Emotic Masked Autoencoder on Dual-views with Attention Fusion for Facial Expression Recognition

Xuan-Bach Nguyen[1*] Hoang-Thien Nguyen[2*] Thanh-Huy Nguyen[3] Nhu-Tai Do[4] Quang Vinh Dinh[5]

[1]Ho Chi Minh City University of Technology, Vietnam

[2]Posts and Telecommunications Institute of Technology, Ho Chi Minh City, Vietnam

[3]Ho Chi Minh City University of Education, Vietnam

[4]University of Economics Ho Chi Minh City-UEH Vietnam

[5]Vietnamese-German University, Vietnam

bach.nguyenspring@hcmut.edu.vn, n21dccn080@student.ptithcm.edu.vn, taidn@ueh.edu.vn

4401101074@student.hcmue.edu.vn, vinh.dq2@vgu.edu.vn

Abstract

Facial Expression Recognition (FER) is a critical task within computer vision with diverse applications across various domains. Addressing the challenge of limited FER datasets, which hampers the generalization capability of expression recognition models, is imperative for enhancing performance. Our paper presents an innovative approach integrating the MAE-Face self-supervised learning (SSL) method and multi-view Fusion Attention mechanism for expression classification, particularly showcased in the 6th Affective Behavior Analysis in-the-wild (ABAW) competition. By utilizing low-level feature information from the ipsilateral view (auxiliary view) before learning the high-level feature that emphasizes the shift in the human facial expression, our work seeks to provide a straightforward yet innovative way to improve the examined view (main view). We also suggest easy-to-implement and no-training frameworks aimed at highlighting key facial features to determine if such features can serve as guides for the model, focusing on pivotal local elements. The efficacy of this method is validated by improvements in model performance on the Aff-wild2 dataset, as observed in both training and validation contexts.

1. Introduction

Facial expressions are key in conveying emotions and are increasingly important in areas like human-computer interaction [27], healthcare [2], and driving safety [42]. Real-world static image datasets like RAF-DB [22], AffectNet [31], and FERPlus [1] have seen advancements in Facial Expression Recognition (FER) due to deep learning. However, recognizing dynamic facial expressions (DFER)
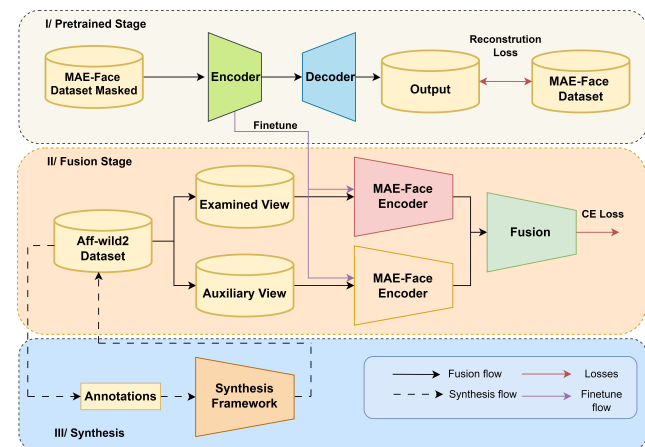

Figure 1. Our proposed pipeline for two-stage pre-training and fine-tuning with fusion, a synthesizing framework to take the informative facial feature with uni-task expression annotations.

from videos remains challenging due to data scarcity, limited dataset diversity, and the complexity of interpreting expressions over time.

Prior works nowadays usually enhance one path forward by leveraging the ample and varied SFER data as a foundational knowledge base to enhance DFER performance, considering the considerable overlap in information between SFER and DFER data. Additionally, effectively modeling the temporal dimension to capture the nuances of facial expressions over time remains a crucial challenge in DFER. While various methods utilizing 3D [7] and 2D Convolutional Neural Networks, combined with Recurrent Neural Networks or Transformer [48] architectures, have shown promise, these techniques often fall short in explicitly modeling the dynamic nature of facial expressions in

videos, thus not fully capitalizing on the temporal information available.

