Leveraging Foundation Models and 3D Facial Reconstruction for Micro-Expression Classification

Le Cong Thuong[1], Hai-Chau Nguyen-Le[1], Tu Nguyen Luu[1], Thi Duyen Ngo[1] and Thanh Ha Le[1,*]

[1]University of Engineering and Technology, Vietnam National University, 144 Xuan Thuy Street, Cau Giay, Hanoi 11300, Vietnam

Abstract

Micro-expressions, characterized by their subtle intensity and brief duration, present significant challenges for automatic recognition systems. Traditional 2D image-based methods often struggle with variations in illumination, pose, and occlusions, limiting their effectiveness. To address these challenges, we propose a hybrid framework that integrates large-scale vision foundation models with advanced 3D facial reconstruction techniques. By combining transferable visual embeddings from models such as RADIOv2.5, and SigLIPv2 with low-dimensional expression coefficients from 3D pipelines like SMIRK, FaceVersev4, and 3DDFAv3, our approach is designed to capture both the rich appearance information from 2D frames and the explicit geometric information from 3D reconstructions crucial for micro-expression analysis. Evaluated on the low-data regime of the 4DME dataset of the public Kaggle Micro-Expression Challenge, the proposed method outperforms every single-modality baseline; one configuration achieves a top-three leaderboard ranking. These findings underscore the synergy between appearance-centric pre-training and geometry-aware modelling, establishing a robust baseline for multimodal micro-expression analysis.[1]

Keywords

micro-expression classification, vision foundation models, 3d facial reconstruction models

1. Introduction

The automatic recognition of micro-expressions—fleeting, involuntary facial movements that betray genuine human emotion—presents a formidable challenge with profound implications for domains ranging from clinical psychology to national security [1]. Their notoriously short duration, subtle intensity, and the inherent scarcity of high-quality annotated data have historically impeded the development of robust recognition systems. Traditional approaches, which primarily operate on 2D image sequences, are often brittle, struggling to disentangle meaningful expressive cues from confounding variations in head pose, illumination, and occlusions, thus limiting their accuracy and generalizability in real world settings.

Two powerful yet largely independent streams of research offer a promising path forward. First, advances in 3D facial modeling provide a robust mechanism to overcome the limitations of 2D analysis. Parametric models like FLAME [2], built from thousands of 3D scans [3], enable the decomposition of faces into interpretable shape, expression, and pose parameters. State-of-the-art reconstruction techniques such as SMIRK [4], FaceVerse [5], and 3DDFA [6] leverage these models to infer detailed 3D geometry and dynamics from standard 2D images, effectively normalizing for pose and lighting variations. The emergence of specialized datasets like 4DME [7], which provides high-fidelity 4D captures of spontaneous micro-expressions, is crucial for training and validating these geometry-aware methods.

Second, the paradigm of foundation vision models, including powerful architectures like CLIP [8], DINOv2 [9], RADIOv2.5 [10], and SigLIPv2 [11], has revolutionized visual representation learning. Pre-trained on web-scale datasets, these models learn exceptionally rich and generalizable features

[1]The evaluation code to validate our results is publicly available at https://github.com/hysup-page/kaggle-4dmr-challenge.git.

4DMR@IJCAI25: International IJCAI Workshop on 1st Challenge and Workshop for 4D Micro-Expression Recognition for Mind Reading, August 29, 2025, Guangzhou, China.

*Corresponding author.

ltha@vnu.edu.vn (T. H. Le)

capable of capturing nuanced visual patterns, making them ideal candidates for detecting the subtle facial movements characteristic of micro-expressions.

While both 3D reconstruction and foundation models offer compelling advantages, they have largely been explored in isolation for this task. The central hypothesis of this work is that a symbiotic fusion of these two paradigms can unlock new levels of performance in micro-expression recognition. We posit that by conditioning powerful foundation models on explicit 3D facial geometry, we can create a system that is not only more accurate but also more robust. This leads to our primary research question: Can the integration of foundation vision models with explicit 3D facial attributes significantly improve the accuracy and generalizability of micro-expression classification?

To this end, our primary contributions are as follows:

1. We present a comprehensive benchmark that systematically evaluates the interplay between leading foundation vision models (RADIOv2.5, SigLIPv2) and state-of-the-art 3D facial reconstruction techniques (SMIRK, FaceVerse, 3DDFA). Using the F1-score on the 4DME dataset, our analysis provides critical insights into the most effective combinations for micro-expression classification.

