Modèle unifié pour la reconnaissance et la génération d'expressions faciales 3D : manipulation de l'espace latent d'un Auto-Encodeur spiral

Hamza BOUZID[1] Lahoucine BALLIHI[2] Olivier LEZORAY[1]

[1]Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France

[2]LRIT-CNRST URAC 29, Mohammed V University in Rabat, Faculty of Sciences, Rabat, Morocco

Résumé – Les expressions faciales jouent un rôle central dans la communication non verbale, et leur intégration en vision par ordinateur est essentielle pour améliorer l'interaction homme-machine. Ce travail aborde deux tâches interconnectées, la reconnaissance et la génération, à travers un modèle unifié traitant les expressions faciales 3D dynamiques. En dissociant les composantes spatiales et temporelles des séquences de maillages, le modèle permet la reconnaissance par analyse temporelle et la génération par synthèse du mouvement. Plus précisément, un Auto-Encodeur léger basé sur des convolutions spirales extrait les caractéristiques géométriques spatiales de chaque maillage 3D. Ces convolutions, optimisées pour les maillages à topologie fixe, garantissent une efficacité computationnelle. Pour la reconnaissance, un Transformer basé sur l'auto-attention capture la dynamique temporelle en exploitant la séquence de caractéristiques. Pour la génération, un modèle de diffusion opère dans l'espace latent afin de synthétiser de nouvelles séquences correspondant aux expressions faciales cibles. Les représentations latentes générées sont ensuite décodées par l'Auto-Encodeur spiral pour reconstruire la sortie dans l'espace des maillages.

Abstract – Facial expressions are crucial for non-verbal communication, making their integration into computer vision essential for enhanced human-computer interaction. This work addresses two interconnected tasks, recognition, and generation, through a unified model that processes 3D dynamic facial expressions. By decoupling spatial and temporal components in mesh sequences, the model enables recognition via temporal analysis and generation through motion synthesis. Specifically, A lightweight Auto-Encoder, built on spiral convolutions, extracts spatial geometric features from each 3D mesh. These convolutions, optimized for fixed-topology mesh data, ensure computational efficiency. For recognition, a self-attention Transformer captures the temporal dynamics of feature sequences. For generation, a diffusion model synthesizes new facial expression sequences in the latent space, which are then decoded by the spiral Auto-Encoder into mesh space.

1 Introduction

Les expressions faciales jouent un rôle clé dans la communication non verbale. Reconnaître et générer des expressions faciales en 3D est crucial pour des domaines comme l'informatique affective et l'interaction humain-machine. La reconnaissance des expressions faciales 3D (3D FER) vise à identifier les émotions à partir de maillages 3D ou de nuages de points, tandis que la génération des expressions faciales 3D (3D FEG) vise à synthétiser des expressions réalistes à partir de conditions données.

3D FER et 3D FEG sont souvent traitées de manière indépendante. Pour la 3D FER, il y a des méthodes classiques qui utilisent des caractéristiques définies manuellement [6], tandis que d'autres projettent les données 3D en 2D pour appliquer des CNN [2]. Récemment, des approches basées sur les nuages de points ont été introduites pour mieux capturer les informations spatiales et temporelles [9]. En 3D FEG, les modèles paramétriques comme les 3D Morphable Models (3DMM) [3] sont utilisés, ainsi que des approches d'apprentissage profond, telles que les Auto-Encodeurs spectraux, les GANs, les Transformers et les modèles de diffusion, pour synthétiser des séquences dynamiques [4, 15, 11, 16, 10].

Cependant, ces méthodes ont des limitations. Elles n'exploitent pas pleinement les avantages des maillages 3D à topologie fixe, qui offrent une structure géométrique cohérente et évitent les défis posés par les nuages de points ou les données de profondeur. Ces maillages nécessitent des architectures adaptées, moins gourmandes en ressources. De plus, les approches existantes sont souvent inefficaces, ce qui limite leur capacité à traiter de longues séquences. Enfin, elles reposent sur des architectures distinctes et des prétraitements spécifiques pour chaque tâche, compliquant leur mise en œuvre et limitant leur évolutivité et leur intégration.