Researchers recently have leveraged Masked Autoencoders (MAE) [24] to pre-train models on extensive, unlabeled facial datasets, enabling the models to learn the underlying probability distribution of human faces and to reconstruct them effectively. The pre-training phase of the encoder block is thus expected to preserve critical human features that distinguish individual faces, thereby enhancing the model's versatility for a variety of downstream tasks. The work of [29] showed that it can improve performance on visual representation by incorporating a two-pass pre-training process and a two-pass fine-tuning process.

Recent studies on facial key points suggest that a focused analysis of facial regions or combining specifically the action units (AUs) [5, 23, 25, 26] or feature point tracking algorithm [3], can enhance the robustness of FER models. This combination is designed to precisely capture facial expressions, thereby directing the model's attention towards emotion-relevant facial regions for more nuanced analysis. With realizing the potential of action unit facial part by extracting and analyzing cropped AU parts of the face without relying on AU-specific annotations, we shift the focus to recognizing expressions based on overall emotional states rather than discrete facial muscle movements. This approach simplifies the labeling process and potentially reduces the subjectivity in annotations.

Many previous works have been studied about the effectiveness of exploiting multi-view networks to boost the performance of classification tasks [33, 40]. Most of these approaches mainly focus on medical applications such as breast cancer diagnosis. Integrating this method with a multi-view analysis capitalizes on the strengths of both static and dynamic expression data. This leads to the creation of a Fusion Multi-View Network that is not only more resilient in interpreting expressions in varied contexts but also reduces the dependency on extensive and precise AU annotations, focusing on the holistic portrayal of emotions. Thus, we aim to address the gaps in Dynamic Facial Expression Recognition (DFER) by building a model grounded in the realistic representation of expressions, reinforced by the insights from the rebuilt smaller static dataset and tailored to the intricacies of dynamic facial expressions by the post-processing. Our approach includes setting up two stages of multi-view analysis. The first stage involves fine-tuning with a pre-trained Masked Autoencoder for Faces (MAE-Face) model, followed by researching various mutations of multi-view networks to understand the effectiveness of fusion at multiple positions in the architecture to robust the feature extracted. We proposed four strategies for fusion types: Mean, Concat, UpDown-Mean, and UpDown-Concat. Overall, our main contributions can be summarized from Fig.1 as follows:

- **A new robust facial synthesis auxiliary view framework**: This network uniquely combines cropped AU analysis with feature point tracking, focusing on emotion-related areas with informative regions that are important for emotional representation.

- **Two-Stage Multi-View Analysis with Pretrained MAE-Face Model**: Incorporating a two-stage analysis, starting with fine-tuning a pre-trained MAE-Face model, enhances the network's ability to interpret complex facial expressions by guiding the model to focus on the important image's parts.

- **Comprehensive Study on Fusion Strategies**: By experimenting with four distinct fusion strategies: Mean, Concat, UpDown-Mean, and UpDown-Concat. We investigate the impact of integrating multi-view analysis at different stages in the network architecture, contributing to a more robust feature extraction and interpretation of emotional expressions.

2. Related Work

2.1. Facial Expression Recognition

Facial expression recognition (FER) is a cornerstone of pattern recognition research. Historically, leveraging fully supervised data has marked significant advancements in FER, as evidenced by methodologies detailed in several studies [4, 5, 43, 49]. Furthermore, multi-view strategies incorporating dual or multiple images, or supplemental augmented views, have proven to bolster performance [6, 37]. Recent trends pivot towards fortifying feature connections across divergent feature spaces. One innovative strategy introduces an attention mechanism comprising spatial and channel attention units that collaboratively hone in on pertinent spatial and channel-specific features [41]. The POSTER framework extends this concept with an integrated facial landmark detector, a foundational image backbone, cross-fusion transformer encoders, and a pyramid network, addressing inter-class similarity, intra-class variability, and sensitivity to scale in unison [30]. In video applications, the S2D model, a multi-view landmark-aware lightweight adaptation for static images, has set new benchmarks [6]. The novel MDF-HF architecture encapsulates this evolution, commencing with meticulous data preprocessing. This includes the transformation of video to image sequences, facial detection and cropping, and alignment, thereby minimizing background distractions. Details on adaptive keyframe selection, dynamic feature extraction, and the nuances of hybrid fusion follow [35].