2. We introduce an integrated framework that achieves high performance, validated by a top-three placement in a recent Kaggle micro-expression classification challenge. This result underscores competitive edge of our proposed fusion of geometric and visual representation learning.

This study marks a significant step towards developing more reliable and principled micro-expression recognition systems, paving the way for more robust systems suitable for applications in human-computer interaction and mental health diagnostics.

2. Related Work

Recent progress in computer vision, particularly foundation models and 3D facial reconstruction techniques, significantly advances micro-expression recognition by addressing challenges related to subtle visual cues and pose variations.

Foundation Models for Image Representation

Modern foundation models, pre-trained on large-scale datasets, provide robust generalizable features critical for micro-expression analysis. While models such as CLIP [8], DINOv2 [9], and SigLIP [12] demonstrate strong general performance, recent advances such as RADIOv2.5 [10] and SigLIPv2 [11] stand out as current state-of-the-art (SOTA) approaches. RADIOv2.5 employs a powerful Vision Transformer (ViT) to extract holistic and detailed dense visual embeddings, whereas SigLIPv2 enhances multilingual vision-language alignment, excelling in zero-shot classification and feature transfer, thus may help effectively capturing subtle micro-expression cues.

3D Facial Reconstruction Models

Advanced 3D reconstruction methods mitigate issues inherent in 2D analyses, such as illumination and pose variations. Prominent models like SMIRK [4], employing a self-supervised neural synthesis approach, and FaceVerse [5], with its fine-grained detail-controllable 3D Morphable Model (3DMM), are particularly effective for capturing subtle facial expressions. Additionally, robust frameworks such as DDFA [6], using cascaded CNNs, provide consistent and precise 3D face alignment critical for analyzing subtle facial deformations and micro-expressions.

While these two fields have progressed in parallel, the optimal strategy for fusing state-of-the-art foundation models with diverse 3D reconstruction pipelines for micro-expression recognition remains an open question. Our work directly addresses this gap by systematically evaluating and proposing an effective fusion architecture.

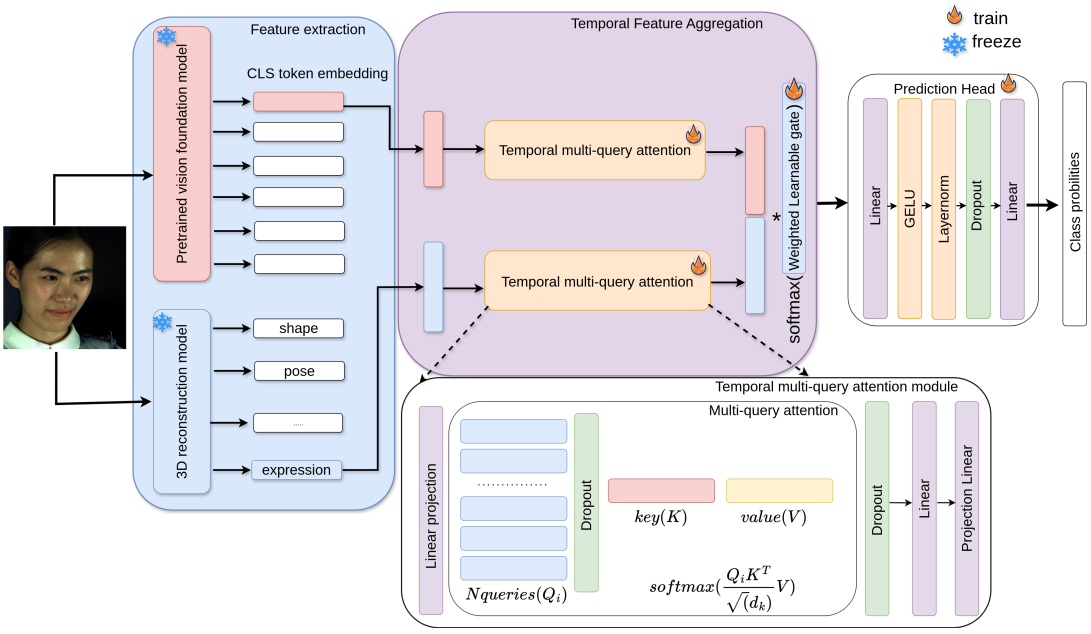

Figure 1: Overview of the proposed hybrid fusion pipeline. For a given sequence, 2D appearance features and 3D expression coefficients are extracted frame-wise. Each modality is processed by a dedicated temporal attention pool (MQA). A learnable gate then computes a weighted sum of the pooled features, which is fed to a final MLP prediction head. Trainable components are highlighted.

