A Unified Approach to Facial Affect Analysis: the MAE-Face Visual Representation

Bowen Ma, Wei Zhang, Feng Qiu, Yu Ding[*]
Virtual Human Group, Netease Fuxi AI Lab
{mabowen01,zhangwei05,qiufeng,dingyu01}@corp.netease.com

Abstract

Facial affect analysis is essential for understanding human expressions and behaviors, encompassing action unit (AU) detection, expression (EXPR) recognition, and valence-arousal (VA) estimation. The CVPR 2023 Competition on Affective Behavior Analysis in-the-wild (ABAW) is dedicated to providing a high-quality and large-scale Aff-wild2 dataset for identifying widely used emotion representations. In this paper, we employ MAE-Face as a unified approach to develop robust visual representations for facial affect analysis. We propose multiple techniques to improve its fine-tuning performance on various downstream tasks, incorporating a two-pass pre-training process and a two-pass fine-tuning process. Our approach exhibits strong results on numerous datasets, highlighting its versatility. Moreover, the proposed model acts as a fundamental component for our final framework in the ABAW5 competition. Our submission achieves outstanding outcomes, ranking first place in the AU and EXPR tracks and second place in the VA track.

1. Introduction

Facial affect analysis plays a vital role in diverse fields such as psychology, neuroscience, computer vision, and human-computer interaction. The advent of deep learning has brought about significant advancements in facial affect analysis, with deep neural networks being employed to learn facial expression representations. However, most existing approaches focus on specific tasks, like facial expression recognition or action unit detection, necessitating the creation of task-specific architectures and training procedures.

In this paper, we adopt MAE-Face as a unified approach for facial affect analysis, originally proposed by [35] for action unit analysis. Drawing inspiration from Masked Autoencoders [13], MAE-Face learns robust visual representations for facial affect via pre-training on a large-scale facial image dataset using a self-supervised learning scheme.

We fine-tune the pre-trained model on various datasets, and experiment with several facial affect analysis tasks, such as action unit (AU) detection, expression (EXPR) recognition, and valence-arousal (VA) estimation. In addition to incorporating the MAE-Face model into these tasks, we explore several methods to further enhance its performance. For example, we propose a two-pass pre-training method for improved model initialization and a two-pass fine-tuning method to stabilize the training process and mitigate overfitting issues.

The 5th Competition on Affective Behavior Analysis in-the-wild (ABAW5) [25] aims to tackle the challenges associated with human affective behavior analysis. To achieve this goal, the competition has developed large-scale multimodal video datasets, namely Aff-wild [23, 26, 65] and Aff-wild2 [20, 22, 24, 27–29]. Aff-wild2 consists of 598 videos with frame-wise annotations for three types of expression representations: Action Units (AU), basic expression categories, and valence-arousal (VA). ABAW5 introduces three challenges concerning the detection of these three expression representations. These datasets have significantly contributed to the progress of facial expression analysis in real-world scenarios and have accelerated the practical implementation of related industries.

Our proposed method exhibits state-of-the-art performance on multiple facial affect analysis task datasets, highlighting the adaptability and robustness of the learned facial feature representation. Moreover, the model presented in this paper serves as the base model in our final framework for the ABAW5 Competition. Our final framework secured first prizes in the AU and EXPR tracks, and second prize in the VA track, underscoring the efficacy of our unified approach and its potential for real-world applications.

2. Related Works

In recent years, there has been significant progress in facial affect analysis. This section presents recent works on

[*] Corresponding Author.

relevant tasks in CVPR2023: ABAW5 competition - AU detection, expression recognition, and VA estimation in the wild. We also discuss state-of-the-art self-supervised learning approaches related to our proposed frameworks.

AU detection. AU detection in the wild faces challenges such as limited identity information and interference from diverse poses, illumination, or occlusions leading to overfitting. Multi-task frameworks, like Zhang et al. [68], Jin et al. [16], and Thinh et al. [51], incorporate auxiliary information as regularization to introduce extra label constraints. Zhang et al. [68, 70] use a pre-trained expression embedding model as the backbone and won ABAW2 and ABAW3. Multi-modal information is also used in ABAW competitions. Zhang et al. [70] fuse vision, acoustic, and text information using a transformer decoder, while Jin et al. [17] use a transformer for multi-modal feature fusion. JAA-Net [48] performs landmarks detection and AU detection simultaneously.

