and Wen Liu. Learning interpretable expression-sensitive features for 3d dynamic facial expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019.
- [81] Qingkai Zhen, Di Huang, Yunhong Wang, and Liming Chen. Muscular movement model-based automatic 3d/4d facial expression recognition. *IEEE Transactions on Multimedia*, 18(7):1438–1450, 2016.
- [82] Iman Abbasnejad, Sridha Sridharan, Dung Nguyen, Simon Denman, Clinton Fookes, and Simon Lucey. Using synthetic data to improve facial expression analysis with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1609–1618, 2017.
- [83] Wei Li, Di Huang, Hui Li, and Yunhong Wang. Automatic 4d facial expression recognition using dynamic geometrical image network. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 24–30. IEEE, 2018.
- [84] Bilal Taha, Munawar Hayat, Stefano Berretti, Dimitrios Hatzinakos, and Naoufel Werghi. Learned 3d shape representations using fused geometrically augmented images: Application to facial expression and action unit detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2900–2916, 2020.
- [85] Muzammil Behzad, Nhat Vo, Xiaobai Li, and Guoying Zhao. Automatic 4d facial expression recognition via collaborative cross-domain dynamic image network. *arXiv preprint arXiv:1905.02319*, 2019.
- [86] Muzammil Behzad, Nhat Vo, Xiaobai Li, and Guoying Zhao. Landmarks-assisted collaborative deep framework for automatic 4d facial expression recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 1–5. IEEE, 2020.
- [87] Mehrdad Behzad, Nguyen Vo, Xiaoming Li, and Guoying Zhao. Towards reading beyond faces for sparsity-aware 3d/4d affect recognition. *Neurocomputing*, 458: 297–307, 2021.
- [88] Frédéric Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H Salesin. Synthesizing realistic facial expressions from photographs. In *ACM SIGGRAPH 2006 Courses*, pages 19–es. 2006.

- [89] Jia-Shing Sheu, Tsu-Shien Hsieh, and Ho-Nien Shou. Automatic generation of facial expression using triangular geometric deformation. *Journal of applied research and technology*, 12(6):1115–1130, 2014.
- [90] Amaryllis Raouzaïou, Nicolas Tsapatsoulis, Kostas Karpouzis, and Stefanos Kollias. Parameterized facial expression synthesis based on mpeg-4. *EURASIP Journal on Advances in Signal Processing*, 2002(10):521048, 2002.
- [91] A Wojdeł and Léon JM Rothkrantz. Parametric generation of facial expressions based on facs. In *Computer Graphics Forum*, volume 24, pages 743–757. Wiley Online Library, 2005.
- [92] Narendra M Patel and Mukesh Zaveri. Parametric facial expression synthesis and animation. *International Journal of Computer Applications*, 3(4):34–40, 2010.
- [93] Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. *ACM SIGGRAPH computer graphics*, 26(2):35–42, 1992.
- [94] Chuan-Kai Yang and Wei-Ting Chiang. An interactive facial expression generation system. *Multimedia Tools and Applications*, 40:41–60, 2008.
- [95] Richard J Harris, Andrew W Young, and Timothy J Andrews. Morphing between expressions dissociates continuous from categorical representations of facial expression in the human brain (2012). In *Facial Expression Recognition*, pages 242–261. Psychology Press, 2016.
- [96] Matthew D Zeiler, Graham W Taylor, Leonid Sigal, Iain Matthews, and Rob Fergus. Facial expression transfer with input-output temporal restricted boltzmann machines. In *Advances in neural information processing systems*, pages 1629–1637, 2011.
- [97] Maryam Sabzevari, Saeed Toosizadeh, Saeed Rahati Quchani, and Vahid Abrishami. A fast and accurate facial expression synthesis system for color face images using face graph and deep belief network. In *2010 International Conference on Electronics and Information Engineering*, volume 2, pages V2–354. IEEE, 2010.
- [98] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.
- [99] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *CoRR*, abs/1511.05644, 2015. URL <http://arxiv.org/abs/1511.05644>.
- [100] J Rafid Siddiqui. Fexgan-meta: Facial expression generation with meta humans. *arXiv preprint arXiv:2203.05975*, 2022.
- [101] Yuqian Zhou and Bertram Emil Shi. Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In *2017 seventh international conference on affective computing and intelligent interaction (ACII)*, pages 370–376. IEEE, 2017.
- [102] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.
- [103] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

