

# AffectVLM: Contrastive Language-Image Learning with Augmented Textual Prompts for 3D/4D Facial Expression Recognition Using Vision-Language Model

Muzammil Behzad<sup>1,\*</sup>, Guoying Zhao<sup>2</sup>

<sup>1</sup> Information & Computer Science Department, King Fahd University of Petroleum & Minerals, Saudi Arabia

<sup>2</sup> Center for Machine Vision and Signal Analysis, University of Oulu, Finland

Email: muzammil.behzad@kfupm.edu.sa, guoying.zhao@oulu.fi

**Abstract**—In this paper, we introduce AffectVLM, a vision-language model designed to integrate multiviews for a semantically rich and visually comprehensive understanding of facial emotions from 3D/4D data. To effectively capture visual features, we propose a joint representation learning framework paired with a novel gradient-friendly loss function that accelerates model convergence towards optimal feature representation. Additionally, we introduce augmented textual prompts to enhance the model’s linguistic capabilities and employ mixed view augmentation to expand the visual dataset. We also develop a Streamlit app for a real-time interactive inference and enable the model for distributed learning. Extensive experiments validate the superior performance of AffectVLM across multiple benchmarks.

## I. INTRODUCTION

Facial Expression Recognition (FER) is a key research area within affective computing, focusing on analyzing and interpreting human emotions through facial analysis, with several applications in human-computer interaction [1], mental health [2], education [3], and more [4]. Given the complexity of the human emotions and the pioneering emotion theory proposed by Ekman & Friesen [5], researchers have developed various models to identify and resolve potential research gaps [6].

The literature presents various methods to learn from the underlying 3D facial geometry. In this regard, the most widely used approaches include local feature-based methods [7], [8], [9], template-based methods [10], [11], curve-based methods [12], [13], and 2D projection-based methods [14], [15]. In recent years, 3D/4D FER has gained a significant attention as it enables deep learning models to extract additional discriminative facial features given the facial depth axis. For example, Yin *et al.* [16] and Sun *et al.* [17] employed Hidden Markov Models (HMM) to capture temporal facial features from 4D facial scans. Likewise, Ben Amor *et al.* [18] demonstrated the effectiveness of a deformation vector field based on Riemannian analysis using a random forest classifier. Similarly, Sandbach *et al.* [19]

utilized Hidden Markov Models (HMM) and GentleBoost to learn free-form representations of 3D frames. Additionally, the authors in [20] represented geometric coordinates and their normals as feature vectors, while another study [21] utilised dynamic local binary patterns (LBP) to recognize facial expressions using a support vector machine (SVM). In a related approach, the authors in [22] extracted features from polar angles and curvatures, proposing a spatio-temporal LBP-based feature extractor for recognition.

Conversely, Li *et al.* [23] introduced an intriguing framework for automatic 4D Facial Emotion Recognition (FER) using a dynamic geometrical image network. They generated geometrical images by calculating differential quantities from the given 3D facial point clouds. The emotion prediction is achieved through score-level fusion of the probability scores derived from various geometrical images.

### A. Motivations

While effective, these traditional methods often rely on manually extracted features and localized cues, limiting performance and adaptability in the real-world scenarios. Inspired by the scaling victory of large language models (LLMs) [24] and large vision language models (LVLMs) [25], we work towards developing a VLM with joint representation learning capabilities to benefit from the underlying facial patterns stored as muscle movements in 3D/4D faces, for a significantly better facial expression recognition (FER). Specifically, instead of using conventional 2D faces only (e.g., [26], [27], [28]), this emotion recognition involves classifying facial expressions from 3D/4D faces with added spatio-temporal features, and the substantial results [29], [30], [31], [9] have validated its advantages. VLMs offer a promising solution by learning joint representations, improving generalization across diverse datasets without manual feature engineering. This highlights the need for advanced approaches that leverage VLMs, which can learn from both visual and textual data [32]. By aligning visual cues with semantic labels, VLMs enhance the robustness of FER systems in varying conditions. However, existing 3D/4D datasets have limited capacity, which hinders the performance of deep learning models. This underscores the importance of improved augmentation strategies, such as spatial transformations, temporal variations, and multiview combinations, to address data scarcity and overfitting.