Expression recognition. The goal of expression recognition is to classify an input image into one of the basic emotion classes, such as happiness or sadness. Zhang et al. [68] utilize the prior expression embedding model and propose a multi-task framework. Phan et al. [40] employ the pre-trained model RegNet [42] as the backbone and add the Transformer [53] structure to extract the temporal information. Kim et al. [19] use Swin transformer [31] as the backbone and exploit the extra auxiliary from the audio modal. Wang et al. [58] propose a semi-supervised framework to predict pseudo-labels for unlabeled data, which helps improve the model's generalization to some extent. Xue et al. [62] develop a CFC network that uses different branches to train the easy-distinguished and hard-distinguished emotion categories.

VA estimation. For VA estimation, multi-task frameworks leveraging the correlation between VA and AU or VA and EXPR are proposed in several studies [7,55,59,68]. Multi-modal frameworks are also common, leveraging hidden features from vision, audio, or text [18,37,45,64,66,69]. The Transformer structure is frequently used for feature fusion in VA tasks [30, 37, 69, 73]. These approaches extract supplementary information from other tasks, particularly for data without VA labels but possessing AU or EXPR labels.

Self-supervised learning. Annotating emotion/AU/VA labels from real-world facial images is time-consuming, hindering the development of affective analysis. SSL methods can exploit knowledge from existing large-scale unlabelled data. Shu et al. [49] improve expression recognition accuracy using contrastive SSL methods (e.g., SimCLR [5]). Ma et al. [35] pre-train the MAE structure on large-scale face images and fine-tune it on AU detection and intensity estimation, achieving state-of-the-art performance on BP4D [71] and DISFA [36]. Zhang et al. [69], Liu et al. [30], and Wang et al. [60] employ pre-trained MAE to extract vision features, securing a top few positions in the ABAW5 competition.

Masked language modeling has recently emerged as a highly effective self-supervised learning technique in natural language processing (NLP). Pre-training models, such as BERT [9], are designed to reconstruct masked tokens within a corpus. On the other hand, auto-regressive models, like GPT-3 [3], are pre-trained to predict the next token in the sequence based on all the preceding tokens. These models leverage self-supervised learning on large-scale datasets to improve their performance significantly.

Masked image modeling has been explored as a potential self-supervised learning technique, similar to masked language modeling. However, this approach has not shown much success [10]. Until recently, inspired by the use of tokens in NLP, BEiT [2] proposed a pre-training framework that predicts visual tokens of missing patches using masked image patches as input. As an alternative, MAE [13] proposes a more straightforward method for image modeling, directly reconstructing the pixels of masked patches. MAE is simpler and faster without requiring a tokenizer. As we aim to develop a facial image representation model, we follow MAE-Face [35] as the approach for our pre-training framework.

3. Method

3.1. Pre-training

Our model is a Vision Transformer (ViT) [10] pre-trained using a self-supervised learning approach, which involves the masking-then-reconstruct procedure from Masked Autoencoder (MAE) [13]. To accomplish this, we split an input image into non-overlapping patches of 16×16, and then mask out a portion of these patches, leaving only the visible patches to be fed into the encoder. The goal of training is to reconstruct the masked patches in pixels, using the visible patches as the input. As suggested by [13], we use a masking ratio of 75%, which speeds up the pre-training process since only 25% of the patches need to be processed by the MAE encoder.

Dataset. We pre-train an MAE-Face model [35] instead of training a general vision model based on the ImageNet dataset [44]. To achieve this, we construct a large-scale facial image dataset by concatenating AffectNet [38], CASIA-WebFace [63], IMDB-WIKI [43], and CelebA [32]. The final dataset includes 2,170,000 face images, and all labels are removed for self-supervised learning.

By pre-training on the facial image dataset, our model learns to predict the global states of a face from only 25% observation of the local patches of a face image. This task is non-trivial and beyond the capabilities of traditional image restoration and inpainting algorithms. The latter typically rely on low-level image features to recover missing pixels, while our pre-training task focuses on building a high-level understanding of the relations between different areas of a face image. This potentially forces the model to learn facial features representing identities, expressions, and poses.

Reconstruction loss. To achieve optimal accuracy during the fine-tuning process, instead of the L2 loss originally proposed by MAE, we use L1 loss for the reconstruction:

$$\mathcal{L}_{recon}(\theta) = \frac{1}{C}\sum_{i=1}^{C} |f_\theta(X_{visible}, i) - Norm(x_i)|_1 \quad (1)$$

where θ represents the model parameters, $|\cdot|_1$ denotes the L1 norm. For the visible patches $X_{visible}$ in an input image, $f_\theta(X_{visible}, i)$ is the model's prediction for the ith masked patch x_i, and C is the number of masked patches in the image. To further improve accuracy, a per-patch pixel normalization is applied to the input patch x_i to form the reconstruction target.