2.2. Learning from unlabeled data

Lately, the realm of self-supervised learning has garnered attention for its ability to extract value from unlabeled
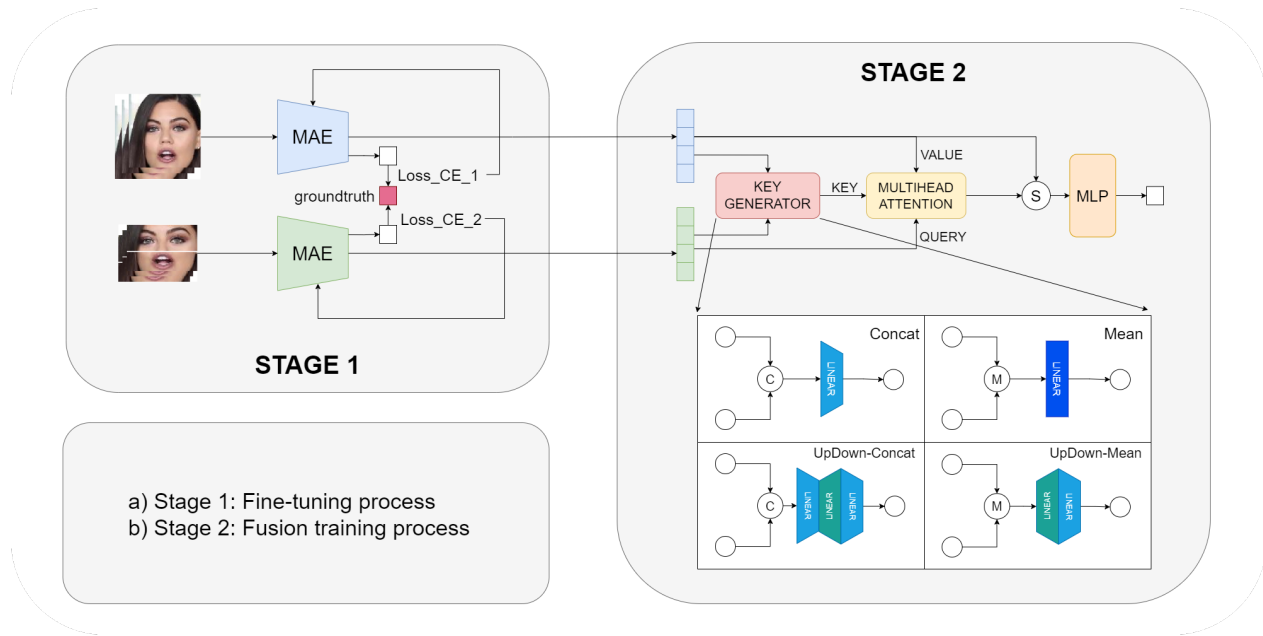
Figure 2. The architecture of our proposed model consists of two stages. First, the pre-trained model MAE is fine-tuned on two different datasets, the original dataset, and the extracted feature dataset. Secondly, we train the feature spaces of two fine-tuned models on the attention fusion models with four methods, including concat, mean, updown-concat, updown-mean, on the key generator module.

data, presenting an innovative avenue to mitigate the challenges of annotating FER datasets. These methods [32] diverge from traditional supervised learning, which relies on human-provided labels, by engaging the model in pretext tasks that do not require annotations. Such tasks include image inpainting [36], jigsaw puzzle solving [34], and contrastive learning [8], among others. For instance, some researchers [24] utilized Masked Autoencoders (MAE) to pretrain models across extensive facial recognition databases for the ABAW competitions, leading to impressive outcomes. Additionally, a semi-supervised approach introduced by [45] employs a dynamic threshold module (DTM) to effectively harness unlabeled data by adjusting the confidence threshold across various classes and throughout different stages of training.