This work was supported by KFUPM (grant EC241013). This work was also supported by the Research Council of Finland (former Academy of Finland) Academy Professor project EmotionAI (grants 336116, 345122, 359854), the University of Oulu Research Council of Finland Profi 7 (grant 352788), and EU HORIZON-MSCA-SE-2022 project ACMod (grant 101130271). The authors would also like to acknowledge the SDAIA-KFUPM Joint Research Center for Artificial Intelligence for computational resources.  
\* indicates corresponding author.

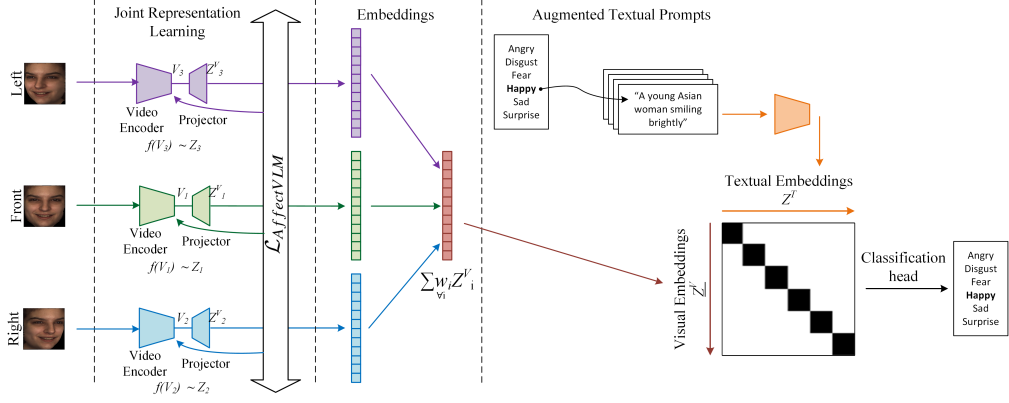


Fig. 1. Overview of our proposed AffectVLM: Affective Vision-Language Model.

## B. Contributions

To the best of our knowledge, existing literature lacks research on 3D/4D FER models utilizing VLMs. This gap arises primarily because implementing such models is not straightforward, accompanied by the complexity and variability in the 3D data structure. In this context, we introduce AffectVLM, a VLM based architecture [33]. The key features of our model are as follows:

- 1) We propose a joint embedding space for coherent feature representation to integrate multiview embeddings.
- 2) We also propose augmented textual prompts to extend the model’s understanding of emotional semantics by augmenting the training labels with extended textual prompts.
- 3) We introduce mixed view augmentation to generate diverse training samples by combining augmentations from multiple views, thereby enriching the dataset.
- 4) We employ a novel gradient-friendly loss function for model’s smoother convergence during training.
- 5) We design AffectVLM with the capability to leverage distributed training for improved scalability.
- 6) We develop a Streamlit application for user inference, enabling interactive access to the model for real-time emotion classification.

Furthermore, it must be noted that AffectVLM is equally applicable to other downstream tasks like, face recognition, anti-spoofing and identity recognition.

## II. AFFECTVLM: AFFECTIVE VISION-LANGUAGE MODEL

Our AffectVLM model’s ability to incorporate multiviews offers a significant advantage with its robust and scalable implementation, making it highly effective and ready for deployment with minimal adjustments.