3.2. Fine-tuning

During the pre-training stage, our model learns a facial feature representation that we expect can be adapted to any face-related recognition task. In this paper, we focus on the model's ability to be adapted to facial affect analysis tasks, including action unit (AU) detection, expression (EXPR) recognition, and valence-arousal (VA) estimation.

To fine-tune the pre-trained MAE-Face model for each downstream task, we use a supervised learning approach on a labeled dataset. The pre-trained model serves as the initialization weights for the fine-tuning model. Firstly, we take the encoder from the pre-trained model and remove the decoder. Then, we add a global average pooling and a fully-connected layer to the encoder's output, which serves as the classification head for the final predictions. Finally, the model is trained on a dataset specific to the given task under supervision.

The overall framework of the proposed method is illustrated in Fig. 1. The pipeline consists of a two-pass pre-training stage and a two-pass fine-tuning stage. Next, we'll explain them in detail.

3.2.1 Two-pass pre-training

The fine-tuning approach is not restricted to relying solely on the pre-trained MAE-Face model. A two-pass pre-training process can also be employed, which can further enhance the model's performance.

For instance, the MAE-Face model serves as a 1st-pass pre-trained model. We can begin by fine-tuning the 1st-pass pre-trained model on a large dataset using supervised learning, resulting in a 2nd-pass pre-trained model. Subsequently, we can fine-tune the 2nd-pass pre-trained model on a smaller dataset to obtain the final model. This two-pass pre-training process can lead to a model that performs better on the smaller dataset. It is noteworthy that the two datasets used in this process do not need to have the same classification labels and can even be related to different facial affect analysis tasks.

Essentially, the MAE-Face model is a pre-trained model based on self-supervised learning. By utilizing supervised learning in the two-pass pre-training process, we can further enhance the model's capabilities before its final fine-tuning.

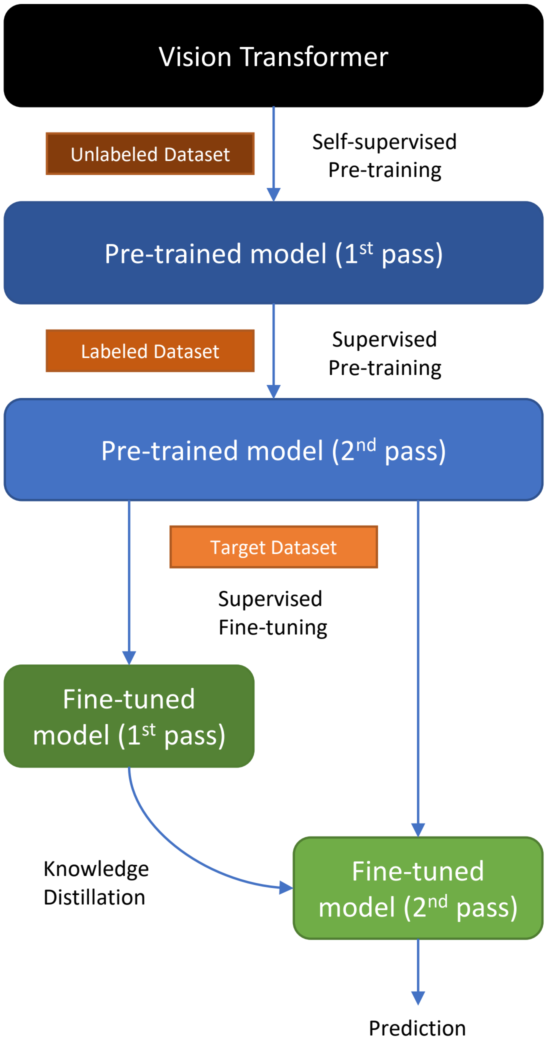

Figure 1. The schematic pipeline of the proposed framework. It consists of two stages: a two-pass pre-training stage and a two-pass fine-tuning stage. During the pre-training stage, facial representations are learned by the model, while the fine-tuning stage enables the model to acquire task-specific prediction capability.

3.2.2 Two-pass fine-tuning

During the fine-tuning process described earlier, we observed that training directly from the hard labels of the original datasets can result in the overfitting of the model. To address this issue, we propose a two-pass fine-tuning process that employs knowledge distillation [1] [14].

Our approach involves training a 1st-pass fine-tuned model using the original labels, which serves as the teacher model. We then train a 2nd-pass fine-tuned model by distilling knowledge from the teacher model. Unlike typical knowledge distillation techniques that distill knowledge from a larger model to a smaller one, we distill knowledge from a model of the same size.