- [104] Xueping Wang, Weixin Li, Guodong Mu, Di Huang, and Yunhong Wang. Facial expression synthesis by u-net conditional generative adversarial networks. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 283–290, 2018.
- [105] Arbish Akram and Nazar Khan. Us-gan: On the importance of ultimate skip connection for facial expression synthesis. *arXiv preprint arXiv:2112.13002*, 2021.
- [106] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [107] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018.
- [108] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5021–5030, 2020.
- [109] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 627–635, 2018.
- [110] Fengchun Qiao, Nai-Ming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. Geometry-contrastive generative adversarial network for facial expression synthesis. *CoRR*, abs/1802.01822, 2018. URL <https://arxiv.org/abs/1802.01822>.
- [111] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 128(5):1455–1484, 2020.
- [112] Wenjun Bai, Changqin Quan, and Zhi-Wei Luo. Data-driven dimensional expression generation via encapsulated variational auto-encoders. *Cognitive Computation*, pages 1–13, 2022.
- [113] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10893–10900, 2020.
- [114] Junshu Tang, Zhiwen Shao, and Lizhuang Ma. Eggan: Learning latent space for fine-grained expression manipulation. *IEEE MultiMedia*, 28(3):42–51, 2021.
- [115] Dejian Li, Wenqian Qi, and Shouqian Sun. Facial landmarks and expression label guided photorealistic facial expression synthesis. *IEEE Access*, 9:56292–56300, 2021.
- [116] Xiaoguang Tu, Yingtian Zou, Jian Zhao, Wenjie Ai, Jian Dong, Yuan Yao, Zhikang Wang, Guodong Guo, Zhifeng Li, Wei Liu, et al. Image-to-video generation via 3d facial dynamics. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1805–1819, 2021.
- [117] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. Vdsm: Unsupervised video disentanglement with state-space modeling and deep mixtures of experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8176–8186, 2021.

- [118] Yongjian Zhao, Xinyan Cao, Siqi Liu, Jinming Che, Wei Ren, Jian Cao, and Jinlong Lin. A facial expression transfer method based on 3dmm and diffusion models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3145–3149. IEEE, 2024.
- [119] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018.
- [120] Jun Ling, Han Xue, Li Song, Shuhui Yang, Rong Xie, and Xiao Gu. Toward fine-grained facial expression manipulation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 37–53. Springer, 2020.
- [121] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. Emotional facial expression transfer from a single image via generative adversarial nets. *Computer Animation and Virtual Worlds*, 29(3-4):e1819, 2018.
- [122] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019.
- [123] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in neural information processing systems*, pages 613–621, 2016.
- [124] Richa Sharma, Manoj Sharma, Ankit Shukla, and Santanu Chaudhury. Conditional deep 3d-convolutional generative adversarial nets for rgb-d generation. *Mathematical Problems in Engineering*, 2021(1):8358314, 2021.
- [125] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, pages 2830–2839, 2017.
- [126] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5264–5273, 2020.
- [127] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1160–1169, 2020.
- [128] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.
- [129] Sonam Gupta, Arti Keshari, and Sukhendu Das. Rv-gan: Recurrent gan for unconditional video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2024–2033, 2022.
- [130] Wei Wang, Xavier Alameda-Pineda, Dan Xu, Pascal Fua, Elisa Ricci, and Nicu Sebe. Every smile is unique: Landmark-guided diverse smile generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7083–7092, 2018.
- [131] Lijie Fan, Wenbing Huang, Chuang Gan, Junzhou Huang, and Boqing Gong. Controllable image-to-video translation: A case study on facial expression generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3510–3517, 2019.