### A. Learning Joint Embedding Space Representations

As shown in Fig. 1, we tailor our model to learn joint embedding space to facilitate coherent feature representation across both visual and textual modalities. We use contrastive language-image pre-training (CLIP) [33] as a backbone VLM

model and extend it for multiviews in our problem statement. Specifically, we aim to achieve joint embedding of visual features from multiple facial views and corresponding textual prompts. The model processes multiview data, thereby capturing frontal, left, and right perspectives, allowing it to learn a rich, holistic representation of emotions from different facial angles by optimizing the loss function jointly with the features from all views in the embedding space. Simultaneously, we integrate augmented textual prompts into the learning process as discussed in the next sub-section. This joint learning framework ensures that different views of the same emotion are aligned in a shared embedding space with their textual counterparts, promoting a deeper understanding of the emotion. This approach not only improves the model’s ability to generalize across unseen data but also enhances its robustness in recognizing emotions from varying viewpoints and contextual diversity. AffectVLM supports multiple model engines such as ViT(16/32) [34], and ResNet variants [35] to learn the underlying embeddings.

### B. Augmented Textual Prompts

To enrich the semantic representation in our model’s training process, we propose Augmented Textual Prompts. Considering the variability in expressions and subject-specific attributes like gender, age, ethnicity, and other contextual details, a single label such as “Happy” may not capture the full spectrum of emotional states. By systematically generating multiple textual variations using predefined templates and adaptive modifications based on emotion and

TABLE I  
DIFFERENT TEXTUAL PROMPTS FOR “HAPPY” EXPRESSION

Prompt ID	Textual Prompt
1	“A young woman with a joyful expression”
2	“An older man looking very happy”
3	“A smiling male full of joy”
4	“A young adult showing happiness”
5	“A middle-aged black woman looking very happy”
6	“Face of an older Asian woman showing a smile”
7	“A young Asian woman smiling brightly”
8	“An older Black female smiling”

metadata, such as “a joyful expression of a young female” or “an older male smiling”, we provide our VLM model with a broader linguistic context. This approach enhances the dataset with richer semantics and improves the model’s ability to generalize across different data samples. The use of diverse prompts facilitates a deeper understanding of facial features. In Table I, we present a list of textual prompts for the “Happy” expression, generated using ChatGPT-3.5. Note that the range of prompts grows proportionally with more available annotations within the dataset. In this method, multiviews use different textual prompts during training to gain more semantic understanding.

### C. Mixed View Augmentation

To address the challenge of limited data in 3D/4D FER [36], [37], we propose Mixed View Augmentation, an innovative method designed to enhance training data diversity by combining multiviews with existing common augmentation techniques such as flipping, rotation, cropping, and scaling. This approach allows us to apply different transformations across various views of the same emotion, creating a more extensive dataset. For instance, we can flip the frontal view of a face while cropping the left view, rather than augmenting solely within a single view. This cross-placing of augmentations significantly increases variability in the training set, facilitating more comprehensive learning of facial features. By employing this strategy, our model is better positioned to generalize across varying viewing conditions and emotional expressions, which ultimately enhances performance. This straightforward yet effective strategy not only reduces the risk of overfitting but also fosters robust learning, enabling the model to adapt to a wide array of visual contexts.

### D. Gradient-friendly Loss Function with Learnable Margin

Using a loss function with different views can enhance the training process by enabling the model to learn more robust features across multiple perspectives. In this context, our model stands out due to its unique loss function that helps the network learn and converge more quickly. Unlike typical approaches, we work to optimize the loss together across all views. We define a shared embedding space  $Z$  where features from different views are projected. Given  $V_1, V_2, V_3$  as features from three different views, the shared embedding can be represented as:

$$Z_i = f_{AffectVLM}(V_i), i \in \{1, 2, 3\}, \quad (1)$$

where  $f_{AffectVLM}$  is our model that projects features from each view into the shared space  $Z$ . We define our loss function as follows:

$$\mathcal{L}_{AffectVLM} \triangleq \underbrace{\sum_{i,j} 1 - \Theta(Z_i, Z_j)}_{\substack{\mathcal{L}_{mc}: \text{ multiview contrastive loss} \\ \text{for same emotion}}} + \underbrace{\sum_{k,l} \max(0, \Theta(Z_k, Z_l) - \alpha)}_{\substack{\text{for different emotion}}} + \underbrace{\max(0, \|Z_{anchor} - Z_{positive}\|_2^2 - \|Z_{anchor} - Z_{negative}\|_2^2 + \alpha)}_{\mathcal{L}_{mt}: \text{ multiview triplet loss}}, \quad (2)$$