The core idea is that the teacher model provides soft labels as additional training targets, which act as a form of regularization in the optimization process. This regularization allows for more training epochs before the model begins to overfit, resulting in a better model than the one that was fine-tuned using only hard labels from the original dataset.

3.2.3 Loss functions

To achieve the best performance for different tasks, we use different loss functions and hyperparameters. In this section, we describe the loss functions used for the three tasks discussed in this paper.

For AU detection, we use Binary Cross Entropy (BCE) as the loss function. For EXPR classification, we use Cross Entropy (CE) as the loss function. For VA estimation, we use Concordance Correlation Coefficient (CCC) loss as the loss function.

To tackle the class imbalance problem in AU and EXPR, we utilize weighted loss. We rescale the loss of each class by a weight, with larger weights applied to the minority classes and smaller weights to the majority classes.

The loss functions for the three tasks are as follows:

$$\mathcal{L}_{AU} = -\frac{1}{C_{au}} \sum_{j=1}^{C_{au}} W_{au_j} [y_j \log \hat{y}_j + (1 - y_j) \log(1 - \hat{y}_j)]. \tag{2}$$

$$\mathcal{L}_{EXPR} = -\frac{1}{C_{expr}} \sum_{j=1}^{C_{expr}} W_{expr_j} z_j \log \hat{z}_j. \tag{3}$$

$$\begin{aligned} \mathcal{L}_{VA} = 1 - CCC(\hat{v}_{batch_i}, v_{batch_i}) \\ +1 - CCC(\hat{a}_{batch_i}, a_{batch_i}) \end{aligned} \tag{4}$$

$$CCC(\mathcal{X}, \hat{\mathcal{X}}) = \frac{2\rho_{\mathcal{X}\hat{\mathcal{X}}}\delta_{\mathcal{X}}\delta_{\hat{\mathcal{X}}}}{\delta_{\mathcal{X}}^2 + \delta_{\hat{\mathcal{X}}}^2 + (\mu_{\mathcal{X}} - \mu_{\hat{\mathcal{X}}})^2}. \tag{5}$$

where C_{au} and C_{expr} are the number of categories for AU and EXPR tasks, respectively. The weights for different categories, represented by W_{au_j} and W_{expr_j}, are inversely proportional to the number of class samples in the training set. $\hat{y}$, $\hat{z}$, $\hat{v}$, and $\hat{a}$ denote the model's predictions for AU, expression category, Valence, and Arousal, respectively. The symbols without hats refer to the ground truth. $\delta_{\mathcal{X}}$ and $\delta_{\hat{\mathcal{X}}}$ indicate the standard deviations of $\mathcal{X}$ and $\hat{\mathcal{X}}$, respectively. $\mu_{\mathcal{X}}$ and $\mu_{\hat{\mathcal{X}}}$ are the corresponding means and $\rho_{\mathcal{X}\hat{\mathcal{X}}}$ is the correlation coefficient.

3.3. Data pre-processing

During both the pre-training and fine-tuning stages, we use RetinaFace [8] to detect 5-point facial landmarks, using the coordinates of the two eyes for face alignment. We then crop a squared image based on the facial bounding box to ensure consistency.

Furthermore, to ensure the quality of the pre-training dataset, we clean up the dataset by removing corrupt files, non-face images, and images with too low resolution.

3.4. Post-processing

The aforementioned procedure predicts the result on an image basis, which works well for image-based facial affect datasets. However, for video-based datasets, applying the proposed method frame by frame may result in fluctuation and inconsistency in the time domain. To address this issue, a straightforward solution is to apply a smoothing filter over the predicted samples of the model output across the time domain. One popular smoothing filter is the Gaussian filter, defined as:

$$h(\sigma)[n] = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{n^2}{2\sigma^2}} \tag{6}$$

$$g(\sigma)[n] = \sum_{k=-\infty}^{\infty} x[k] h_{\sigma}[n-k] \tag{7}$$

where n is the sample index, $x[n]$ is the input sequence, and $g(\sigma)[n]$ is the filtered output sequence. σ is the standard deviation of the Gaussian filter.

However, directly applying a Gaussian filter on the predicted samples may lead to over-smoothing, eliminating all high-frequency components in the time domain, which is undesirable as facial affect can change rapidly in some scenarios.

