Self-Supervised Learning via Multi-view Facial Rendezvous for 3D/4D Affect Recognition

Muzammil Behzad, Guoying Zhao*
Center for Machine Vision and Signal Analysis, University of Oulu, Finland
Email: {muzammil.behzad,guoying.zhao}@oulu.fi

***Abstract*— In this paper, we present Multi-view Facial Rendezvous (MiFaR): a novel multi-view self-supervised learning model for 3D/4D facial affect recognition. Our self-supervised learning architecture has the capability to learn collaboratively via multi-views. For each view, our model learns to compute the embeddings via different encoders and robustly aims to correlate two distorted versions of the input batch. We additionally present a novel loss function that not only leverages the correlation associated with the underlying facial patterns among multi-views but it is also robust and consistent towards different batch sizes. Finally, our model is equipped with distributed training to ensure better learning along with computational convenience. We conduct extensive experiments and report ablations to validate the competence of our model on widely-used datasets for 3D/4D FER.**

I. Introduction

Self-supervised learning aims to achieve the capability of learning useful feature representations from the provided input data on its own without the need of manually-injected human annotations. The recent advances in this regard [1], [2], [3] advocate that self-learned representations can be as competitive as the supervised representations. Most of these methods follow a similar underlying theme where the aim is to learn representations that are invariant to various distortion conditions [4]. Typically, this is achieved by maximizing the similarity of distorted samples and finding correlated patterns to help cluster similar batches of data.

Inspired by the scaling victory of self-supervised learning, we work towards developing a self-supervised learning architecture to exploit the underlying similarity stored as correlated patterns in 3D/4D faces for effective facial expression recognition (FER). Specifically, contrary to the 2D faces (e.g., [5], [6], [7], [8]), such expression recognition involves predicting emotions from 3D/4D faces with complementary spatial and temporal facial features, and the significant results [9], [10], [11], [12] have proven its merits.

Literature contains several methods to learn from the underlying 3D facial geometry. However, the most popular approaches are divided into local feature-based [13], [14], [12], template-based [15], [16], [17], curve-based [18], [19] and 2D projections-based [20], [21] methods. Over the past years, 4D FER attracted a lot of interest by allowing deep learning models to learn discriminative facial features. For instance, Yin *et al.* [22] and Sun *et al.* [23] utilized Hidden Markov Models (HMM) to learn temporal facial features via 4D facial scans. Similarly, Ben Amor *et al.* [24] used the random forest classifier to demonstrate that a deformation vector field based on Riemannian analysis can yield effective performance. Likewise, Sandbach *et al.* [25] relied on HMM and GentleBoost for learning the free-form representations of the 3D frames. Furthermore, the authors in [26] represented geometrical coordinates and its normal as feature vectors, and as dynamic local binary patterns (LBP) in another work [27] for recognizing facial expressions with support vector machine (SVM). In a similar way, the authors in [28] acquire features from polar angles and curvatures, and proposed a spatio-temporal LBP-based feature extractor for recognition.

On the other hand, Li *et al.* [29] proposed an interesting framework for automatic 4D FER via dynamic geometrical image network. They produced geometrical images by computing the differential quantities from the provided 3D facial point-clouds. The emotion prediction is then a combination of score-level fusion from the probability scores of different geometrical images. Another latest work [30] takes into account the sparse coding-based representations of LBP difference. Firstly, the authors used mesh-local binary pattern difference to extract appearance and geometric features, and then applied sparse coding to recognize facial expressions. Importantly, although all these works demonstrate desirable performance, the use of manually and locally extracted cues make these solutions potentially inconvenient.

A. Motivations

Recently, some works [31], [32] have applied self-supervised methods for 2D FER but to the best of our knowledge, the literature is missing work on self-supervised models for 3D/4D FER because utilizing such models is not straightforward and trivial. This is mainly due to the complexity and variations in the data structure of the 3D data, hence, asking for appropriate self-supervised methods. Most importantly, it is worth mentioning that a good framework should look beyond the existent learning representations to formulate a robust FER system. For instance, despite the fact that multi-views acquired from an input 3D point-cloud contain highly correlated patterns along with local dependencies, their contribution is frequently overlooked. Similarly, another desirable feature of robust frameworks is the capability of leveraging prominent emotion cues to explore its significance in the improvement of performance.