Pour répondre à ces défis, cet article propose un modèle unifié pour reconnaître et générer simultanément les expressions faciales. Il repose sur l'utilisation de séquences de maillages pour modéliser les dynamiques faciales 3D. Notre approche découple les dimensions spatiales et temporelles afin d'améliorer l'efficacité et la généralisation. Dans un premier temps, nous entraînons un Auto-Encodeur basé sur des convolutions spirales [4] afin d'extraire des représentations latentes compactes et informatives des maillages 3D et permettent leur reconstruction. Cet Auto-Encodeur génère ainsi des représentations indépendantes des tâches, contenant des informations géométriques générales exploitables pour différentes applications. Pour la 3D FER, un réseau Transformer [14] traite les séquences latentes pour capturer les dépendances temporelles via un mécanisme d'attention et classifier le mouvement 4D. Le choix de Transformer repose sur ses capacités de parallélisation et de scalabilité, permettant le traitement efficace de longues séquences. Pour la 3D FEG, nous utilisons un modèle de diffusion latent [8] afin d'apprendre l'espace latent de l'Auto-Encodeur spiral. Cette approche permet de générer de nouvelles séquences latentes correspondant à des expressions faciales, qui sont ensuite décodées en maillages 3D. L'utilisa-

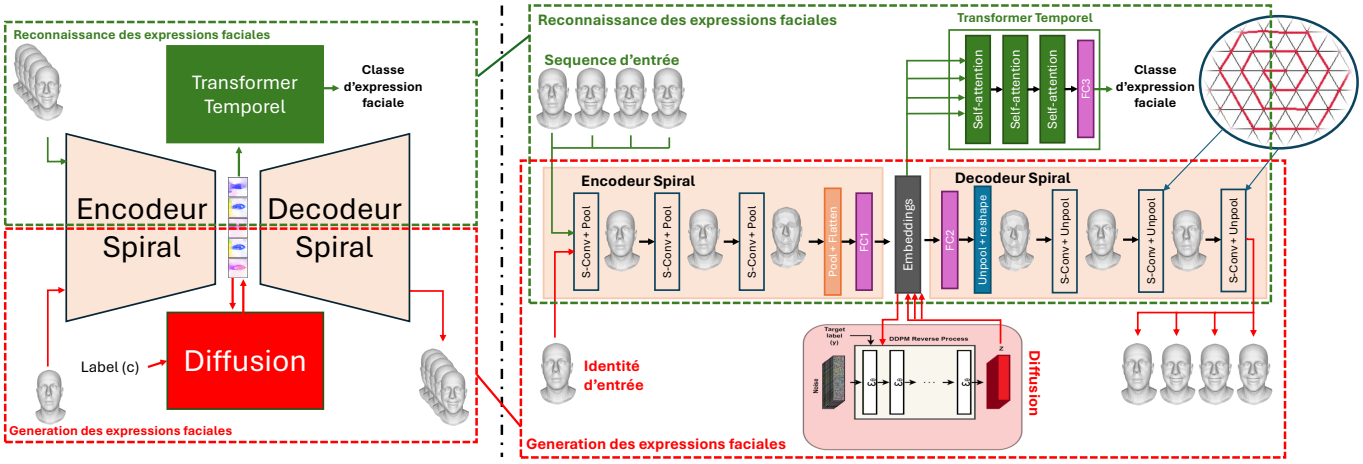

FIGURE 1 : Présentation générale du modèle proposé. À gauche, nous donnons un aperçu global, tandis qu'à droite, nous présentons des détails plus précis sur chaque composant.

tion de modèles de diffusion est inspirée par leur capacité à modéliser des distributions complexes à grande variance.

2 Méthode proposée

La figure 1 présente le schéma de la méthode proposée, qui repose sur deux étapes : l'apprentissage des représentations latentes et la modélisation des dynamiques spatiales et temporelles. Un Auto-Encodeur à convolutions spirales extrait tout d'abord les représentations latentes des maillages 3D. Pour la reconnaissance, un Transformer les analyse pour capturer les dépendances temporelles et classifier le mouvement 4D, tandis qu'un modèle de diffusion génère de nouvelles séquences latentes, décodées ensuite en maillages 3D pour la génération.