3. Methodology

Our framework tackles micro-expression recognition in two stages (Fig. 1). First, we distil a *hybrid* representation that marries geometry-aware 3D expression coefficients with appearance-rich 2D embeddings. Second, a lightweight attention-based temporal encoder models the evolution of these features and yields the final class probabilities.

3.1. Hybrid Feature Extraction

To faithfully capture the low-intensity, transient muscle activations that characterise micro-expressions, we fuse **(i)** 3D expression parameters that preserve subtle geometric displacements and **(ii)** 2D appearance cues that encode photometric and texture patterns.

3.1.1. 3D expression coefficients

Given an input frame, a generic 3D Morphable Model (3DMM) decomposes the face into *shape* ($\boldsymbol{\beta}$), *expression* ($\boldsymbol{\psi}$), and *pose* ($\boldsymbol{\theta}$). Because micro-expressions are chiefly conveyed through deformations of the facial musculature, we retain only the expression vector $\boldsymbol{\psi} \in \mathbb{R}^{d_{3D}}$. To evaluate the influence of the underlying reconstruction engine, we extract $\boldsymbol{\psi}$ with three complementary 3DMM pipelines:

- **FaceVersev4** [5] ($d_{3D} = 171$) — a high-capacity PCA model that excels at fine-grained geometry.
- **3DDFAv3** [6] ($d_{3D} = 64$) — a cascaded CNN regressor designed for real-time tracking.
- **SMIRK** [4] ($d_{3D} = 50$) — a self-supervised network explicitly optimised for micro-expression cues.

While these pipelines produce expression vectors ($\boldsymbol{\psi}$) of varying dimensions, they are all designed to represent facial muscle activations as low-dimensional blendshape coefficients. Our framework is designed to be agnostic to the specific 3D basis, learning modality-specific dynamics for each.

3.1.2. 2D appearance embeddings

We complement geometry with holistic appearance features extracted by Vision Transformers (ViTs). For each frame, the [CLS] token is harvested from two large-scale, pre-trained models:

- **RADIOv2.5** [10] ($d_{2D} = 1538$) — trained on a broad corpus, capturing long-range dependencies.
- **SigLIPv2** [11] ($d_{2D} = 1024$) — jointly vision-language pre-trained, sensitive to localised appearance changes.

3.1.3. Temporal Pooling and Fusion

Let the per-frame appearance feature be $\mathbf{f}_t^{\text{2D}} \in \mathbb{R}^{d_{\text{2D}}}$ and the geometric feature be $\mathbf{f}_t^{\text{3D}} \in \mathbb{R}^{d_{\text{3D}}}$. Each modality's sequence of features, $\mathbf{F} = \{\mathbf{f}_t\}_{t=1}^{T}$, is processed by a dedicated Multi-Query Attention (MQA) block [13] to model temporal dynamics.

Multi-Query Attention. We use Multi-Query Attention (MQA) to efficiently summarize the temporal dynamics within each feature sequence. MQA reduces the computational complexity of standard attention by utilizing a single key ($\mathbf{K}$) and value ($\mathbf{V}$) projection, which is shared across all query heads.

Given an input sequence $\mathbf{F} \in \mathbb{R}^{T \times d_{in}}$, where T is the sequence length, it is first projected into key and value matrices, $\mathbf{K} = \mathbf{F}\mathbf{W}_K$ and $\mathbf{V} = \mathbf{F}\mathbf{W}_V$, where $\mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{in} \times d_k}$. A separate set of learnable query vectors, forming a query matrix $\mathbf{Q} \in \mathbb{R}^{N_q \times d_k}$, then interact with this shared representation. The number of queries, N_q, is a modality-specific hyperparameter, denoted as N_q^{3D} for geometry and N_q^{2D} for appearance.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{1}$$

The output is a matrix of N_q feature vectors, with each vector representing a different learned summary of the sequence. We then apply mean pooling across these vectors to produce a single, fixed-size feature descriptor ($\mathbf{z}_{\text{2D}}$ and $\mathbf{z}_{\text{3D}}$) for each modality, effectively capturing its temporal characteristics.