This work was supported by Infotech Oulu, and the Academy of Finland. As well, the financial supports from Riitta ja Jorma J. Takanen Foundation and Tauno Tönning Foundation are acknowledged.
* indicates corresponding author.

B. Contributions

Our proposed architecture is substantially significant to tailor multi-view data from point-clouds. Precisely, the salient features of our method are as following:

1) We propose a novel multi-view self-supervised learning architecture with the capability to learn collaboratively via multi-views, hence the name: *Multi-view Facial Rendezvous*.
2) The multi-views in our method independently aim to maximize the representations' similarity via distorted versions of batches of data sampled from the dataset.
3) Importantly, instead of relying on similar embeddings, we propose to extract representation embeddings via different encoders for each view. This significantly helps the network to learn the underlying patterns in each view and distinguish inter-class patterns.
4) We also formulate an innovative loss function to facilitate better learning via cross-correlation among multi-view patterns and gradient updates across multi-views during backward propagation.
5) For computational and practical convenience, we equip our code with faster distributed training performance.

To the best of our knowledge, this is the pioneer multi-view self-supervised learning architecture for 3D/4D facial data. Additionally, our pre-trained networks can also be used for other 3D/4D tasks such as, face recognition, face anti-spoofing, etc.

II. Proposed Method

Since our self-supervised learning model is capable of incorporating multi-views, one tremendous benefit of this incredibly robust setup is its effective and scalable implementation – which can be used almost out of the box.

A. Multi-view Facial Rendezvous (MiFaR)

We present an overview of our proposed Multi-view Facial Rendezvous (MiFaR) model in Fig. 1. The model inputs 2D images in multi-views extracted from the dataset. For each view, the idea is to independently compute the cross-correlation matrix between the embeddings of two identical networks fed with two different distorted versions of the same data, and then try to collaboratively make variants of this matrix closer to identity matrix.

Motivated by recent works [3], [4], our model first creates pairs of different distorted versions of the input data $[X_l^\alpha, X_l^\beta]$, $[X_f^\alpha, X_f^\beta]$ and $[X_r^\alpha, X_r^\beta]$ for left, front and right view, respectively. However, these distortions are obtained from a set of distribution of data augmentations, i.e., $\mathcal{T}_l, \mathcal{T}_f$ and $\mathcal{T}_r$. More importantly, to avoid over-fitting, these augmentations are different for each view. The distorted data is fed to corresponding encoders to yield output embeddings $[Y_l^\alpha, Y_l^\beta]$, $[Y_f^\alpha, Y_f^\beta]$ and $[Y_r^\alpha, Y_r^\beta]$ for computing cross-correlation matrices. Having different encoders and augmentation distributions not only realize the network to penalize it more when the embeddings have a higher anti-correlation, but it also makes it prune to over-fitting by learning only from a specific distribution. More importantly, this helps the network to generalize well to unseen data. For each view, this solution ensures that the embedding vectors from the two identical networks are similar, while minimizing the irrelevant information among these embeddings. This conceptually robust network model makes it competitive even towards the state-of-the-art methods for supervised learning for 3D/4D FER.

B. Loss Function

Our self-supervised learning framework distinguishes itself from the other frameworks by its unique and novel loss function that helps the network learn and converge faster. Unlike common approaches, we aim to collaboratively optimize the loss via each view. We formulate our loss function as following:

$$\mathcal{L}_{MiFaR} \triangleq \overbrace{\underbrace{\sum_i \frac{(1-\mathcal{C}_{w,ii})^2}{2}}_{\text{weighted invariance}} + \underbrace{\frac{\mu_w}{2}\sum_i\sum_{i\neq j}\mathcal{C}_{w,ij}^2}_{\text{weighted redundancy}}}^{\mathcal{L}_w\text{: weighted loss}} + \underbrace{\sum_\theta \frac{\theta}{2}\left(\overbrace{\sum_x (1-\mathcal{C}_{\theta,xx})^2}^{\text{view invariance}} + \overbrace{\mu_\theta \sum_x\sum_{x\neq y}\mathcal{C}_{\theta,xy}^2}^{\text{view redundancy}}\right)}_{L_\theta\text{: view loss}}, \tag{1}$$