- [132] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.
- [133] Tao Yang, Peiran Ren, Xuansong Xie, Xian-Sheng Hua, and Lei Zhang. Beyond a video frame interpolator: A space decoupled learning approach to continuous image transition. In *European Conference on Computer Vision*, pages 738–755. Springer, 2022.
- [134] Chuin Hong Yap, Moi Hoon Yap, Adrian K Davison, and Ryan Cunningham. Self-supervised spontaneous latent-based facial expression sequence generation. *IEEE Open Journal of Signal Processing*, 4:304–312, 2023.
- [135] I Sutskever. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.
- [136] Naima Otberdout, Mohamed Daoudi, Anis Kacem, Lahoucine Ballihi, and Stefano Berretti. Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):848–863, 2020.
- [137] Sandeep Manandhar and Auguste Genovesio. An invertible video gan in a conditional setting. 2023.
- [138] Bitan Azari and Angelica Lim. Emostyle: One-shot facial expression editing using continuous emotion parameters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6385–6394, 2024.
- [139] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023.
- [140] Hamza Bouzid and Lahoucine Ballihi. Facenhance: Facial expression enhancing with recurrent ddpm. *arXiv preprint arXiv:2406.09040*, 2024.
- [141] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.
- [142] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.
- [143] Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, and Xiaofei Wu. Co-speech gesture video generation via motion-decoupled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2263–2273, 2024.
- [144] Cuifeng Shen, Yulu Gan, Chen Chen, Xiongwei Zhu, Lele Cheng, Tingting Gao, and Jinzhi Wang. Decouple content and motion for conditional image-to-video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4757–4765, 2024.
- [145] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023.
- [146] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016.

- [147] Alan Brunton, Timo Bolkart, and Stefanie Wuhler. Multilinear wavelets: A statistical shape space for human faces. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*, pages 297–312. Springer, 2014.
- [148] Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. Dictionary learning based 3d morphable model construction for face recognition with varying expression and pose. In *2015 International Conference on 3D Vision*, pages 509–517. IEEE, 2015.
- [149] Marcel Lüthi, Thomas Gerig, Christoph Jud, and Thomas Vetter. Gaussian process morphable models. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1860–1873, 2017.
- [150] Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus Magnor, and Christian Theobalt. Sparse localized deformation components. *ACM Transactions on Graphics (TOG)*, 32(6):1–10, 2013.
- [151] Claudio Ferrari, Stefano Berretti, Pietro Pala, and Alberto Del Bimbo. A sparse and locally coherent morphable face model for dense semantic correspondence across heterogeneous 3d faces. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6667–6682, 2021.
- [152] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.
- [153] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 704–720, 2018.
- [154] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020.
- [155] Isaak Lim, Alexander Dielen, Marcel Campen, and Leif Kobbelt. A simple approach to intrinsic correspondence learning on unstructured 3d meshes. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.
- [156] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhler, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5):1–38, 2020.
- [157] Claudio Ferrari, Stefano Berretti, Pietro Pala, and Alberto Del Bimbo. Rendering realistic subject-dependent expression images by learning 3dmm deformation coefficients. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.
- [158] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):78–92, 2017.
- [159] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022.
- [160] Xiyi Chen, Marko Mihajlovic, Shaofei Wang, Sergey Prokudin, and Siyu Tang. Morphable diffusion: 3d-consistent diffusion for single-image avatar creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10359–10370, 2024.

- [161] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017.
- [162] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. Expnet: Landmark-free, deep, 3d facial expressions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 122–129. IEEE, 2018.
- [163] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (ToG)*, 36(4):1–12, 2017.
- [164] Xiangyu Liang, Wenlin Zhuang, Tianyong Wang, Guangxing Geng, Guangyue Geng, Haifeng Xia, and Siyu Xia. Cstalk: Correlation supervised speech-driven 3d emotional facial animation generation. *arXiv preprint arXiv:2404.18604*, 2024.
- [165] Kim Sung-Bin, Lee Hyun, Da Hye Hong, Suekyeong Nam, Janghoon Ju, and Tae-Hyun Oh. Laughtalk: Expressive 3d talking head generation with laughter. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6404–6413, 2024.
- [166] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–9, 2024.
- [167] Dan Zeng, Han Liu, Hui Lin, and Shiming Ge. Talking face generation with expression-tailored generative adversarial network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1716–1724, 2020.
- [168] Rolandos Alexandros Potamias, Jiali Zheng, Stylianos Ploumpis, Giorgos Bouritsas, Evangelos Ververas, and Stefanos Zafeiriou. Learning to generate customized dynamic 3d facial expressions. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 278–294. Springer, 2020.
- [169] Kaifeng Zou, Boyang Yu, and Hyewon Seo. 3d facial expression generator based on transformer vae. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2550–2554. IEEE, 2023.
- [170] Naima Otberdout, Claudio Ferrari, Mohamed Daoudi, Stefano Berretti, and Alberto Del Bimbo. 3d to 4d facial expressions generation guided by landmarks. *arXiv preprint arXiv:2105.07463*, 2021.
- [171] Naima Otberdout, Claudio Ferrari, Mohamed Daoudi, Stefano Berretti, and Alberto Del Bimbo. Generating multiple 4d expression transitions by learning face landmark trajectories. *IEEE Transactions on Affective Computing*, 15(2):566–578, 2023.
- [172] Naima Otberdout, Claudio Ferrari, Mohamed Daoudi, Stefano Berretti, and Alberto Del Bimbo. Sparse to dense dynamic 3d facial expression generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20385–20394, 2022.
- [173] Kaifeng Zou, Sylvain Faisan, Boyang Yu, Sébastien Valette, and Hyewon Seo. 4d facial expression diffusion model. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.
- [174] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