To address this issue, we propose a post-processing algorithm that smooths the predictions while compensating for high-frequency components. We define a function, $MD(x[n], y[n])$, that returns the element with the smaller absolute value between $x[n]$ and $y[n]$, as follows:

$$MD(x[n], y[n]) = \begin{cases} x[n] & \text{if } |x[n]| \leq |y[n]| \\ y[n] & \text{if } |x[n]| > |y[n]| \end{cases} \quad (8)$$
$$\text{for } n = 0, 1, 2, \ldots$$

With a input sequence $x[n]$, we define the filtered output sequence, $y[n]$, as follows:

$$\begin{aligned} y[n] &= g(\sigma_1)[n] \\ &\quad + MD(x[n] - g(\sigma_2)[n], x[n] - g(\sigma_3)[n]) \end{aligned} \quad (9)$$

where σ_1, σ_2, and σ_3 are the standard deviation of the Gaussian filters. We set the limitation that $\sigma_1 > \sigma_2$ and $\sigma_1 > \sigma_3$.

The filtered output sequence, $y[n]$, is obtained by adding the Gaussian filtered output sequence, $g(\sigma_1)[n]$, to the output of the MD function applied to the difference between the input sequence, $x[n]$, and two Gaussian filtered sequences with different standard deviations, σ_2 and σ_3, respectively. The MD function ensures that the high-frequency components that were eliminated by the Gaussian filter are compensated for, while still smoothing the overall output.

Our proposed post-processing algorithm can be applied to any facial affect recognition model that operates on video-based datasets and can help to improve the stability and consistency of the output results in the time domain.

4. Experiments

4.1. Experimental Setting

Pre-training. Our model is pre-trained for 800 epochs with 40 warmup epochs using the AdamW optimizer [34] and a weight decay of 0.05. Random cropping is applied for data augmentation, and the Transformer blocks are initialized with Xavier Uniform [12]. The batch size is set to 4096, and the learning rate is 2.4e-3 with cosine annealing [33].

Fine-tuning. For each task, our model is fine-tuned for 50 epochs with 5 warmup epochs using the AdamW optimizer and a weight decay of 0.05. RandAug(9, 0.5) [6] is used for data augmentation. Drop path [15] of 0.1 is applied for regularization. The batch size is set to 512, and the learning rate is 2e-4 with cosine annealing.

All implementations are created using PyTorch [39] and trained on 8 NVIDIA A30 GPUs.

4.2. Evaluation metrics

For AU detection, we measure the performance using F1-score, where the F1-score is calculated by taking the average of the F1-score on each class.

For EXPR classification, we measure the performance using F1-score or Top-1 accuracy. Furthermore, we also examine Top-5 accuracy and Top-10 accuracy for the Emo135 dataset.

For VA estimation, we measure the performance by calculating Concordance Correlation Coefficient (CCC) for valence and arousal respectively.

The definitions for each track in the ABAW5 challenge are described as follows:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

$$S_{AU} = \frac{1}{N_{au}} \sum F1_{au_i} \quad (11)$$

$$S_{EXPR} = \frac{1}{N_{expr}} \sum F1_{expr_i} \quad (12)$$

$$S_{VA} = 0.5 * (CCC(\hat{v}, v) + CCC(\hat{a}, a)) \quad (13)$$

where the definition of CCC refers to equ. 5

4.3. Results on AffectNet

AffectNet [38] is a widely used dataset for facial expression classification, comprising approximately 440k face images with manual annotations of seven discrete facial expressions and the intensity of valence and arousal.

We fine-tune our model on AffectNet for both 8-class and 7-class tracks, with results shown in Tab. 1. We utilize one-pass pre-training and two-pass fine-tuning. Our approach achieves 66.65% Top-1 accuracy in 8-class and 69.51% Top-1 accuracy in 7-class, surpassing previous works by more than 3%.

In addition to the ViT-Base [10] backbone, we incorporate a ViT-Tiny [52] variant to evaluate performance in a smaller model size. The ViT-Tiny model demonstrates reasonably good performance, with about 1% lower accuracy than the ViT-Base model, making it more suitable for deployment in applications where speed is a concern.

4.4. Results on Emo135

Emo135 [4] is a dataset comprising 135 emotion categories with a total of 696,168 facial images. Each emotion category contains between 994 and 12,794 facial images labeled with emotions. Due to the higher number of class labels, it presents a more challenging expression classification task compared to AffectNet.

We fine-tune our model on Emo135 and present the results in Tab. 2. We utilize one-pass pre-training and two-pass fine-tuning. Our approach outperforms the baseline results reported by Chen et al. [4] by a significant margin.