where  $\Theta(\cdot)$  represents a similarity function such as Cosine similarity or Euclidean distance. Here, the multiview contrastive loss function  $\mathcal{L}_{mc}$  is used to enforce similarity for the same emotion across views and dissimilarity for different emotions. Similarly, the multiview triplet loss  $\mathcal{L}_{mt}$  is used to further enhance the learning process by explicitly constructing anchor, positive, and negative samples, where  $Z_{anchor}$  is the shared representation from one view,  $Z_{positive}$  is the representation of the same emotion from another view, and  $Z_{negative}$  is from a different emotion. The margin  $\alpha$  is a hyperparameter that helps ensure a gap between positive and negative pairs. In our setup, we initialize  $\alpha$  as a learnable parameter, which allows the model to adapt the margin dynamically during training, ensuring model generalization while catering to the needs of various diverse datasets.

### E. Performance Scalability with Distributed Learning

Our model’s ability to train in a distributed environment is a significant feature that enhances its performance. We utilize PyTorch’s distributed communication package (`torch.distributed`) alongside NVIDIA’s Collective Communications Library (NCCL) to efficiently distribute workloads across available resources. Our implementation dynamically determines the optimal training setup such as, multi-GPU, single-GPU, or CPU, based on resource availability, employing the `ddp` strategy in `pytorch-lightning` when multiple GPUs are present. To mitigate synchronization issues across GPUs or nodes, we dynamically allocate network `<port>` and `<IP>` addresses to establish flexible TCP communication sockets. This approach facilitates efficient multi-process parallelism, ensuring resource-efficient training of our AffectVLM model.

### F. Interactive Inference with Streamlit

To make our model more user friendly, we have developed a `streamlit` application that serves as a service interface for real-time inference using our AffectVLM model. Users can upload multiple images for immediate processing, specifically three multiview images, enabling on-the-fly emotion classification based on the learned features from the joint embedding space. This architecture not only enhances user experience but also demonstrates the model’s scalability and adaptability to diverse input modalities in 3D/4D expression recognition tasks. Importantly, this user application opens directions for accelerated research development especially within real-time affective computing. We aim to publish this interactive application soon.

## III. RESULTS AND DISCUSSIONS

We validate our model using the Bosphorus [38], BU-3DFE [39], BU-4DFE [16] and BP4D-Spontaneous [40] datasets. Consistent with prior works [23], [41], [42], [43], we generate multiview 2D images from 3D/4D point clouds and apply rank pooling [44] for video data to create dynamic images. A 10-fold subject-independent cross-validation is employed for all experiments.

TABLE II

ACCURACY (%) COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE BU-3DFE SUBSET I AND SUBSET II, AND BOSPHORUS DATASETS.

Method	Subset I ( $\uparrow$ )	Method	Subset II ( $\uparrow$ )	Bosphorus ( $\uparrow$ )
Zhen <i>et al.</i> [30]	84.50 (8.13 $\uparrow$ )	Li <i>et al.</i> [9]	80.42 (6.14 $\uparrow$ )	79.72 (10.02 $\uparrow$ )
Yang <i>et al.</i> [31]	84.80 (7.83 $\uparrow$ )	Yang <i>et al.</i> [31]	80.46 (6.10 $\uparrow$ )	77.50 (12.24 $\uparrow$ )
Li <i>et al.</i> [9]	86.32 (6.31 $\uparrow$ )	Li <i>et al.</i> [14]	81.33 (5.23 $\uparrow$ )	80.00 (9.74 $\uparrow$ )
Li <i>et al.</i> [14]	86.86 (5.77 $\uparrow$ )	Sui <i>et al.</i> [45]	-	82.06 (7.68 $\uparrow$ )
Oyedotun <i>et al.</i> [15]	89.31 (3.32 $\uparrow$ )	Li <i>et al.</i> [46]	-	86.77 (2.97 $\uparrow$ )
<b>AffectVLM (Ours)</b>	<b>92.63</b>	<b>AffectVLM (Ours)</b>	<b>86.56</b>	<b>89.74</b>