- [175] Michael J Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba, and Julien Budynek. The japanese female facial expression (jaffe) database. In *Proceedings of third international conference on automatic face and gesture recognition*, pages 14–16, 1998.
- [176] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.
- [177] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.
- [178] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4. IEEE, 2010.
- [179] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.
- [180] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013.
- [181] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2106–2112. IEEE, 2011.
- [182] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 423–426, 2015.
- [183] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal ‘in-the-wild’ challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017.
- [184] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019.
- [185] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006.
- [186] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE, 2013.
- [187] Dawood Al Chanti and Alice Caplier. Deep learning for spatio-temporal modeling of dynamic spontaneous emotions. *IEEE Transactions on Affective Computing*, 12(2):363–376, 2018.

- [188] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.
- [189] Ahmad Ali, Yanmin Zhu, and Muhammad Zakarya. Exploiting dynamic spatio-temporal correlations for citywide traffic flow prediction using attention based neural networks. *Information Sciences*, 577:852–870, 2021.
- [190] Ahmad Ali, Yanmin Zhu, and Muhammad Zakarya. Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. *Neural networks*, 145:233–247, 2022.
- [191] Yunseok Jang, Gunhee Kim, and Yale Song. Video prediction with appearance and motion conditions. In *International Conference on Machine Learning*, pages 2225–2234. PMLR, 2018.
- [192] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015. ISBN 1-901725-53-7. doi: 10.5244/C.29.41. URL <https://dx.doi.org/10.5244/C.29.41>.
- [193] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.
- [194] Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.
- [195] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.
- [196] Abien Fred Agarap. Deep learning using rectified linear units (relu). *CoRR*, abs/1803.08375, 2018. URL <http://arxiv.org/abs/1803.08375>.
- [197] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. URL <http://arxiv.org/abs/1502.03167>.
- [198] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015. URL <http://arxiv.org/abs/1505.00853>.
- [199] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 387–403, 2018.
- [200] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.
- [201] Qing Zhu, Qirong Mao, Hongjie Jia, Ocquaye Elias Nii Noi, and Juanjuan Tu. Convolutional relation network for facial expression recognition in the wild with few-shot learning. *Expert Systems with Applications*, 189:116046, 2022.
- [202] Wheidima Carneiro de Melo, Eric Granger, and Miguel Bordallo Lopez. Facial expression analysis using decomposed multiscale spatiotemporal networks. *Expert Systems with Applications*, 236:121276, 2024.