Learnable Gating. To fuse the modalities, two trainable gating parameters $\boldsymbol{\gamma} = (\gamma_{\text{3D}}, \gamma_{\text{2D}})$ are used to compute a dynamic weighting. These are SOFTMAX-normalised, $\boldsymbol{g} = \text{softmax}(\boldsymbol{\gamma})$, and the final descriptor is the weighted sum: $\tilde{\mathbf{z}} = g_{\text{3D}}\mathbf{z}_{\text{3D}} + g_{\text{2D}}\mathbf{z}_{\text{2D}}$.

Prediction head. Finally, a lightweight MLP takes the fused descriptor $\tilde{\mathbf{z}}$ and produces class logits:

$$\hat{\mathbf{y}} = \text{Linear}\Big(\text{Drop}\big(\text{LN}\big(\text{GELU}\big(\text{Linear}(\tilde{\mathbf{z}})\big)\big)\big)\Big). \tag{2}$$

Here LN and Drop denote Layer Normalisation and dropout, respectively. The trainable components of this fusion architecture are highlighted in Fig. 1.

4. Experiments

4.1. Experimental setup

Dataset. We follow the official protocol of the 4DMR subset of 4DME [7]: 100 micro-expression sequences (*train*) and 28 sequences (*test*) from 24 culturally diverse subjects, each already trimmed to the active interval. To manage class imbalance within the multi-label setting, we followed a stratified 3-fold cross-validation strategy that preserves the distribution of label combinations in each split [14].[1]

[1]Specifically, we used the MultilabelStratifiedKFold implementation from the iterative-stratification Python library, available at https://github.com/trent-b/iterative-stratification.

Pre-processing. Each frame is vertically halved to expose left–right asymmetry, then a MediaPipe face detector [15] on the first frame fixes a square crop that is reused for the entire clip, guaranteeing temporal alignment. Crops are resized to 1024×1024 (Lanczos) before 2D feature extraction.

Metric. Unless stated otherwise we report **Macro-F1**—the unweighted mean of per-class F1—averaged over three folds.

4.2. Implementation details

All models share the optimiser and schedule of Table 1. Sequences are uniformly resampled to 18 frames; nearest-frame duplication fills shortages. Feature-level dropout regularises both modalities. All models were trained on a single NVIDIA RTX 3090 GPU.

Table 1
Training hyper-parameters (identical for 2D, 3D, and fusion models).

Optimiser	AdamW	Epochs	1000
Loss	Focal (γ=2, label smoothing=0.05)	Batch size	32
Lr	5×10^{-4}	Seq. length	18
N_q^{3D} (3D queries)	32	N_q^{2D} (2D queries)	512
Aug. dropout	0.1	Head dropout	0.3

4.3. Results

Our primary results on the 4DMR are summarized in Table 2. The findings clearly demonstrate the superiority of our proposed hybrid fusion model.

Table 2
Macro-F1 (%) on the 4DMR test set. **Bold** = best within group.

Model	**Mean**	**Std**
3D geometry		
SMIRK	38.67	3.11
3DDFAv3	42.99	7.44
FaceVersev4	**53.52**	4.68
2D appearance		
RADIOv2.5	**51.69**	2.03
SigLIPv2	41.83	4.17
Hybrid (ours)		
FaceVersev4 + RADIOv2.5 (dual pool + gate)	**54.79**	2.73

Unimodal baselines. Within the 3D family, FaceVersev4 dominates, underscoring the value of a rich PCA basis for capturing sub-millimetre vertex motion. For the appearance branch, RADIOv2.5 leads. This is further supported by our ablation study (Table 3), where a naïve concatenation of features processed by a single temporal encoder performs poorly (36.23% F1), likely due to over-parameterisation and the model's inability to learn modality-specific temporal dynamics.

Hybrid fusion. Combining the two best unimodal encoders with our dual-pool and softmax gate architecture achieves the highest mean F1-score (54.79%) and, more importantly, demonstrates superior training stability. The variance is reduced to nearly a third of that of the FaceVerse-only model, confirming that appearance features add complementary information and lead to a more robust system, even in the low-data regime.

4.4. Fusion–strategy ablation

Table 3
Effect of fusion design on Macro-F1 (%). All models fuse FaceVersev4 and RADIOv2.5. Mean ± Std over five seeds.