Nonetheless, it's important to note that these pretext tasks are primarily designed for general image classification purposes, such as identifying object species, and may not be as effective in directly extracting features that are critical for recognizing facial expressions in FER tasks.

3. Methodology

In this section, we will describe our proposed approach in detail in Fig.2. At the core of our model lies an architecture composed of several critical components. Initially, we employ a pre-trained Masked Autoencoder (MAE-Face) as a primary feature extractor to capture nuanced facial expressions and affective cues. This base is further refined by

fine-tuning the MAE-Face on the Aff-wild2 dataset, ensuring our feature extraction is highly tailored to the intricacies of affective behavior. To process these features effectively, we incorporate self-attention blocks, which allow our model to discern the most relevant affective signals dynamically. The integration of these elements culminates in a Multilayer Perceptron (MLP) that serves as the final prediction mechanism, offering a precise and reliable analysis of emotional expressions.

3.1. Pre-processing

3.1.1 Clean data

First of all, in order to uphold the integrity and quality of the training dataset, we implement a rigorous data-cleaning process. This involves the systematic removal of corrupt files, non-face images, and those with insufficient resolution, thereby ensuring the robustness and reliability of the dataset for subsequent analyses and model training.

3.1.2 Synthesis Framework

The Dlib framework [9] is applied to filter out all the facial images that do not have sufficient key points for eyes, eyebrows, nose, and mouth. In the fine-tuning phases, face images are cropped by a percentage of frame sizes to extract features of the eyes and mouth. The eyes are extracted by cutting from 0.35 to 0.55 in height and 0.2 to 0.8 in width, and the mouth is cropped with the same length but at 0.7 to

0.9 in height Fig.3. Finally, image transformation is applied to resize the face image patches.

3.2. MAE-Face

The proposed model employs a Vision Transformer (ViT) pre-trained via a self-supervised learning method, leveraging a Masked Autoencoder (MAE) for the process called MAE-Face [28]. Specifically, input images are segmented into 16×16 patches, of which 75% are masked, focusing the training on reconstructing these patches from the remaining visible ones. The model is pre-trained on a comprehensive facial image dataset, including AffectNet [31], CASIA-WebFace [44], IMDB-WIKI [38], and CelebA [47], totaling 2,170,000 images, to enhance its capability in understanding facial features without relying on labels. Post pre-training, the model transitions to fine-tuning for downstream task Expression recognition on the Aff-wild2 dataset. We extracted the feature based on the pre-training and fine-tuned the MAE face for further processing.

3.3. Fusion Attention block for Emotion Recognition

In the approach employed, a fusion-based methodology is utilized for enhancing facial emotion recognition (FER) through the integration of two pre-trained models. The methodology makes use of an attention-based network, concentrating on both self-attention and local attention, to effectively combine the strengths of these models.

3.3.1 Fusion Attention Network

The proposed fusion attention network incorporates a multilayer perceptron (MLP) for combining features from two emotion recognition models. By concatenating or adding and then downsampling these features, the network applies self-attention and local attention mechanisms in the Multihead Attention Block to refine the feature representation for emotion classification.

3.3.2 Self-Attention and Local Attention

The self-attention mechanism operates by computing the attention scores based on the dot product of features from pre-trained networks, serving as the Key (K) and Query (Q), followed by a softmax function. The local attention mechanism employs convolution operations to extract more refined features. The attention formula is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right) \otimes V \quad (1)$$

where K is the Key, Q is the Query, V is the Value, and d_K is the dimensionality of the Key. This approach not only

leverages the distinct advantages of both attention mechanisms but also enhances the model's ability to recognize complex emotional states from facial expressions.

3.3.3 Skip Connection

To bolster the model further, a skip connection is introduced to preserve and reinforce prominent features, thereby enhancing the learning process and robustness of the model. The skip connection allows for the direct flow of information by adding the input directly to the output of the attention block. The formula for a skip connection, represented in LaTeX code, is as follows:

$$\text{Output} = \text{Attention}(Q, K, V) + \text{X} \quad (2)$$

Where x represents the stronger input feature. This addition of the input to the output of the attention mechanism helps in mitigating the issue of vanishing gradients, allowing for deeper network architectures without the loss of relevant feature information through the layers.