### A. Performance on 3D FER

Following existing protocols [14], [15], the BU-3DFE dataset with 101 subjects is divided into Subset I, which includes two higher intensity levels, and Subset II, which contains all four intensity levels. For the Bosphorus dataset, 65 subjects performed six expressions. Table II shows our model’s performance, achieving 92.63% on Subset I, surpassing state-of-the-art results [15], and 89.74% on the Bosphorus dataset. Notably, for Subset II, we outperform the current best method with an accuracy of 86.56%, demonstrating our model’s ability to effectively learn expressions.

### B. Performance on 4D FER

We conducted extensive experiments on the BU-4DFE dataset, which consists of posed video clips from 101 subjects with six facial expressions. As shown in Table III, our model achieves a high accuracy of 99.36%, outperforming all competing methods by a significant margin. Compared to the top-performing state-of-the-art method [51], which achieved 96.50%, our model outperforms it by 2.86%, highlighting the effectiveness of our multiview architecture with its innovative loss function and augmentation strategies.

### C. Towards Spontaneous 4D FER

We evaluated our model on the BP4D-Spontaneous dataset, which includes 41 subjects showing spontaneous expressions, including nervousness and pain. As shown in Table IV, our method outperforms [49] by 5.99% and [52] by 4.02% in recognition accuracy. Additionally, we conducted cross-dataset evaluations [40], [53] by using BU-4DFE for

TABLE III

PERFORMANCE (%) COMPARISON OF 4D FER WITH THE STATE-OF-THE-ART METHODS ON THE BU-4DFE DATASET.

Method	Experimental Settings	Accuracy ( $\uparrow$ )
Sandbach <i>et al.</i> [19]	6-CV, Sliding window	64.60 (34.76 $\uparrow$ )
Fang <i>et al.</i> [21]	10-CV, Full sequence	75.82 (23.54 $\uparrow$ )
Xue <i>et al.</i> [47]	10-CV, Full sequence	78.80 (20.56 $\uparrow$ )
Sun <i>et al.</i> [17]	10-CV, -	83.70 (15.66 $\uparrow$ )
Zhen <i>et al.</i> [48]	10-CV, Full sequence	87.06 (12.30 $\uparrow$ )
Yao <i>et al.</i> [49]	10-CV, Key-frame	87.61 (11.75 $\uparrow$ )
Fang <i>et al.</i> [20]	10-CV, -	91.00 (8.36 $\uparrow$ )
Li <i>et al.</i> [23]	10-CV, Full sequence	92.22 (7.14 $\uparrow$ )
Ben Amor <i>et al.</i> [18]	10-CV, Full sequence	93.21 (6.15 $\uparrow$ )
Zhen <i>et al.</i> [41]	10-CV, Full sequence	94.18 (5.17 $\uparrow$ )
Bejaoui <i>et al.</i> [50]	10-CV, Full sequence	94.20 (5.15 $\uparrow$ )
Zhen <i>et al.</i> [41]	10-CV, Key-frame	95.13 (4.23 $\uparrow$ )
Behzad <i>et al.</i> [51]	10-CV, Full sequence	96.50 (2.86 $\uparrow$ )
<b>AffectVLM (Ours)</b>	10-CV, Full sequence	<b>99.36</b>

TABLE IV

ACCURACY (%) COMPARISON ON THE BP4D-SPONTANEOUS DATASET.