Additionally, we also test different ViT model variants as the backbone. The results reveal that ViT-Base [10] significantly outperforms ViT-Tiny [52], while ViT-Large [10]

Method	Metric: Top-1 Acc. (%)	
	AffectNet-8	AffectNet-7
ESR-9 [50]	59.3	-
RAN [57]	59.5	-
Georgescu et al. [11]	59.58	63.31
VGG-FACE [21]	60.40	-
PSR [54]	60.68	-
Distilled student [47]	61.60	65.40
Pourmirzaei et al. [41]	61.72	-
DAN [61]	62.09	65.69
MT-ArcRes [27]	63	-
Savchenko et al. [46]	63.03	66.29
Ours (ViT-Tiny)	65.23	68.46
Ours (ViT-Base)	**66.65**	**69.51**

Table 1. AffectNet (EXPR): The results of the models that are trained and evaluated on the 8-class or 7-class task. We also include the results from several state-of-the-art works for comparison. The best results are in bold.

Method	Metric			
	F1-score	Top-1 Acc. (%)	Top-5 Acc. (%)	Top-10 Acc. (%)
Chen et al. [4]	0.247	28.3	66.4	78.7
Ours (ViT-Tiny)	0.3203	35.44	75.65	86.61
Ours (ViT-Base)	0.3753	38.64	80.82	89.71
Ours (ViT-Large)	**0.3791**	**39.26**	**80.93**	**89.86**

Table 2. Emo135 (EXPR): The results of the models that are trained on the training set and evaluated on the test set. We also compare our result to the baseline from Chen et al. [4]. The best results are in bold.

Val set	Metric
	F1-score
Official	0.5553
5-fold Avg.	0.5548
fold-1	0.5617
fold-2	0.5843
fold-3	0.5606
fold-4	0.5274
fold-5	0.5402

Table 3. ABAW5 (AU): The results of our models that are trained and evaluated on different folds.

Val set	Metric	
	F1-score	Top-1 Acc. (%)
Official	0.4460	58.51
5-fold Avg.	0.4403	56.45
fold-1	0.4306	52.95
fold-2	0.4642	59.41
fold-3	0.4299	57.74
fold-4	0.5028	59.44
fold-5	0.3738	52.72

Table 4. ABAW5 (EXPR): The results of our models that are trained and evaluated on different folds.

offers only a marginal advantage over ViT-Base. Therefore, for Emo135, ViT-Base represents the optimal balance between performance and speed.

4.5. Results on ABAW5 validation set

The Affective Behavior Analysis in-the-wild (ABAW5) competition provides the Aff-Wild2 dataset, which is a large-scale video-based dataset comprising three emotion representations: action units, expression categories, and valence-arousal.

We assess the effectiveness of our method on all three tracks. It is important to note that a separate model is fine-tuned on the specific data for each track. All models employ ViT-Base as the backbone. We utilize two-pass pre-training and two-pass fine-tuning, where the 2nd-pass pre-training is based on the AffectNet dataset. Apart from the official validation split, we also evaluate the performance using 5-fold cross-validation. The results are presented in Tab. 3, Tab. 4, and Tab. 5.

The results obtained from the AU and EXPR tracks

Val set	Metric: CCC		
	Valence	Arousal	VA Avg.
Official	0.4652	0.6177	0.5414
5-fold Avg.	0.5914	0.5507	0.6321
fold-1	0.5811	0.5500	0.6122
fold-2	0.5691	0.5382	0.6001
fold-3	0.6175	0.5618	0.6731
fold-4	0.5598	0.5093	0.6103
fold-5	0.6293	0.5941	0.6646

Table 5. ABAW5 (VA): The results of our models that are trained and evaluated on different folds.

demonstrate that the proposed method has the ability to generalize well on different validation splits. This is evident from the fact that the results obtained from the official validation set are quite similar to those obtained from the 5-fold average. Therefore, we can conclude that the proposed method is robust and can perform well on different validation sets. However, it is worth mentioning that the results obtained from the VA track indicate a relatively large gap, which suggests that the VA track is more sensitive to validation splits. In the future, more experiments should be conducted to further investigate the sensitivity of the VA track to validation splits and to determine whether the proposed method can be further improved to perform better on different validation sets.

4.6. Results on ABAW5 test set

In the ABAW5 competition, the proposed model in this paper serves as a strong baseline for our final method. Our final approach leverages multi-modal information from vision, acoustic, and text modalities, as well as various post-processing techniques, to improve the capability and robustness in analyzing the Aff-Wild2 dataset.

Our method won first place in the AU Detection Challenge with an F1-score of 0.5549 (Tab. 6). For the Expression Classification Challenge, our method also won first place with an F1-score of 0.4121 (Tab. 7). Furthermore, we secured second place in the VA Estimation Challenge with a CCC score of 0.6372 (Tab. 8).