- [203] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2669–2676, 2020.
- [204] Yan Yan, Ying Huang, Si Chen, Chunhua Shen, and Hanzi Wang. Joint deep learning of facial expression synthesis and recognition. *IEEE Transactions on Multimedia*, 22(11):2792–2807, 2019.
- [205] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dcfac: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12715–12725, 2023.
- [206] Arbish Akram and Nazar Khan. Us-gan: on the importance of ultimate skip connection for facial expression synthesis. *Multimedia Tools and Applications*, 83(3):7231–7247, 2024.
- [207] Wei Wei, Chunna Tian, Stephen John Maybank, and Yanning Zhang. Facial expression transfer method based on frequency analysis. *Pattern Recognition*, 49: 115–128, 2016.
- [208] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [209] Hamza Bouzid and Lahoucine Ballihi. Facial expression video generation based-on spatio-temporal convolutional gan: Fev-gan. *Intelligent Systems with Applications*, 16:200139, 2022.
- [210] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18197–18207, 2022.
- [211] Chao Yang and Ser-Nam Lim. Unconstrained facial expression transfer using style-based generator. *arXiv preprint arXiv:1912.06253*, 2019.
- [212] Tiankai Hang, Huan Yang, Bei Liu, Jianlong Fu, Xin Geng, and Baining Guo. Language-guided face animation by recurrent stylegan-based generator. *IEEE Transactions on Multimedia*, 2023.
- [213] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.
- [214] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.
- [215] Tianyi Zhang, Zheng Wang, Jing Huang, Mohiuddin Muhammad Tasnim, and Wei Shi. A survey of diffusion based image generation models: Issues and their solutions. *arXiv preprint arXiv:2308.13142*, 2023.
- [216] Ximeng Sun, Huijuan Xu, and Kate Saenko. Twostreamvan: Improving motion modeling in video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2744–2753, 2020.
- [217] Kangfu Mei and Vishal Patel. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 9117–9125, 2023.

- [218] Yale Song, David Demirdjian, and Randall Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 500–506. IEEE, 2011.
- [219] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pages 363–370. Springer, 2003.
- [220] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.
- [221] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. <https://github.com/open-mmlab/mmpose>, 2020.
- [222] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014.
- [223] Tao Fang, Xinwei Zhao, Sameer K Shah, and Ioannis A Kakadiaris. 4d facial expression recognition. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1594–1601. IEEE, 2011.
- [224] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1290–1297. IEEE, 2012.
- [225] Hassen Drira, Boulbaba Ben Amor, Mohamed Daoudi, Anuj Srivastava, and Stefano Berretti. 3d dynamic expression recognition based on a novel deformation vector field and random forest. In *2012 21st International Conference on Pattern Recognition (ICPR)*, pages 1104–1107. IEEE, 2012.
- [226] Boulbaba Ben Amor, Hassen Drira, Stefano Berretti, Mohamed Daoudi, and Anuj Srivastava. 4-d facial expression recognition by learning geometric deformations. *IEEE Transactions on Cybernetics*, 44(12):2443–2457, 2014.
- [227] Mohamed E Hussein, Marwan Torki, Mohammad A Gawayyed, and Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Twenty-third international joint conference on artificial intelligence*, 2013.
- [228] Chen Chen, Kui Liu, and Nasser Kehtarnavaz. Real-time human action recognition based on depth motion maps. *Journal of real-time image processing*, 12:155–163, 2016.
- [229] Zhifei Li, Zhonglong Zheng, Feilong Lin, Howard Leung, and Qing Li. Action recognition from depth sequence using depth motion maps-based local ternary patterns and cnn. *Multimedia Tools and Applications*, 78:19587–19601, 2019.
- [230] Lichen Wang, Zhengming Ding, Zhiqiang Tao, Yunyu Liu, and Yun Fu. Generative multi-view human action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6212–6221, 2019.
- [231] Olalekan Agbolade, Azree Nazri, Razali Yaakob, Abdul Azim Ghani, and Yoke Kqueen Cheah. 3-dimensional facial expression recognition in human using multi-points warping. *BMC bioinformatics*, 20:1–15, 2019.