3.4. Loss function

The cross-entropy loss function is commonly used in classification tasks and calculates the difference between the predicted probability distribution and the true distribution of the classes.

The standard cross-entropy loss function is given by:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \log(\hat{y}_{ij}) \quad (3)$$

where N is the number of batch samples, C is the number of predicted classes (eight for Aff-Wild2), and y_{ij}, $\hat{y}_{ij}$ represent the ground-truth label and predicted scores, respectively.

3.5. Post-processing

Given that the Aff-Wild2 dataset is derived from the entirety of video frames and facial expressions typically evolve gradually over time, rapid changes in expressions across adjacent frames are uncommon. Furthermore, following the removal of non-face, corrupted, and low-resolution images, a sliding window approach is adopted for post-processing prediction results to ensure smoother labels. We chose a window size is 50 for our approach.

4. Experiment

In this section, we will provide a detailed description of the used datasets, the experiment setup, and the experimental results.
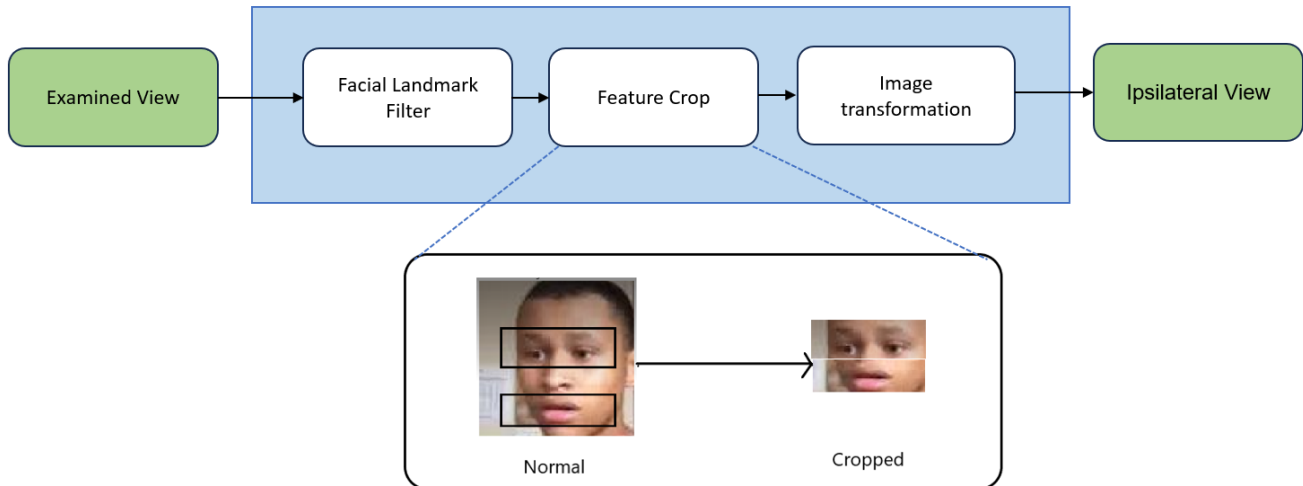
Figure 3. After the images are cropped and aligned in the Aff-wild2 dataset, the image parts that contain just the mouth and eye are extracted for further processing.