where μ_w and μ_θ are the positive constants trading off the importance of the invariance and redundancy terms of the weighted and view loss, respectively in (1). Here, $\theta = \{20°, 0°, -20°\}$ is the rotation angle for each view, while $\mathcal{C}_w$ refers to the weighted cross-correlation matrix computed as $\mathcal{C}_w = {}_l\mathcal{C}_l + {}_f\mathcal{C}_f + {}_r\mathcal{C}_r$. The variables ${}_l$, ${}_f$ and ${}_r$ are weights of the left, front and right view, respectively, while $\mathcal{C}_l$, $\mathcal{C}_f$ and $\mathcal{C}_r$ are the corresponding cross-correlation matrices computed between the outputs of two identical networks for each view:

$$\mathcal{C}_{\theta,ab} \triangleq \frac{\sum_k y^\alpha_{\theta,k,a} y^\beta_{\theta,k,b}}{\sqrt{\sum_k (y^\alpha_{\theta,k,a})^2}\sqrt{\sum_k (y^\beta_{\theta,k,b})^2}}, \forall \theta, \tag{2}$$

where a, b index the vector dimensions of the the networks' outputs, while k indexes the batch samples. Theoretically from (1), we demonstrate the learning-friendly nature of our loss function which also advocates its efficiency for faster network training. In fact, the weighted and view invariance terms aim at pushing its diagonal elements closer to the diagonal elements of the identity matrix, hence, making the embeddings invariant to the applied distortions for effective learning. Similarly, the redundancy terms try to decorrelate the different vector components of the embeddings, hence, removing irrelevant information for faster learning.

C. Distributed Performance

An important feature of our model is its ability to learn in a distributed manner. For faster training capabilities, we use PyTorch's distributed communication package (`torch.distributed`), that helps distribute the processing requirements over available resources. To facilitate

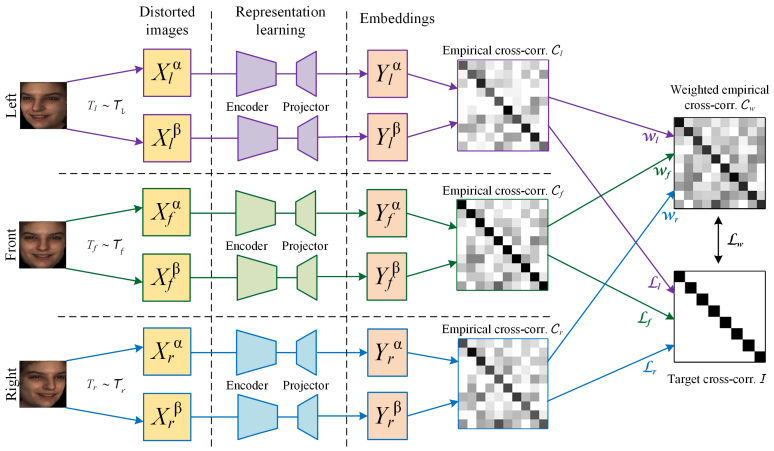

Fig. 1. Illustration of our proposed multi-view facial rendezvous (MiFaR) model.

this, we utilize NVIDIA Collective Communications Library (NCCL). Importantly, to avoid inevitable synchronization issues of distributed processes, especially in case of multiple GPUs/nodes, we dynamically allocate available and accessible network's `<port>` and `<IP>` to form a dynamic TCP communication socket. This solution allows efficient communication primitives towards multi-process parallelism for resource-efficient 3D/4D FER.

III. Results and Discussions

We use Bosphorus [33], BU-3DFE [34], BU-4DFE [22] and BP4D-Spontaneous [35] datasets to validate our model. Following previous works [29], [36], [37], [38], [39], we first compute projected 2D images in multi-views from the 3D/4D point cloud data. For video data, we use rank pooling [40] for each view [37] to form dynamic images which are then fed to the model. As far as we know, there are no available self-supervised methods for 3D/4D FER yet, so we compare our method with all the supervised methods. Importantly, a 10-fold subject-independent cross-validation (CV) is used for the experiments. Lastly, we bypass FC layers of encoders in training, and finetune it for the downstream task.