- [232] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.
- [233] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020.
- [234] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1112–1121, 2020.
- [235] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021.
- [236] Anis Kacem, Mohamed Daoudi, Boulbaba Ben Amor, Stefano Berretti, and Juan Carlos Alvarez-Paiva. A novel geometric framework on gram matrix trajectories for human behavior understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):1–14, 2018.
- [237] Duc-Phong Nguyen, Marie-Christine Ho Ba Tho, and Tien-Tuan Dao. Enhanced facial expression recognition using 3d point sets and geometric deep learning. *Medical & Biological Engineering & Computing*, 59(6):1235–1244, 2021.
- [238] Yi He, Keren Fu, Peng Cheng, and Jianwei Zhang. Facial expression recognition with geometric scattering on 3d point clouds. *Sensors*, 22(21):8293, 2022.
- [239] Imen Hamrouni Trimech, Ahmed Maalej, and Najoua Essoukri Ben Amara. Facial expression recognition using 3d points aware deep neural network. *Traitement du Signal*, 38(2), 2021.
- [240] Xing Li, Qian Huang, Zhijian Wang, Tianjin Yang, Zhenjie Hou, and Zhuang Miao. Real-time 3-d human action recognition based on hyperpoint sequence. *IEEE Transactions on Industrial Informatics*, 19(8):8933–8942, 2022.
- [241] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and Mohan Kankanhalli. Pstnet: Point spatio-temporal convolution on point cloud sequences. *arXiv preprint arXiv:2205.13713*, 2022.
- [242] Xiaoyu Zhu, Po-Yao Huang, Junwei Liang, Celso M de Melo, and Alexander G Hauptmann. Stmt: A spatial-temporal mesh transformer for mocap-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1526–1536, 2023.
- [243] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.
- [244] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005.
- [245] Abhinanda R Punnakal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021.

- [246] Saeed Ghorbani, Kimia Mahdavian, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F Troje. Movi: A large multi-purpose human motion and video dataset. *Plos one*, 16(6):e0253157, 2021.
- [247] Nikolaus F Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2(5):2–2, 2002.
- [248] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. URL <https://amass.is.tue.mpg.de>.
- [249] Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1057–1060, 2012.
- [250] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.
- [251] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2017.
- [252] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1012–1020, 2017.
- [253] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, pages 579–583. IEEE, 2015.
- [254] Bo Li, Yuchao Dai, Xuelian Cheng, Huahui Chen, Yi Lin, and Mingyi He. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 601–604. IEEE, 2017.
- [255] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297, 2017.
- [256] Van-Nam Hoang, Thi-Lan Le, Thanh-Hai Tran, Van-Toi Nguyen, et al. 3d skeleton-based action recognition with convolutional neural networks. In *2019 international conference on multimedia analysis and pattern recognition (MAPR)*, pages 1–6. IEEE, 2019.
- [257] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.
- [258] Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5323–5332, 2018.
- [259] Yancheng Wang, Yang Xiao, Fu Xiong, Wenxiang Jiang, Zhiguo Cao, Joey Tianyi Zhou, and Junsong Yuan. 3dv: 3d dynamic voxel for action recognition in depth video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 511–520, 2020.

- [260] Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. Rank pooling for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):773–787, 2016.
- [261] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.
- [262] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14204–14213, 2021.
- [263] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [264] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.
- [265] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [266] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. *Advances in Neural Information Processing Systems*, 34:11960–11973, 2021.
- [267] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.
- [268] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.
- [269] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.
- [270] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [271] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.
- [272] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.
- [273] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024.
- [274] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

- [275] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018.
- [276] Ron Banner, Itay Hubara, Elad Hoffer, and Daniel Soudry. Scalable methods for 8-bit training of neural networks. *Advances in neural information processing systems*, 31, 2018.
- [277] Brian Chmiel, Liad Ben-Uri, Moran Shkolnik, Elad Hoffer, Ron Banner, and Daniel Soudry. Neural gradients are near-lognormal: improved quantized and sparse training. *arXiv preprint arXiv:2006.08173*, 2020.
- [278] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [279] Jingtao Zhan, Qingyao Ai, Yiqun Liu, Jia Chen, and Shaoping Ma. Capability-aware prompt reformulation learning for text-to-image generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2145–2155, 2024.
- [280] Yiman Zhang, Hanting Chen, Xinghao Chen, Yiping Deng, Chunjing Xu, and Yunhe Wang. Data-free knowledge distillation for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7852–7861, 2021.
- [281] Kaiwen Cui, Yingchen Yu, Fangneng Zhan, Shengcai Liao, Shijian Lu, and Eric P Xing. Kd-dlgan: Data limited image generation via knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3872–3882, 2023.
- [282] Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. Spiralnet++: A fast and highly efficient mesh convolution operator. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.
- [283] Hao Zhu and Piotr Koniusz. Simple spectral graph convolution. In *International conference on learning representations*, 2021.
- [284] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [285] Federico Nocentini, Thomas Besnier, Claudio Ferrari, Sylvain Arguillere, Stefano Berretti, and Mohamed Daoudi. Beyond fixed topologies: Unregistered training and comprehensive evaluation metrics for 3d talking heads. *arXiv preprint arXiv:2410.11041*, 2024.
- [286] Federico Nocentini, Thomas Besnier, Claudio Ferrari, Sylvain Arguillere, Stefano Berretti, and Mohamed Daoudi. Scantalk: 3d talking heads from unregistered scans. *arXiv preprint arXiv:2403.10942*, 2024.