Other top-performing methods such as SituTech and SZFaceU also use MAE pre-trained on facial image datasets as the feature extractor. SituTech and CtyunAI also incorporate multi-modal information for expression classification, combining vision and audio features. HFUT-MAC uses POSTER2 as the feature extractor and a transformer for temporal feature integration. HSE-NN-SberAI applies EfficientNet and MLP for classification. For the VA Estimation Challenge, CBCR leverages TCN for temporal feature capture and channel attention network (CAN) for feature fusion, while other teams' methods are similar to those in the other challenges.

Team	Test Set	
	Rank	F1-score
PRL [56]	#5	0.5101
SZFaceU [60]	#4	0.5128
USTC-IAT-United [64]	#3	0.5144
SituTech [30]	#2	0.5422
Ours	#1	**0.5549**

Table 6. Final competition results (average F1-score) on the AU test set of ABAW5.

Team	Test Set	
	Rank	F1-score
HSE-NN-SberAI [45]	#5	0.3292
HFUT-MAC [72]	#4	0.3337
CtyunAI [73]	#3	0.3532
SituTech [30]	#2	0.4072
Ours	#1	**0.4121**

Table 7. Final competition results (average F1-score) on the EXPR test set of ABAW5.

Team	Test Set	
	Rank	CCC
HFUT-MAC [72]	#5	0.5342
CtyunAI [73]	#4	0.5666
CBCR [67]	#3	0.5913
SituTech [30]	#1	**0.6414**
Ours	#2	0.6372

Table 8. Final competition results (average CCC) on the VA test set of ABAW5.

In summary, our proposed method outperforms other top-performing methods in the Action Unit Detection and Expression Classification Challenges, while achieving competitive results in the VA Estimation Challenge. Our approach leverages the MAE-Face pre-training and multi-modal information to improve performance. Other top teams also employ similar techniques, highlighting the effectiveness of such methods.

5. Conclusion

This paper presents MAE-Face as a unified approach to facial affect analysis. By pre-training on a large-scale facial image dataset, MAE-Face learns a robust visual representation for facial affect. The pre-trained model is subsequently fine-tuned on different datasets for specific downstream tasks.

Our proposed method demonstrates state-of-the-art performance on various facial affect analysis task datasets. The results for different tasks stem from the same pre-trained model and fine-tuning procedure, showcasing the robustness of the learned facial feature representation and its adaptability to distinct facial affect analysis tasks.

In the ABAW5 competition, the model proposed in this paper serves as the fundamental building block in our final framework, offering a robust baseline for visually analyzing facial affective behavior on a frame-by-frame basis. By incorporating sequential information and multimodal analysis in our final framework, we won first prizes in the AU and EXPR tracks, and second prize in the VA track.

Acknowledgments

The experiments and the data management and storage are supported by NetEase Fuxi Youling platform, based on Fuxi Agent-Oriented Programming (AOP) system that is carefully designed to facilitate task modeling. This work is also supported by the 2022 Hangzhou Key Science and Technology Innovation Program (No. 2022AIZD0054), and the Key Research and Development Program of Zhejiang Province (No. 2022C01011).

References

[1] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014. 4

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[4] Keyu Chen, Xu Yang, Changjie Fan, Wei Zhang, and Yu Ding. Semantic-rich facial emotional expression recognition. *IEEE Transactions on Affective Computing*, 13(4):1906–1916, 2022. 5, 6

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 5

[7] Didan Deng, Liang Wu, and Bertram E. Shi. Iterative distillation for better uncertainty estimates in multitask emotion recognition, 2021. 2

[8] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 4

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5

[11] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Marius Popescu. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, 7:64827–64836, 2019. 6

[12] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 5

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1, 2

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[15] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016. 5

[16] Yue Jin, Tianqing Zheng, Chao Gao, and Guoqiang Xu. A multi-modal and multi-task learning method for action unit and expression recognition. *arXiv preprint arXiv:2107.04187*, 2021. 2

[17] Yue Jin, Tianqing Zheng, Chao Gao, and Guoqiang Xu. A multi-modal and multi-task learning method for action unit and expression recognition. *arXiv preprint arXiv:2107.04187*, 2021. 2

[18] Vincent Karas, Mani Kumar Tellamekala, Adria Mallol-Ragolta, Michel Valstar, and Björn W. Schuller. Continuous-time audiovisual fusion with recurrence vs. attention for in-the-wild affect recognition. 2022. 2

[19] Jun-Hwa Kim, Namho Kim, and Chee Sun Won. Facial expression recognition with swin transformer, 2022. 2