A. Performance on 3D FER

Following existing protocols [20], [21], the BU-3DFE dataset with 101 subjects is grouped into: Subset I – the standard dataset including expressions with two higher levels of intensities, and Subset II – rarely applied in 3D FER, containing all four levels of intensities except the 100 neutral samples. In Bosphorus dataset, only 65 subjects perform the six expressions with each subject. Table I summarizes the accuracy results from our extensive experiments for 3D FER advocating the effectiveness of our method. Specifically, for Subset I and Bosphorus datasets, we show that, although slightly behind by 0.78% and 1.16%, respectively, our model nearly reaches the results by state-of-the-art method [21]. With Subset II, however, we surpass the prediction results by 1.34% when compared against the most efficient state-of-the-art supervised method. These results highlight that despite being a self-supervised method, our model is capable of learning expressions effectively.

TABLE I

Accuracy (%) comparisons with state-of-the-art methods on the BU-3DFE Subset I and Subset II, and Bosphorus datasets.

Method	Subset I (↑↓)
Zhen *et al.*[10]	84.50 (4.03↑)
Yang *et al.*[11]	84.80 (3.73↑)
Li *et al.*[12]	86.32 (2.21↑)
Li *et al.*[20]	86.86 (1.67↑)
Oyedotun *et al.*[21]	89.31 (0.78↓)
MiFaR (Ours)	**88.53**

Method	Subset II (↑↓)	Bosphorus (↑↓)
Li *et al.*[12]	80.42 (2.25↑)	79.72 (0.88↓)
Yang *et al.*[11]	80.46 (2.21↑)	77.50 (1.34↑)
Li *et al.*[20]	81.33 (1.34↑)	80.00 (1.16↓)
MiFaR (Ours)	**82.67**	**78.84**

B. Performance on 4D FER

To compare the performance on 4D FER, we carry out extensive experiments on the popular BU-4DFE dataset which contains posed video clips of 101 subjects with six facial expressions. In Table II, we demonstrate that with additional information from temporal domain in terms of 4D facial data, our model attains higher recognition results and reaches the accuracies of competing supervised methods for 4D FER. We report that our model outperform most of the methods by a considerable margin thanks to its effective multi-view

TABLE II

Performance (%) comparison of 4D FER with the state-of-the-art methods on the BU-4DFE dataset.

Method	Experimental Settings	Accuracy (↑↓)
Sandbach *et al.*[25]	6-CV, Sliding window	64.60 (31.16↑)
Fang *et al.*[27]	10-CV, Full sequence	75.82 (19.94↑)
Xue *et al.*[41]	10-CV, Full sequence	78.80 (16.96↑)
Sun *et al.*[23]	10-CV, -	83.70 (12.06↑)
Zhen *et al.*[42]	10-CV, Full sequence	87.06 (8.7↑)
Yao *et al.*[43]	10-CV, Key-frame	87.61 (8.15↑)
Fang *et al.*[26]	10-CV, -	91.00 (4.76↑)
Li *et al.*[29]	10-CV, Full sequence	92.22 (3.54↑)
Ben Amor *et al.*[24]	10-CV, Full sequence	93.21 (2.55↑)
Zhen *et al.*[36]	10-CV, Full sequence	94.18 (1.58↑)
Bejaoui *et al.*[30]	10-CV, Full sequence	94.20 (1.56↑)
Zhen *et al.*[36]	10-CV, Key-frame	95.13 (0.63↑)
Behzad *et al.*[37]	10-CV, Full sequence	96.50 (0.74↓)
MiFaR (Ours)	10-CV, Full sequence	**95.76**

architecture and collaborative nature of the proposed loss function. In the table, we also show the difference in accuracies against each method. Specifically, by comparing with the most accurate state-of-the-art supervised method [37], our method lag behind by merely 0.74%, reaching the highest accuracy of 95.76%. Such superior performance illustrates that our novel self-supervised method offers a tremendous solution via collaboration among learned embeddings from multi-views, without the need of manually-injected labels.

C. Towards Spontaneous 4D FER

For spontaneous 4D FER, we use the BP4D-Spontaneous dataset which contains a total of 41 subjects showing spontaneous expressions with two additional expressions: nervousness and pain. In Table III, we compare the recognition performance and also report the results of cross-dataset evaluation. For recognition, our method outperforms [43] by 0.55%, while remains slightly behind by 1.42% when compared against [44].