2.1 Auto-Encodeur à convolutions spirales

L'Auto-encodeur spiral (S-AE) transforme un maillage 3D M_i en une représentation latente compacte z_i tout en permettant une reconstruction $\bar{M}_i$ proche de l'original. Contrairement aux méthodes classiques comme PCA et LDA, qui peinent à capturer les déformations complexes des données 3D, l'apprentissage profond est plus adapté. Cependant, appliquer des convolutions sur des maillages 3D irréguliers est difficile en raison de l'absence de coordonnées globales. Pour y remédier, nous utilisons les convolutions spirales [4], qui exploitent efficacement la connectivité des maillages à topologie fixe. Ces convolutions favorisent le partage de poids, réduisent le nombre de paramètres et rendent le réseau plus léger et plus facile à optimiser.

L'encodeur prend un maillage M_i en entrée à travers trois couches de convolutions spirales avec sous-échantillonnage, suivies d'une couche entièrement connectée (FC1) pour obtenir la représentation latente z_i de taille 512. Le décodeur inverse ce processus en utilisant FC2 et trois couches de convolution spirale avec sur-échantillonnage pour reconstruire le maillage $\bar{M}_i$.

Le réseau est entraîné en minimisant la perte de reconstruction (MAE) :

$$\text{S-AE}^* = \arg\min_{\text{S-AE}} \mathcal{L}_{\text{S-AE}}(M_i, \bar{M}_i) = \arg\min_{\text{S-AE}} \|M_i - \bar{M}_i\|_1.$$

En assurant une reconstruction précise, le S-AE capture les caractéristiques essentielles de la forme 3D, notamment la structure, l'expression, la pose et l'apparence.

2.2 Transformer temporel

Inspirés par le succès des Transformers [14] en NLP et en vision par ordinateur, nous utilisons l'auto-attention pour capturer le contexte temporel dans des séquences de maillages. Ce mécanisme permet un traitement parallèle des séquences et modélise efficacement les dépendances à long terme.

Étant donné une séquence de l représentations latentes $Z = [z_1, z_2, ..., z_l]$, encodant les informations spatiales, chaque z_i est utilisé comme un token pour le Transformer. L'auto-attention (SA) extrait les requêtes (Q), clés (K) et valeurs (V) via des transformations linéaires :

$$Q = W_q \cdot Z, \quad K = W_k \cdot Z, \quad V = W_v \cdot Z.$$

Les poids d'attention, qui représentent les relations temporelles, sont obtenus comme suit :

$$attention(Q, K) = softmax\left(\frac{Q^T \cdot K}{\sqrt{C_K}}\right),$$

et la sortie de l'auto-attention est calculée par :

$$F_{out} = V \cdot attention(Q, K).$$

Pour améliorer l'apprentissage, nous employons une auto-attention multi-tête, où plusieurs mécanismes d'auto-attention, chacun avec des poids indépendants, sont appliqués puis concaténés. Ce mécanisme multi-tête est appliqué trois fois avant qu'une couche entièrement connectée (FC3) ne finalise la classification de l'action.

Le modèle est entraîné en minimisant la fonction de perte d'entropie croisée multi-classes :

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

où $\mathbf{y}$ représente les étiquettes des classes réelles, et $\bar{\mathbf{y}}$ est la distribution de probabilité prédite.

2.3 Modèle de diffusion : génération

Nous utilisons un modèle de diffusion [8] basé sur un UNet 2D, où l'une des dimensions représente la structure spatiale du visage et l'autre son évolution temporelle. Ce modèle apprend la génération de mouvements dans l'espace latent appris par le S-AE. L'entraînement suit un cadre conditionnel, où le modèle

est guidé par la représentation latente du maillage d'entrée z_{in} et l'étiquette de mouvement cible c. Ainsi, le mouvement généré reste cohérent avec l'identité du sujet et la catégorie d'action visée. L'objectif est de modéliser la distribution conditionnelle p_θ, permettant la synthèse du mouvement sous des contraintes spécifiques.

Phase d'entraînement : Le modèle de diffusion fonctionne en ajoutant progressivement du bruit à la séquence d'encodages de mouvement $q(x_t|x_{t-1})$, en commençant par :

$$X_0 = Z = \text{S-AE}([M_1, M_2, ..., M_l]); \quad X_0 \in \mathbb{R}^{512\times l}$$

où Z désigne les encodages latents associés à la séquence de mouvements d'entrée. Le processus de débruitage qui suit est alors :

$$p_\theta(X_{t-1}|X_t, z_{in}, c) = \mathcal{N}\left(X_{t-1}; \mu_\theta(X_t, z_{in}, c, t), \sigma_\theta(X_t, z_{in}, c, t)^2\mathbf{I}\right).$$