[20] 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. 1

[21] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 128:1455–1484, 2020. 6

[22] 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. 1

[23] 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. 1

[24] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1

[25] 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. *arXiv preprint arXiv:2303.01498*, 2023. 1

[26] 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. 1

[27] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1, 6

[28] 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. 1

[29] 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. 1

[30] Chuanhe Liu, Xinjie Zhang, Xiaolong Liu, Tenggan Zhang, Liyu Meng, Yuchen Liu, Yuanyuan Deng, and Wenqiang Jiang. Multi-modal expression recognition with ensemble method, 2023. 2, 7

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2

[33] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[35] 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. 1, 2

[36] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 2

[37] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Yuan Cheng, Meng Wang, Chuanhe Liu, and Qin Jin. Multimodal emotion estimation for in-the-wild videos, 2022. 2

[38] 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. 2, 5

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[40] Kim Ngan Phan, Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. Expression classification using concatenation of deep neural network for the 3rd abaw3 competition, 2022. 2

[41] Mahdi Pourmirzaei, Gholam Ali Montazer, and Farzaneh Esmaili. Using self-supervised auxiliary tasks to improve fine-grained facial representation. *arXiv preprint arXiv:2105.06421*, 2021. 6

[42] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. 2

[43] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018. 2

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2

[45] Andrey V. Savchenko. Emotieffnet facial features in uni-task emotion recognition in video at abaw-5 competition, 2023. 2, 7

[46] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143, 2022. 6

[47] Liam Schoneveld, Alice Othmani, and Hazem Abdelkawy. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, 146:1–7, 2021. 6

[48] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaanet: joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129(2):321–340, 2021. 2

[49] Yuxuan Shu, Xiao Gu, Guang-Zhong Yang, and Benny Lo. Revisiting self-supervised contrastive learning for facial expression recognition. *arXiv preprint arXiv:2210.03853*, 2022. 2

[50] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient facial feature learning with wide ensemble-based con-

volutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5800–5809, 2020. 6

[51] Phan Tran Dac Thinh, Hoang Manh Hung, Hyung-Jeong Yang, Soo-Hyung Kim, and Guee-Sang Lee. Emotion recognition with incomplete labels using modified multi-task learning technique. *arXiv preprint arXiv:2107.04192*, 2021. 2

[52] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 5

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[54] Thanh-Hung Vo, Guee-Sang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 8:131988–132001, 2020. 6

[55] Manh Tu Vu and Marie Beurton-Aimar. Multitask multi-database emotion recognition, 2021. 2

[56] Tu Vu, Van Thong Huynh, and Soo Hyung Kim. Vision transformer for action units detection, 2023. 7

[57] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 6

[58] Lingfeng Wang and Shisen Wang. A multi-task mean teacher for semi-supervised facial affective behavior analysis. *arXiv preprint arXiv:2107.04225*, 2021. 2

[59] Lingfeng Wang, Shisen Wang, Jin Qi, and Kenji Suzuki. A multi-task mean teacher for semi-supervised facial affective behavior analysis, 2021. 2

[60] Zihan Wang, Siyang Song, Cheng Luo, Yuzhi Zhou, shiling Wu, Weicheng Xie, and Linlin Shen. Spatio-temporal au relational graph representation learning for facial action units detection, 2023. 2, 7

[61] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270*, 2021. 6

[62] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. 2

[63] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2

[64] Jun Yu, Renda Li, Zhongpeng Cai, Gongpeng Zhao, Guochen Xie, Jichao Zhu, and Wangyuan Zhu. Local region perception and relationship learning combined with feature fusion for facial action unit detection, 2023. 2, 7

[65] 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. 1

[66] Su Zhang, Ruyi An, Yi Ding, and Cuntai Guan. Continuous emotion recognition using visual-audio-linguistic information: A technical report for abaw3, 2022. 2

[67] Su Zhang, Ziyuan Zhao, and Cuntai Guan. Multimodal continuous emotion recognition: A technical report for abaw5, 2023. 7

[68] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, Yu Ding, Runze Wu, Tangjie Lv, and Changjie Fan. Prior aided streaming network for multi-task affective analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3539–3549, 2021. 2

[69] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Facial affective analysis based on mae and multi-modal information for 5th abaw competition, 2023. 2

[70] Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2428–2437, 2022. 2

[71] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 2

[72] Ziyang Zhang, Liuwei An, Zishun Cui, Ao xu, Tengteng Dong, Yueqi Jiang, Jingyi Shi, Xin Liu, Xiao Sun, and Meng Wang. Facial affect recognition based on transformer encoder and audiovisual fusion for the abaw5 challenge, 2023. 7

[73] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on tcn and transformer, 2023. 2, 7