Features	Accuracy	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise	Other	Marco F1
Not crop (normal)	**0.512**	**0.624**	**0.295**	**0.188**	0.015	**0.515**	**0.448**	**0.287**	0.502	**0.359**
Eye	0.257	0.390	0.001	0.000	**0.412**	0.134	0.012	0.052	0.213	0.157
Nose	0.239	0.071	0.049	0.117	0.033	0.244	0.121	0.063	0.402	0.137
Mouth	0.324	0.388	0.072	0.011	0.406	0.250	0.246	0.110	0.433	0.239
Eye + Mouth	0.491	0.559	0.257	0.023	0.019	0.482	0.233	0.257	**0.559**	0.304
Eye + Mouth + Nose	0.443	0.551	0.192	0.015	0.233	0.512	0.289	0.165	0.495	0.302

Table 1. Results on fine-tuning process of the validation set off Aff-widl2

4.1. Dataset

Aff-wild2 [10–21, 46] is a large-scale video database for ABAW competitions. It annotated 548 videos, around 2.7M frames, into eight pre-defined categories: anger, disgust, fear, happiness, sadness, surprise, neutral, and others. Thanks to the release of this database, we conduct experiments to explore the effectiveness of our method in the ABAW challenge. In our paper, we obtain 8,000 labeled images for each class category from the Aff-Wild2 dataset through uniform sampling.

4.2. Setup

All training face images are resized to 224×224 pixels, our proposed method is implemented with the PyTorch toolbox on NVIDIA Tesla A6000 GPUs. The model is fine-tuned with the Adam optimizer for 100 iters. The learning rate for the fine-tuning process is set at 1e-4. The batch size is set to 512. The average f1 Score across all eight categories on the validation set is reported.

4.3. Metrics

The average F1 Score across all eight categories on the validation set is measured as a performance assessment.

$$P = \sum_{i=1}^{8} \frac{F1_i}{8} \quad (4)$$

where $F1$ denotes f1-score, is calculated by:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (5)$$

4.4. Results

In the fine-tuning process, we experience many methods with our approach as shown in Table 1. We have tried training the model with just eyes, nose, or mouth. After that, we try other cases by combining these objects, such as eye and mouth or eye and mouth and nose, to analyze the importance and influence of each feature on the emotions. However, we prioritize the normal feature model, which remains the original image, and the eye-mouth model, which uses the eye and mouth of the face, to feed into the fusion training process. The ratio gets each feature:

Fusion	Accuracy	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise	Other	Marco F1
Mean	0.527	0.615	0.349	0.231	0.012	0.527	0.454	0.264	0.553	0.376
Concat	**0.529**	**0.622**	**0.364**	0.241	0.018	**0.538**	0.432	**0.271**	**0.554**	**0.380**
UpDown + Mean	0.518	0.613	0.282	0.265	0.018	0.524	0.453	0.252	0.539	0.368
UpDown + Concat	0.520	0.614	0.261	**0.274**	**0.033**	0.520	**0.457**	0.268	0.549	0.372

Table 2. Results on fusion process of the validation set of Aff-wild2

- Eye: width $\in$ [0.2, 0.8], height $\in$ [0.35, 0.55]

- Mouth: width $\in$ [0.2, 0.8], height $\in$ [0.7, 0.9]

- Nose: width $\in$ [0.4, 0.6], height $\in$ [0.2, 0.8]

Next, we utilize the feature embeddings of two distinct models to train the fusion attention model. We change the key generator in many cases, such as concat methods (includes 1 dense layer), mean method (1 dense layer), UpDown-Mean(2 dense layers), and UpDown-Concat (3 dense layers). Finally, we conclude that the concat method is the main approach of our model. We can see the results in Table 2.

4.5. Ablation Studies

An ablation study was conducted to assess the contribution of different facial regions and fusion methods to the performance of a facial expression recognition model. The study's results are delineated in two tables, presenting the outcomes of the fine-tuning process and the fusion strategies, respectively. From the first table, it's evident that dif-

Method	F1
w/o. MAE finetune, Fusion, Post-processing	0.236
w/o. Fusion,Post-processing	0.359
w/o. Post-processing	0.380
Ours	0.401

Table 3. Results of overall method of the validation set of Aff-wild2

ferent facial features contribute variably to the accuracy of the model. When the model does not use any cropped features and utilizes the full face (normal), it achieves a baseline accuracy of 0.512. However, when isolating the features, the 'Eyes' alone result in a lower accuracy (0.257), which suggests that while the eyes are important for recognizing some emotions, they are not sufficient on their own however they can highlight where the expression is 'Fear'. The 'Nose' and 'Mouth' features also show limited effectiveness when used independently, with accuracies of 0.239 and 0.324, respectively.