Ici, μ_θ et σ_θ représentent respectivement la moyenne et la variance prédites, conditionnées par l'encodage bruité actuel x_t, l'encodage de référence z_{in}, l'étiquette cible c, et le pas de temps t. Le processus de diffusion est défini avec $T = 1000$ étapes, assurant une évolution progressive du bruit vers une représentation affinée. La fonction de perte optimise le modèle en minimisant l'erreur entre le bruit réel ε et le bruit prédit ε_θ, garantissant ainsi une génération précise du mouvement :

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

Phase d'inférence : Le modèle de diffusion génère des encodages de mouvement directement à partir du bruit en le débruitant progressivement, conditionné par z_{in} et c, et en l'affinant itérativement de $t = T = 1000$ à $t = 1$. Les encodages générés $\bar{Z} = \{\bar{z}_{1\ldots l}\}$ sont ensuite utilisés par le décodeur du S-AE pour reconstruire la séquence finale de maillages.

3 Résultats expérimentaux

3.1 Jeux de données

Nous évaluons notre méthode sur plusieurs jeux de données : pour les expressions faciales, nous utilisons **MUG Facial Expression** [1], qui contient des vidéos de 86 personnes et sept expressions de base, que nous enregistrons sur un template 3D à l'aide d'EMOCA [5], ainsi que **COMA** [13], qui propose des séquences 4D de déformations faciales. Nous effectuons également des tests sur la reconnaissance des actions humaines avec **MoVi** [7], qui contient 1864 séquences de 86 individus exécutant 20 actions, et **Babel** [12], une combinaison de 15 jeux de données AMASS, avec 43.5 heures d'enregistrements couvrant 250 actions, en utilisant les 60 classes les plus fréquentes.

3.2 Reconnaissance des expressions faciales

Pour le jeu de données MUG, l'enregistrement des vidéos en maillages 3D génère des visages avec des transitions fluides portant les performances de classification à **91.07%**. De plus, nous mettons en évidence la flexibilité des caractéristiques extraites par l'Auto-Encodeur spiral, soulignant leur potentiel d'utilisation pour diverses tâches sans nécessiter de réentraînement. En exploitant ces représentations indépendantes des tâches, plusieurs classifieurs sont utilisés pour prédire les expressions faciales, l'identité des individus et le genre. Le tableau 1 indique des précisions de 91.07% pour la reconnaissance des expressions faciales, 98.21% pour l'identification des individus et 88.33% pour la prédiction du genre. Ces résultats démontrent la capacité de l'Auto-Encodeur à générer des descripteurs robustes et polyvalents, facilitant l'optimisation des ressources et du temps de calcul.

TABLE 1 : Précision de reconnaissance des expressions faciales, de l'identité et du genre (%) sur MUG.

Reconnaissance		
expressions	**identité**	**genre**
91.07%	98.21%	88.33%

Nous avons également évalué le modèle proposé pour la reconnaissance des actions humaines sur les jeux de données MoVi et Babel, en le comparant aux méthodes de l'état de l'art. Le tableau 2 présente les résultats de cette évaluation. Pour le jeu de données MoVi, notre méthode atteint une précision de 95.42%, tandis que pour Babel, elle obtient une précision de 70.36% en Top-1 et 89.12% en Top-5, surpassant ainsi les autres modèles. En comparaison avec les approches existantes, notre méthode démontre une performance compétitive, illustrant sa robustesse et son efficacité pour la reconnaissance d'actions humaines, même sur des jeux de données complexes et variés.

TABLE 2 : Précision de reconnaissance des actions (%) sur les jeux de données MoVi et Babel.

		MoVi	**Babel**	
Méthode	**Entrée**	Top-1(%)	Top-1(%)	Top-5(%)
2s-AGCN-FL (CVPR'19)	Squelette 3D	-	49.62	79.12
2s-AGCN-CE (CVPR'19)	Squelette 3D	-	63.57	86.77
CTR-GCN (ICCV'21)	Squelette 3D	-	67.30	88.50
MS-G3D (CVPR'20)	Squelette 3D	-	67.43	87.99
P4Transformer (CVPR'21)	Nuage de points	91.25	63.54	86.55
PSTNet (ICLR'21)	Nuage de points	88.75	61.94	84.11
SequentialPointNet (TII'22)	Nuage de points	**98.84**	62.92	84.58
STMT (CVPR'23)	Maillage	-	67.65	88.68
Notre modèle	Maillage	95.42	**70.36**	**89.12**