Fusion Strategy	F1 (%)
Concat 2D + 3D → single pool	36.23 ± 6.16
Dual pool (no gate)	49.34 ± 6.91
Dual pool + **softmax gate** (ours)	**54.79 ± 2.73**

Take-aways. *(i)* Processing each modality with its own temporal attention pool before fusion provides a massive performance gain over a naïve concatenation. *(ii)* Adding a simple, learnable gating mechanism provides a further significant bump in accuracy while substantially stabilising training (i.e., reducing variance), yielding the best accuracy-variance trade-off with minimal computational overhead.

5. Conclusion

This work has tackled a longstanding challenge in micro-expression recognition by rigorously assessing the integration of high-fidelity 3D reconstruction with state-of-the-art vision-foundation models. Our extensive benchmark delivers a critical insight: 3D geometric parameters and 2D appearance embeddings offer complementary, rather than overlapping, information, forming a cornerstone for robust classification. This finding emphasizes the value of a multimodal strategy to fully capture the nuanced dynamics of micro-expressions.

While our study benchmarks two powerful foundation models, a limitation is that we have not explored the full breadth of available 2D feature extractors. A valuable direction for future work is a more systematic investigation across different families of models. This could include other general-purpose self-supervised models (e.g., SimCLR [16], MAE [17], DINOv2 [9]), supervised models (e.g., ViT [18], ConvNeXt [19]). Furthermore, incorporating features from models pre-trained specifically on the face and expression domain, such as FaRL [20], and SVFAP [21], could yield significant performance gains by leveraging domain-specific knowledge.

Future work will also focus on enhancing model interpretability by visualizing attention maps against psychological cues like FACS Action Units. We will extend this framework to other subtle behavior analysis tasks, such as pain or deception detection, and evaluate its robustness in data-scarce, zero-shot learning contexts.

Acknowledgements

This work was supported by the HORIZON-MSCA-SE-2022 PhySU-Net 241 project ACMod (grant 101130271).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-o3, and Gemini-2.5 to rephrase sentences and paragraphs in order to improve clarity, conciseness, and style. After using this tool, the author(s) carefully reviewed and edited the content as needed and take full responsibility for the final version of the publication.

References

[1] A. J. R. Kumar, B. Bhanu, Micro-expression classification based on landmark relations with graph attention convolutional network, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1511–1520.

[2] T. Li, T. Bolkart, M. J. Black, H. Li, J. Romero, Learning a model of facial shape and expression from 4d scans., ACM Trans. Graph. 36 (2017) 194–1.

[3] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, et al., 3d morphable face models—past, present, and future, ACM Transactions on Graphics (ToG) 39 (2020) 1–38.

[4] G. Retsinas, P. P. Filntisis, R. Danecek, V. F. Abrevaya, A. Roussos, T. Bolkart, P. Maragos, 3d facial expressions through analysis-by-neural-synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 2490–2501.

[5] L. Wang, Z. Chen, T. Yu, C. Ma, L. Li, Y. Liu, Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 20333–20342.

[6] Z. Wang, X. Zhu, T. Zhang, B. Wang, Z. Lei, 3d face reconstruction with the geometric guidance of facial part segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1672–1682.

[7] X. Li, S. Cheng, Y. Li, M. Behzad, J. Shen, S. Zafeiriou, M. Pantic, G. Zhao, 4dme: A spontaneous 4d micro-expression dataset with multimodalities, IEEE Transactions on Affective Computing 14 (2022) 3031–3047.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[9] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).

[10] G. Heinrich, M. Ranzinger, Y. Hongxu, Y. Lu, J. Kautz, A. Tao, B. Catanzaro, P. Molchanov, Radiov2. 5: Improved baselines for agglomerative vision foundation models, in: Proc. CVPR, volume 2, 2025, p. 6.

[11] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, et al., Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, arXiv preprint arXiv:2502.14786 (2025).

[12] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 11975–11986.

[13] N. Shazeer, Fast transformer decoding: One write-head is all you need, arXiv preprint arXiv:1911.02150 (2019).

[14] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2011, pp. 145–158.

[15] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al., Mediapipe: A framework for building perception pipelines, arXiv preprint arXiv:1906.08172 (2019).

[16] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PmLR, 2020, pp. 1597–1607.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16000–16009.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[19] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976–11986.

[20] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, F. Wen, General facial representation learning in a visual-linguistic manner, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 18697–18709.

[21] L. Sun, Z. Lian, K. Wang, Y. He, M. Xu, H. Sun, B. Liu, J. Tao, Svfap: Self-supervised video facial affect perceiver, IEEE Transactions on Affective Computing (2024).