More importantly, following the experimental settings in [35], [45], we also show cross-dataset evaluations to highlight our model's generalizability and robustness. For this, the BU-4DFE dataset is used for training, while a subset of the BP4D-Spontaneous dataset (i.e., Task 1 and Task 8, containing happy and disgust expressions) is used for validation. As illustrated in the table, our method achieves a notable accuracy of 79.05%, leaving behind [35] by 8.05%, while lagging behind by [45] by merely 2.65%, thereby, demonstrating promising results indicating its robustness. Consequently, our model shows the potential to be generalized well to spontaneous situations making it desirable for real-world scenarios.

D. Ablations

Distributed Learning: We mainly use NVIDIA Ampere A100 GPUs (machine-C) for experiments, but we also compare the distributed performance on GP100GL, NVIDIA Tesla P100-PCIE GPUs (machine-A) and Xeon Gold 6230, NVIDIA Volta V100 GPUs (machine-B). In Fig. 2, we compare the average time to complete one epoch in training. With 1000 epochs, the reported training times for 256×256, 512×512 and 1024×1024 on machine-C are roughly 32.5 hours, 21.1 hours, and 17.7 hours, respectively. This shows that distributed processing produces better results by leveraging available resources efficiently.

Loss Function: We also discuss the ablation of our loss function illustrated in Fig. 3. Specifically, we compare the results from our proposed loss function $\mathcal{L}_{MiFaR}$ with the view loss $\mathcal{L}_{\theta}$ to show the role of our collaborative strategy.

TABLE III

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

(A) RECOGNITION

Method	Accuracy (↑↓)
Yao *et al.*[43]	86.59 (0.55↑)
Danelakis *et al.*[44]	88.56 (1.42↓)
MiFaR (Ours)	**87.14**

(B) CROSS-DATASET EVALUATION

Method	Accuracy (↑↓)
Zhang *et al.*[35]	71.00 (8.05↑)
Zhen *et al.*[45]	81.70 (2.65↓)
MiFaR (Ours)	**79.05**

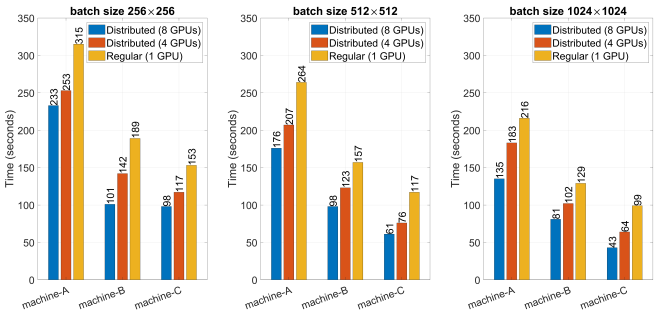

Fig. 2. Comparison of distributed learning on different machines.

We notice that our proposed loss function results in better performance by converging to a smaller loss value from the beginning. To validate the consistency and robustness of $\mathcal{L}_{MiFaR}$, we conduct experiments with different batch sizes on machine-C, i.e., 256×256, 512×512 and 1024×1024. It can be clearly seen that $\mathcal{L}_{MiFaR}$ shows dominant performance by exploiting collaboration across multi-views via the weighted cross-correlation term in $\mathcal{L}_{w}$ that helps identify similar patterns in the 3D facial structure.

Effect of Embeddings: Similarly in Fig. 3, we show the effect of using same and different encoders/embeddings for multi-views on the model's performance. We find that while using same embeddings could be equally useful in converging the loss, using different embeddings not only yields this faster but also helps achieve lower loss. More importantly, as demonstrated, this trend holds true for all batch sizes used in the experiments.

IV. CONCLUSION

We presented Multi-view Facial Rendezvous (MiFaR): a novel multi-view self-supervised learning model for 3D/4D facial affect recognition. MiFaR is equipped with the capability to learn collaboratively via multi-views in a self-supervised fashion. Our proposed loss function leverages the correlation associated with the underlying facial patterns among multi-views. With the help of several experiments, we showed that our model not only demonstrates a superior performance, but it is also robust towards 3D/4D FER.

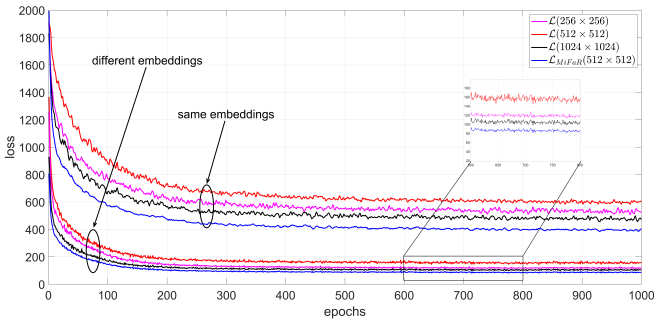

Fig. 3. Ablation comparisons of the loss under different scenarios.