Ensuite, nous démontrons l'efficacité et la scalabilité de notre modèle en l'entraînant sur des séquences de différentes longueurs issues du jeu de données MoVi. Les résultats, présentés dans le tableau 3, montrent que notre modèle optimise l'utilisation de la mémoire, avec une allocation initiale modeste de $480 + 157$MB (S-AE+T) pour 12 maillages, qui s'étend progressivement à $480 + 1480$MB pour 960 maillages, tout en maintenant un taux de reconnaissance à un niveau constant. Cela démontre l'efficacité mémoire de notre modèle, confirmant son adaptabilité fluide à des séquences plus longues avec des augmentations marginales de mémoire GPU.

TABLE 3 : Précision de la reconnaissance des actions (%) et utilisation mémoire/par lot sur le jeu de données MoVi.

Nombre de frames	12	48	96	960
Mem S-AE+T (MB)	480+157	480+187	480+230	480+1480
Précision (%)	93.75	95.42	95.00	94.58

3.3 Génération des expressions faciales

La figure 2 présente des exemples de séquences générées par notre modèle, entraîné sur le jeu de données MUG, chacune

représentant une expression faciale de base. Les séquences générées illustrent des transitions continues, fluides et naturelles entre les expressions faciales, tout en atteignant l'expression cible.

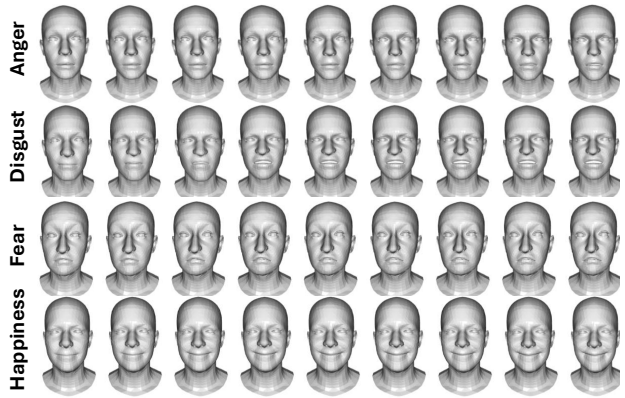

FIGURE 2 : Exemples générés lors de l'évaluation sur MUG.

La figure 3 montre des exemples de séquences générées par notre modèle, entraîné sur le jeu de données COMA, chacune représentant une expression faciale extrême et asymétrique. Ces résultats démontrent la capacité de notre modèle à traiter des tâches plus complexes que la génération des six expressions faciales de base, en générant douze classes (classes de COMA) d'expressions plus expressives et extrêmes.

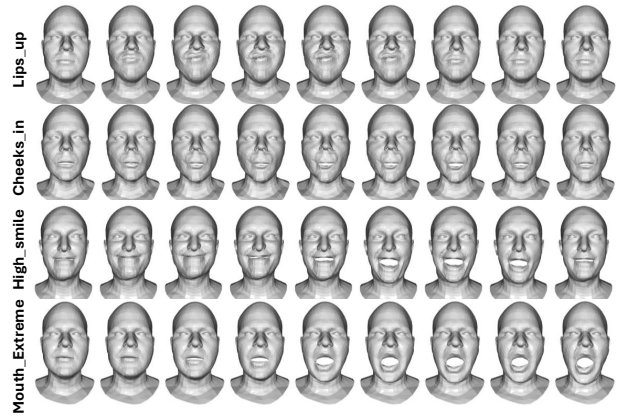

FIGURE 3 : Exemples générés lors de l'évaluation sur COMA.

NB : La tâche de génération est en cours d'évaluation, et des résultats quantitatifs détaillés seront fournis ultérieurement.

4 Discussion & conclusions

Nos expériences démontrent que le modèle proposé unifie efficacement la reconnaissance et la génération au sein d'un même cadre, en exploitant le même extracteur de caractéristiques et le même reconstructeur de maillage (S-AE). En reconnaissance, nos résultats mettent en évidence l'efficacité mémoire du modèle sur des séquences de longueurs variées, le distinguant de nombreuses approches de pointe qui se limitent à des séquences plus courtes. De plus, l'extracteur de caractéristiques génère des représentations flexibles pouvant être utilisées pour plusieurs tâches, y compris la reconnaissance des expressions faciales, de l'identité et du genre, sans nécessiter de réentraînement.