Combining features improves the model's performance, with 'Eye + Mouth' significantly improving the accuracy to 0.491. The combination of all three features ('Eye + Mouth

+ Nose') further enhances the performance, yielding an accuracy of 0.443. This implies that a holistic representation of facial features, encompassing multiple regions, is more effective for emotion recognition tasks.

The second table explores the effectiveness of different feature fusion strategies. 'Mean' fusion leads to an accuracy of 0.527, while 'Concat' fusion slightly outperforms it with an accuracy of 0.529, suggesting that concatenation of features might be slightly more beneficial than averaging them. Furthermore, advanced fusion strategies, such as 'UpDown + Mean' and 'UpDown + Concat', yield comparable results with accuracies of 0.518 and 0.520, respectively. These methods appear to be robust, offering a balance between detail retention and abstraction.

As shown in Table 3, by implementing a sliding window approach with a window size of 50, a technique informed by the study in [39], our models attained an overall F1 macro score of 0.401. This indicates that selecting an optimal smoothing window size, denoted as k, can significantly enhance the performance metrics observed on the validation dataset.

The performance of our method in the 6th ABAW challenge as shown in Table 4, while not reaching the high mark set by the leading team, Netease Fuxi AI Lab, does manage to exceed the baseline. This indicates a promising direction for our approach, suggesting that with further development, there's potential to bridge the gap to the SOTA performance.

Teams	Test
Netease Fuxi AI Lab	**0.5005**
CtyunAI	0.3625
USTC-IAT-United	0.3534
HSEmotion	0.3414
M2-Lab-Purdue	0.3228
KBS-DGU	0.3005
SUN-CE	0.2877
Ours	0.2797
CAS-MAIS	0.265
IMLAB	0.2296
baseline	0.2250
Ours (Post challenge)	0.2866

Table 4. Result on the test set of 6th ABAW challenge

5. Conclusion

In this study, we present an efficient approach for evaluating the performance of the Fusion MAE-Face technique in uni-task expression analysis using the Aff-Wild2 dataset. Our method improves the learning of facial expressions by refining preprocessing steps that focus on facial features while reducing interference from background pixels. The model initially learns low-level features from two symmetric perspectives separately and then combines these features to learn high-level information through a process of feature fusion. Our tests on the Aff-Wild2 dataset show that starting with pre-trained MAE-Face weights and integrating fusion attention mechanisms significantly enhances the model's ability to identify meaningful features. This allows our method to outperform baseline models with less than 10% of the training data. Our methodology is innovative and provides new insights into focusing on crucial local details and exploring how different views can enhance model performance.

Acknowledgement

References

[1] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 279–283, 2016. 1

[2] Carmen Bisogni, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer. Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Transactions on Industrial Informatics*, 18(8):5619–5627, 2022. 1

[3] Daniel Canedo and António JR Neves. Facial expression recognition using computer vision: A systematic review. *Applied Sciences*, 9(21):4678, 2019. 2

[4] Jia-Ren Chang, Yong-Sheng Chen, and Wei-Chen Chiu. Learning facial representations from the cycle-consistency of face. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9680–9689, 2021. 2

[5] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991, 2021. 2

[6] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *arXiv preprint arXiv:2312.05447*, 2023. 2

[7] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 445–450, 2016. 1

[8] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 3

[9] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 3

[10] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 5

[11] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2023. 5

[12] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2023. 5

[13] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 5

[14] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 5

[15] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 5

[16] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2023. 5

[17] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. *arXiv preprint arXiv:2402.19344*, 2024. 5

[18] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 5

[19] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 5

[20] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 5

[21] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 5

[22] Shan Li, Weihong Deng, and JunPing Du. Reliable crowd-sourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 1

[23] Ximan Li, Weihong Deng, Shan Li, and Yong Li. Compound expression recognition in-the-wild with au-assisted meta multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5734–5743, 2023.