Résumé

Les expressions faciales jouent un rôle crucial dans la communication non verbale humaine, et leur intégration dans les systèmes de vision par ordinateur est essentielle pour améliorer les interactions homme-machine. L'analyse automatique des expressions faciales a depuis longtemps attiré l'attention de la communauté de la vision par ordinateur, ouvrant des perspectives dans des domaines tels que la sécurité, la robotique et l'AR/VR. Cette thèse explore deux applications aux défis interconnectés : la reconnaissance et la génération d'expressions faciales, avec des extensions vers les actions humaines. Notre objectif est de développer des modèles d'apprentissage profond qui capturent efficacement à la fois les informations spatiales et l'évolution temporelle, tout en optimisant l'utilisation des ressources. Nous présentons d'abord FEV-GAN, un modèle qui utilise des GAN convolutifs spatio-temporels pour générer des vidéos d'expressions faciales à partir d'images neutres tout en préservant l'identité. Bien que performant, ce modèle produit des vidéos de basse résolution. Pour pallier cette limitation, nous introduisons FacEnhance, qui exploite un modèle de diffusion pour améliorer la résolution et la qualité des vidéos générées. Nous proposons ensuite LMDM, un modèle de diffusion latente pour générer des expressions faciales et des actions humaines, en séparant les espaces de contenu et de mouvement afin de produire des mouvements variés. Enfin, nous explorons le domaine 3D avec SpATr, un modèle basé sur des convolutions spirales et des Transformers pour la classification des expressions faciales et des actions humaines dans des séquences de maillages 3D. Nous présentons également LM3DM, un modèle de diffusion latente pour la génération d'expressions faciales en 3D, qui applique directement la diffusion dans l'espace latent afin de capturer les dynamiques des expressions. Ce travail approfondit la compréhension, la synthèse et l'application des communications non verbales humaines en vision par ordinateur. Les travaux futurs viseront à améliorer la génération de vidéos haute résolution et de scans 3D.

Mots-clefs : Reconnaissance des expressions faciales ; Génération des expressions faciales ; Apprentissage profond ; Les réseaux antagonistes génératifs ; Modèles de diffusion.

Abstract

Facial expressions play a crucial role in non-verbal human communication, and their integration into computer vision systems is essential for improving human-computer interactions. The automatic analysis of facial expressions has long attracted the attention of the computer vision community, opening up potential in fields such as security, robotics, and AR/VR. This thesis explores two intersecting tasks: facial expression recognition and generation, with some solutions extended to human actions. Our goal is to develop deep learning models that effectively capture both spatial information and temporal evolution, while optimizing resource usage. First, we present FEV-GAN, a model that uses spatio-temporal convolutional GANs to generate facial expression videos from neutral images while preserving identity. Although effective, this model produces low-resolution videos. To address this limitation, we introduce FacEnhance, which leverages a diffusion model to enhance the resolution and quality of the generated videos. Next, we propose LMDM, a latent diffusion model for generating facial expressions and human actions, by separating content and motion spaces to produce varied movements. Finally, we explore the 3D domain with SpATr, a model based on spiral convolutions and Transformers for classifying facial expressions and human actions in 3D mesh sequences. We also present LM3DM, a latent diffusion model for generating 3D facial expressions, which directly applies diffusion within the latent space to capture the dynamics of expressions. This work deepens the understanding, synthesis, and application of human non-verbal communication in computer vision. Future efforts will focus on improving the generation of high-resolution videos and 3D scans.

Keywords: Facial Expression Recognition; Facial Expression Generation; Deep Learning; Generative Adversarial Networks; Diffusion Models.