(A) RECOGNITION		(B) CROSS-DATASET EVALUATION	
Method	Accuracy ( $\uparrow$ )	Method	Accuracy ( $\uparrow$ )
Yao <i>et al.</i> [49]	86.59 (5.99 $\uparrow$ )	Zhang <i>et al.</i> [40]	71.00 (15.07 $\uparrow$ )
Danelakis <i>et al.</i> [52]	88.56 (4.02 $\uparrow$ )	Zhen <i>et al.</i> [53]	81.70 (4.37 $\uparrow$ )
<b>AffectVLM (Ours)</b>	<b>92.58</b>	<b>AffectVLM (Ours)</b>	<b>86.07</b>

training and BP4D-Spontaneous (Task 1 and Task 8) for validation, achieving an accuracy of 86.07%. Our model surpasses [40] by 15.07% and [53] by 4.37%, demonstrating the model’s robustness and strong generalization to spontaneous expressions, making it well-suited for real-world applications.

### D. Performance Upgrade with Distributed Learning

We use NVIDIA GeForce RTX 3090 Ti to demonstrate our model’s efficiency on lower-end multi-GPU setups. As shown in Fig. 2, distributed training with 3 GPUs achieves average per-epoch times of 84s, 73s, and 50s for respective batches, reducing total training time to 11.67h, 10.13h, and 6.94h over 500 epochs, validating our model’s scalability.

### E. Ablation Study

As shown in Fig. 3, our ablation study demonstrates that augmented textual prompts and mixed-view augmentation significantly upgrades the performance, confirming their role in enhancing semantic understanding and generalization. Our AffectVLM model consistently achieves the highest accuracy, validating the effectiveness of our joint representation learning framework.

## IV. CONCLUSION

We presented AffectVLM, a vision-language model designed to integrate multiviews for a more comprehensive understanding of facial emotions from 3D/4D data. AffectVLM employed joint learning of feature representations from multiple views in a shared embedding space, optimizing them collectively. Additionally, we introduced augmented textual prompts to enhance the model’s linguistic capabilities and a novel gradient-friendly loss function to improve convergence during training while using distributed learning.

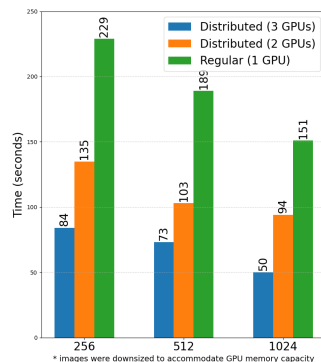


Fig. 2. Performance comparisons of distributed learning.

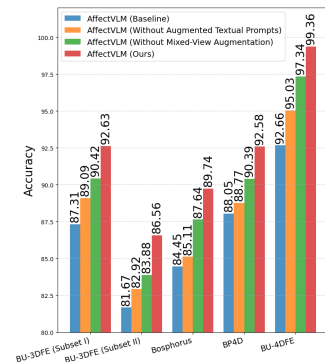


Fig. 3. Ablation study showing the impact of individual components.

## ETHICAL IMPACT STATEMENT

Our research on 3D/4D facial expression recognition leverages advanced vision-language models to enhance emotion recognition, a technology with significant potential for applications in healthcare, security, and human-computer interaction. While these models are designed to benefit society by improving user experience and enabling more empathetic interactions, they also pose potential ethical risks, such as misuse for surveillance or privacy invasion, bias, and misinterpretation in diverse social and cultural contexts. Given these risks, it is imperative to consider how this technology might inadvertently affect individuals' privacy, autonomy, and personal data security.

To mitigate potential risks, we have used publicly available datasets where participants have consented to data usage for research purposes. Additionally, our models are designed with transparent limitations, including explicit disclaimers on the scope of their applications to discourage misuse, and we encourage future work to undergo continuous ethical evaluation. Furthermore, we recognize the importance of including a diverse representation in training datasets to reduce bias and avoid disproportionate impacts on any particular group.

The potential benefits of our research include advancements in mental health monitoring, adaptive educational systems, and user-centric AI applications that support well-being and accessibility. By prioritizing transparency, privacy, and ethical governance, we believe our work balances innovation with careful risk mitigation and aims to foster societal benefit responsibly.