Concernant la génération, notre modèle est efficace pour synthétiser des expressions faciales avec des mouvements naturels, fluides et variés. Cependant, malgré ces résultats prometteurs, le processus de génération est encore en phase de développement et présente certaines limites, telles que la présence occasionnelle d'artefacts, d'erreurs dans les séquences générées, ainsi que des expressions de très faible intensité.

Bien que notre modèle montre un fort potentiel, il présente certaines contraintes. Notamment, il est conçu pour des maillages à topologie fixe, ce qui limite son adaptabilité à des structures de maillage variables. De plus, ses performances en reconnaissance dépendent de la qualité de l'enregistrement des maillages, ce qui peut affecter sa robustesse dans des applications en conditions réelles. Surmonter ces défis constitue une perspective importante pour les recherches futures.

Références

[1] N. AIFANTI, C. PAPACHRISTOU et A. DELOPOULOS : The mug facial expression database. *In WIAMIS*, pages 1–4. IEEE, 2010.

[2] M. BEHZAD, N. VO, X. LI et G. ZHAO : Landmarks-assisted collaborative deep framework for automatic 4d facial expression recognition. *In FG*, pages 1–5. IEEE, 2020.

[3] V. BLANZ et T. VETTER : A morphable model for the synthesis of 3d faces. *In Seminal Graphics Papers*, pages 157–164. 2023.

[4] G. BOURITSAS, S. BOKHNYAK, S. PLOUMPIS, M. BRONSTEIN et S. ZAFEIRIOU : Neural 3d morphable models : Spiral convolutional networks for 3d shape representation learning and generation. *In ICCV*, pages 7213–7222, 2019.

[5] R. DANĚČEK, M. J. BLACK et T. BOLKART : Emoca : Emotion driven monocular face capture and animation. *In CVPR*, pages 20311–20322, 2022.

[6] H. DRIRA, B. B. AMOR, M. DAOUDI, A. SRIVASTAVA et S. BERRETTI : 3d dynamic expression recognition based on a novel deformation vector field and random forest. *In ICPR*, pages 1104–1107. IEEE, 2012.

[7] S. GHORBANI, K. MAHDAVIANI, A. THALER, K. KORDING, D. J. COOK, G. BLOHM et N. F. TROJE : Movi : A large multi-purpose human motion and video dataset. *PLOS ONE*, 16(6):e0253157, 2021.

[8] J. HO, A. JAIN et P. ABBEEL : Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[9] X. LI, Q. HUANG, Z. WANG, T. YANG, Z. HOU et Z. MIAO : Real-time 3-d human action recognition based on hyperpoint sequence. *TII*, 19(8):8933–8942, 2022.

[10] N. OTBERDOUT, C. F., M. DAOUDI, S. BERRETTI et A. DEL BIMBO : Sparse to dense dynamic 3d facial expression generation. *In CVPR*, pages 20385–20394, 2022.

[11] R. A. POTAMIAS, J. ZHENG, S. PLOUMPIS, G. BOURITSAS, E. VERVERAS et S. ZAFEIRIOU : Learning to generate customized dynamic 3d facial expressions. *In ECCV*, pages 278–294. Springer, 2020.

[12] A. R PUNNAKKAL, A. CHANDRASEKARAN, N. ATHANASIOU, A. QUIROS-RAMIREZ et M. J. BLACK : Babel : bodies, action and behavior with english labels. *In CVPR*, pages 722–731, 2021.

[13] A. RANJAN, T. BOLKART, S. SANYAL et M. J. BLACK : Generating 3d faces using convolutional mesh autoencoders. *In ECCV*, pages 704–720, 2018.

[14] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N GOMEZ, Ł. KAISER et I. POLOSUKHIN : Attention is all you need. *NeurIPS*, 30, 2017.

[15] K. ZOU, S. FAISAN, B. YU, S. VALETTE et H. SEO : 4d facial expression diffusion model. *TOMM*, 2023.

[16] K. ZOU, B. YU et H. SEO : 3d facial expression generator based on transformer vae. *In ICIP*, pages 2550–2554. IEEE, 2023.