[24] Yifan Li, Haomiao Sun, Zhaori Liu, Hu Han, and Shiguang Shan. Affective behaviour analysis using pretrained model with facial prior. In *European Conference on Computer Vision*, pages 19–30. Springer, 2022. 2, 3

[25] Liqian Liang, Congyan Lang, Yidong Li, Songhe Feng, and Jian Zhao. Fine-grained facial expression recognition in the wild. *IEEE Transactions on Information Forensics and Security*, 16:482–494, 2020. 2

[26] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE, 2013. 2

[27] Zhentao Liu, Min Wu, Weihua Cao, Luefeng Chen, Jianping Xu, Ri Zhang, Mengtian Zhou, and Junwei Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE CAA J. Autom. Sinica*, 4(4):668–676, 2017. 1

[28] Bowen Ma, Rudong An, Wei Zhang, Yu Ding, Zeng Zhao, Rongsheng Zhang, Tangjie Lv, Changjie Fan, and Zhipeng Hu. Facial action unit detection and intensity estimation from self-supervised representation. *arXiv preprint arXiv:2210.15878*, 2022. 4

[29] Bowen Ma, Wei Zhang, Feng Qiu, and Yu Ding. A unified approach to facial affect analysis: the mae-face visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5923–5932, 2023. 2

[30] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, and Aibin Huang. Poster++: A simpler and stronger facial expression recognition network. *arXiv preprint arXiv:2301.12149*, 2023. 2

[31] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 1, 4

[32] Ba Hung Ngo, Ba Thinh Lam, Thanh Huy Nguyen, Quang Vinh Dinh, and Tae Jong Choi. Dual dynamic consistency regularization for semi-supervised domain adaptation. *IEEE Access*, 2024. 3

[33] Thanh-Huy Nguyen, Quang Hien Kha, Thai Ngoc Toan Truong, Ba Thinh Lam, Ba Hung Ngo, Quang Vinh Dinh, and Nguyen Quoc Khanh Le. Towards robust natural-looking mammography lesion synthesis on ipsilateral dual-views breast cancer analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2564–2573, 2023. 2

[34] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 3

[35] Bei Pan, Kaoru Hirota, Yaping Dai, Zhiyang Jia, Edwardo F Fukushima, and Jinhua She. Adaptive key-frame selection-based facial expression recognition via multi-cue dynamic features hybrid fusion. *Information Sciences*, 660:120138, 2024. 2

[36] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3

[37] Andrés Romero, Juan León, and Pablo Arbeláez. Multi-view dynamic facial action unit detection. *Image and Vision Computing*, 122:103723, 2022. 2

[38] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 10–15, 2015. 4

[39] Andrey V Savchenko. Emotieffnet facial features in uni-task emotion recognition in video at abaw-5 competition. *arXiv preprint arXiv:2303.09162*, 2023. 6

[40] Toan TN Truong, Huy T Nguyen, Thinh B Lam, Duy VM Nguyen, and Phuc H Nguyen. Delving into ipsilateral mammogram assessment under multi-view network. In *International Workshop on Machine Learning in Medical Imaging*, pages 367–376. Springer, 2023. 2

[41] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics*, 8(2):199, 2023. 2

[42] Torsten Wilhelm. Towards facial expression analysis in a driver assistance system. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–4. IEEE, 2019. 1

[43] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2021. 2

[44] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 4

[45] Jun Yu, Zhongpeng Cai, Renda Li, Gongpeng Zhao, Guochen Xie, Jichao Zhu, Wangyuan Zhu, Qiang Ling, Lei Wang, Cong Wang, et al. Exploring large-scale unlabeled faces to enhance facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5803–5810, 2023. 3

[46] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops*

(CVPRW), 2017 IEEE Conference on, pages 1980–1987. IEEE, 2017. 5

[47] Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 70–85. Springer, 2020. 4

[48] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021. 1

[49] Ce Zheng, Matias Mendieta, and Chen Chen. Poster: A pyramid cross-fusion transformer network for facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3146–3155, 2023. 2