## REFERENCES

- [1] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human-computer interaction applications," *Neural Computing and Applications*, vol. 35, no. 32, pp. 23311–23328, 2023.
- [2] N. M. Foteinopoulou and I. Patras, "Learning from label relationships in human affect," in *Proceedings of the 30th ACM International Conference on Multimedia*, vol. 33 of *MM '22*, ACM, 2022.
- [3] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, "Affective computing in education: A systematic review and future research," *Computers & Education*, vol. 142, 2019.
- [4] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE TPAMI*, vol. 38, no. 8, pp. 1548–1568, 2016.
- [5] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.
- [6] Y.-J. Liu, B. Wang, L. Gao, J. Zhao, R. Yi, M. Yu, Z. Pan, and X. Gu, "4d facial analysis: A survey of datasets, algorithms and applications," *Computers & Graphics*, vol. 115, pp. 423–445, 2023.
- [7] H. Li et al., "3d facial expression recognition via multiple kernel learning of multi-scale local normal patterns," in *ICPR*, 2012.
- [8] X. Li, Tao Jia, and H. Zhang, "Expression-insensitive 3d face recognition using sparse representation," in *CVPR*, pp. 2575–2582, 2009.
- [9] H. Li, H. Ding, D. Huang, Y. Wang, X. Zhao, J.-M. Morvan, and L. Chen, "An efficient multimodal 2d+ 3d feature-based approach to automatic facial expression recognition," *CVIU*, pp. 83–92, 2015.
- [10] I. Mpiiperis, S. Malassiotis, and M. G. Strintzis, "Bilinear models for 3-d face and facial expression recognition," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 3, pp. 498–511, 2008.
- [11] X. Zhao, D. Huang, E. Dellandréa, and L. Chen, "Automatic 3d facial expression recognition based on a bayesian belief net and a statistical facial feature model," in *ICPR*, pp. 3724–3727, 2010.
- [12] C. Samir et al., "An intrinsic framework for analysis of facial surfaces," *IJCV*, 2009.
- [13] A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Shape analysis of local facial patches for 3d facial expression recognition," *Pattern Recognition*, vol. 44, no. 8, pp. 1581–1589, 2011.
- [14] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2d+3d facial expression recognition with deep fusion convolutional neural network," *IEEE Transactions on Multimedia*, vol. 19, 2017.
- [15] O. K. Oyedotun, G. Demisse, A. E. R. Shabayek, D. Aouada, and B. Ottersten, "Facial expression recognition via joint deep learning of rgb-depth map latent representations," in *ICCVW*, 2017.
- [16] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *FG*, 2013.
- [17] Y. Sun, X. Chen, M. Rosato, and L. Yin, "Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 3, pp. 461–474, 2010.
- [18] B. B. Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava, "4-d facial expression recognition by learning geometric deformations," *IEEE transactions on cybernetics*, vol. 44, 2014.
- [19] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3d facial expression dynamics," *Image and Vision Computing*, 2012.
- [20] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3d/4d facial expression analysis: An advanced annotated face model approach," *Image and vision Computing*, vol. 30, no. 10, 2012.
- [21] T. Fang, X. Zhao, S. K. Shah, and I. A. Kakadiaris, "4d facial expression recognition," in *ICCVW*, 2011.
- [22] M. Reale, X. Zhang, and L. Yin, "Nebula feature: A space-time feature for posed and spontaneous 4d facial behavior analysis," in *FG*, 2013.
- [23] W. Li, D. Huang, H. Li, and Y. Wang, "Automatic 4d facial expression recognition using dynamic geometrical image network," in *FG*, 2018.
- [24] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," 2024.
- [25] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," 2024.
- [26] Y. Tian, J. Cheng, Y. Li, and S. Wang, "Secondary information aware facial expression recognition," *IEEE SPL*, 2019.