REFERENCES

[1] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *NeurIPS*, 2020.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[3] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - a new approach to self-supervised learning," in *NeurIPS*, 2020.

[4] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *ICML*, 2021.

[5] R. Gao, F. Yang, W. Yang, and Q. Liao, "Margin loss: Making faces more separable," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 308–312, 2018.

[6] Y. Tian, J. Cheng, Y. Li, and S. Wang, "Secondary information aware facial expression recognition," *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1753–1757, 2019.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] X. Yang, D. Huang, Y. Wang, and L. Chen, "Automatic 3d facial expression recognition using geometric scattering representation," in *IEEE FG*, 2015.

[12] 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," *Computer Vision and Image Understanding*, vol. 140, pp. 83–92, 2015.

[13] H. Li et al., "3d facial expression recognition via multiple kernel learning of multi-scale local normal patterns," in *ICPR*, 2012.

[14] X. Li, Tao Jia, and H. Zhang, "Expression-insensitive 3d face recognition using sparse representation," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2575–2582, 2009.

[15] O. Ocegueda, T. Fang, S. K. Shah, and I. A. Kakadiaris, "Expressive maps for 3d facial expression recognition," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1270–1275, 2011.

[16] I. Mpiperis, 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.

[17] 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.

[18] C. Samir et al., "An intrinsic framework for analysis of facial surfaces," *IJCV*, 2009.

[19] 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.

[20] 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, no. 12, pp. 2816–2831, 2017.

[21] 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*, pp. 3161–3168, 2017.

[22] 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.

[23] 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.

[24] 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, no. 12, pp. 2443–2457, 2014.

[25] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3d facial expression dynamics," *Image and Vision Computing*, 2012.

[26] 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, pp. 738–749, 2012.

[27] T. Fang, X. Zhao, S. K. Shah, and I. A. Kakadiaris, "4d facial expression recognition," in *ICCVW*, 2011.

[28] M. Reale, X. Zhang, and L. Yin, "Nebula feature: A space-time feature for posed and spontaneous 4d facial behavior analysis," in *IEEE FG*, 2013.

[29] W. Li, D. Huang, H. Li, and Y. Wang, "Automatic 4d facial expression recognition using dynamic geometrical image network," in *IEEE FG*, 2018.

[30] H. Bejaoui, H. Ghazouani, and W. Barhoumi, "Sparse coding-based representation of lbp difference for 3d/4d facial expression recognition," *Multimedia Tools and Applications*, vol. 78, no. 16, pp. 22773–22796, 2019.

[31] L. Lu, L. Tavabi, and M. Soleymani, "Self-supervised learning for facial action unit recognition through temporal consistency," in *BMVC*, 2020.

[32] S. Athar, Z. Shu, and D. Samaras, "Self-supervised deformation modeling for facial expression editing," in *FG*, 2020.

[33] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *European workshop on biometrics and identity management*, 2008.

[34] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *FG*, 2006.

[35] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692 – 706, 2014.

[36] 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.

[37] M. Behzad, N. Vo, X. Li, and G. Zhao, "Automatic 4d facial expression recognition via collaborative cross-domain dynamic image network," in *British Machine Vision Conference*, 2019.

[38] M. Behzad, N. Vo, X. Li, and G. Zhao, "Landmarks-assisted collaborative deep framework for automatic 4d facial expression recognition," in *FG*, 2020.

[39] M. Behzad, N. Vo, X. Li, and G. Zhao, "Towards reading beyond faces for sparsity-aware 3d/4d affect recognition," *Neurocomputing*, 2021.

[40] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," *IEEE TPAMI*, vol. 40, no. 12, pp. 2799–2813, 2017.

[41] M. Xue et al., "Automatic 4d facial expression recognition using dct features," in *WACV*, 2015.

[42] 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.

[43] 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.*, vol. 14, 2018.

[44] A. Danelakis, T. Theoharis, I. Pratikakis, and P. Perakis, "An effective methodology for dynamic 3d facial expression retrieval," *Pattern Recognition*, vol. 52, 2016.

[45] 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.