- [27] P. Jiang, B. Wan, Q. Wang, and J. Wu, "Fast and efficient facial expression recognition using a gabor convolutional network," *IEEE Signal Processing Letters*, vol. 27, pp. 1954–1958, 2020.
- [28] M. Hu, Q. Chu, X. Wang, L. He, and F. Ren, "A two-stage spatiotemporal attention convolution network for continuous dimensional emotion recognition from facial video," *IEEE Signal Processing Letters*, vol. 28, pp. 698–702, 2021.
- [29] H. Li, J.-M. Morvan, and L. Chen, "3d facial expression recognition based on histograms of surface differential quantities," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 483–494, Springer, 2011.
- [30] Q. Zhen, D. Huang, Y. Wang, and L. Chen, "Muscular movement model-based automatic 3d/4d facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1438–1450, 2016.
- [31] X. Yang, D. Huang, Y. Wang, and L. Chen, "Automatic 3d facial expression recognition using geometric scattering representation," in *IEEE FG*, 2015.
- [32] Y. Zhai, H. Bai, Z. Lin, J. Pan, S. Tong, Y. Zhou, A. Suhr, S. Xie, Y. LeCun, Y. Ma, and S. Levine, "Fine-tuning large vision-language models as decision-making agents via reinforcement learning," 2024.
- [33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.
- [34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.
- [35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.
- [36] M. Behzad, N. Vo, X. Li, and G. Zhao, "Towards reading beyond faces for sparsity-aware 3d/4d affect recognition," *Neurocomputing*, vol. 458, pp. 297–307, 2021.
- [37] M. Behzad and G. Zhao, "Self-supervised learning via multi-view facial rendezvous for 3d/4d affect recognition," in *FG 2021*, IEEE, 2021.
- [38] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *European workshop on biometrics and identity management*, 2008.
- [39] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *FG*, 2006.
- [40] X. Zhang et al., "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image & Vision Comput.*, 2014.
- [41] Q. Zhen, D. Huang, H. Drira, B. B. Amor, Y. Wang, and M. Daoudi, "Magnifying subtle facial motions for effective 4d expression recognition," *IEEE Transactions on Affective Computing*, 2017.
- [42] M. Behzad, X. Li, and G. Zhao, "Landmarks-assisted collaborative deep framework for automatic 4d facial expression recognition," in *FG*, 2020.
- [43] M. Behzad, X. Li, and G. Zhao, "Disentangling 3d/4d facial affect recognition with faster multi-view transformer," *IEEE Signal Processing Letters*, 2021.
- [44] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," *IEEE TPAMI*, 2017.
- [45] M. Sui et al., "Afnm: Adaptive fusion network with masks for 2d+ 3d facial expression recognition," in *ICIP*, 2023.
- [46] H. Li et al., "Drfer: Learning disentangled representations for 3d facial expression recognition," *IEEE FG*, 2024.
- [47] M. Xue et al., "Automatic 4d facial expression recognition using dct features," in *WACV*, 2015.
- [48] Q. Zhen et al., "Muscular movement model-based automatic 3d/4d facial expression recognition," *IEEE TMM*, 2016.
- [49] Y. Yao, D. Huang, X. Yang, Y. Wang, and L. Chen, "Texture and geometry scattering representation-based facial expression recognition in 2d+3d videos," *ACM Trans. Mult. Comput. Commun. Appl.*, 2018.
- [50] H. Bejaoui, H. Ghazouani, and W. Barhoumi, "Sparse coding-based representation of lbp difference for 3d/4d facial expression recognition," *Multimedia Tools and Applications*, 2019.
- [51] M. Behzad, N. Vo, X. Li, and G. Zhao, "Automatic 4d facial expression recognition via collaborative cross-domain dynamic image network," in *BMVC*, British Machine Vision Association Press, 2019.
- [52] A. Danelakis et al., "An effective methodology for dynamic 3d facial expression retrieval," *Pattern Recognition*, 2016.
- [53] Q. Zhen, D. Huang, H. Drira, B. B. Amor, Y. Wang, and M. Daoudi, "Magnifying subtle facial motions for effective 4d expression recognition," *IEEE Transactions on Affective Computing*, 2017.