Spatial-Temporal Graph-Based AU Relationship Learning for Facial Action Unit Detection

Zihan Wang[1,2], Siyang Song[3], Cheng Luo[1,2], Yuzhi Zhou[1,2], Shiling Wu[1,2],
Weicheng Xie[1,2] and Linlin Shen[1,2] *
[1]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University
[2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
[3]School of Computing and Mathematical Sciences, University of Leicester
{2018112013, luocheng2020, 2200271004, 2110276070}@email.szu.edu.cn,
{wcxie, llshen}@szu.edu.cn, ss1535@leicester.ac.uk

Abstract

This paper presents our Facial Action Units (AUs) detection submission to the fifth Affective Behavior Analysis in-the-wild Competition (ABAW). Our approach consists of three main modules: (i) a pre-trained facial representation encoder which produce a strong facial representation from each input face image in the input sequence; (ii) an AU-specific feature generator that specifically learns a set of AU features from each facial representation; and (iii) a spatio-temporal graph learning module that constructs a spatio-temporal graph representation. This graph representation describes AUs contained in all frames and predicts the occurrence of each AU based on both the modeled spatial information within the corresponding face and the learned temporal dynamics among frames. The experimental results show that our approach outperformed the baseline and the spatio-temporal graph representation learning allows our model to generate the best results among all ablated systems. Our model ranks at the 4th place in the AU recognition track at the 5th ABAW Competition. Our code is publicly available at https://github.com/wzh125/ABAW-5.

1. Introduction

Human Facial Action Units (AUs), as a widely-used description for facial muscle movements, play a significant role in human behavior understanding [11,22,42,44]. Facial AUs are annotated according to the anatomical characteristics of multiple facial muscle movement based on Facial Action Coding System (FACS) [8]. Compared to categorical facial expressions, AUs are more objective and comprehensive representation of facial expressions, and thus drew increasing attentions in recent years [33]. However, AU detection is a challenging multi-label classification task as AUs are subtle movement of facial muscles, and different facial muscles have different ranges of movement, which are affected by various person-specific factors (e.g., gender and age) as well as contexts (e.g., background, illumination ,and occlusion).

The Action Unit Detection Challenge of the 5th ABAW Competition [20] is based on the Aff-Wild2 [16–19, 21–24, 55] database. Some of the AU detection approaches in the previous ABAW Competitions [16, 17, 24] fuse multi-modal features including video and audio to provide multi-dimensional information to predict AUs' occurrence [13, 14, 50, 58]. Meanwhile, other studies found that AU detection performance can be benefited from multi-task learning [3, 13, 36, 56], i.e., jointly conducting expression recognition or valence/arousal estimation provides helpful cues for AU detection. Moreover, temporal models such as GRU [6] or Transformer [48] are also introduced to model temporal dynamics among consecutive frames [37, 50]. While AUs' activation status in each facial display are highly correlated, their relationships provide crucial cues for their occurrence recognition. Meanwhile, the annotations of AUs in the Aff-Wild2 database exhibit a notable imbalance (e.g.,samples of AU7,10,25 are far more than that of AU15,23,24,26 and some AUs only appear on certain identities.), which can result in the training of a biased model that are predisposed to learn AU patterns that have been annotated more frequently in the training set. However, to the best of our knowledge, there is no previous study can jointly address both problems.

Recent studies show that the graph representation is powerful for modelling the underlying relationship among AUs [31, 45, 46]. In particular, task-specific multi-dimensional edge features shows strong capability in ex-

*Corresponding Author

plicitly describing the relationship between each pair of AUs. To overcome the overfitting problem, Ma et al. [32]introduces a robust facial representation model MAE-Face for AU analysis. But both of them ignore the temporal information. Considering that the Aff-Wild2 database consists of videos, there is a certain relationship between different frames in a video and the adjacent frames are relatively similar. Therefore, AU detection can be benefited from temporal information.

In this paper, we propose a spatio-temporal facial AU graph representation learning framework for the AU Detection Challenge at the 5th ABAW Competition. Our framework starts with pre-training a masked autoencoder (MAE) [9] from a set of face databases. This way, the pre-trained MAE can produce a strong facial representation from each input face image. Based on the facial representation, a AU-specific Feature Generator learns specific representation for each AU, which is considered as the node feature in the spatio-temporal AU graph. Then, a spatio-temporal graph learning (STGL) module is introduced to jointly model the spatio-temporal relationships among AUs of all face frames. Specifically, the update of a AU node is not only related to its spatial neighbours in the same frame, but also the nodes in its specific AU sequence, so that the relationships between different AUs and the temporal information of a specific AU sequence can interact and jointly guide the graph to learn representation for each AU node. The main contributions of this work are listed as follow:

- We pre-train a MAE model based on human face databases, which can generate a strong facial representation from each input facial display, to overcome the data imbalance problem in action units detection.
- We propose a spatio-temporal graph learning module to model spacial relationships between different AUs and temporal dependencies among different frames.
- The proposed method achieves significant improvement over the baseline and ranks at the 4th place in the AU Detection Challenge at the 5th ABAW Competition.

2. Related Work

In this section, we systematically review previous AU detection approaches, which are categorized into two types: Non-graph and graph-based AU detection approaches.

2.1. Non-graph based AU detection approaches

Since each AU's activation can only appear in a sparse facial region [15], several facial region-based methods are proposed [10, 15, 26, 27, 40]. For example, Li et al. [27] first crop key facial regions from face images, and then learn deep features for each facial region individually. Jacob et al. [10] set an attention module to enforce the model to focus on the facial regions corresponding to activated AUs. There are several studies [2, 12, 37, 41, 43] introduce temporal information as the facial muscle movement is a dynamic process. Chu et al. [2] use CNN to extract feature from each frame and model the temporal sequence by LSTM. Nguyen et al. [37] utilize Transformer to add temporal information in frame sequence and they ranked 3rd in the AU detection challenge at the ABAW competition 2022. The methods summarised above are built on typical supervised learning, whose generalization capabilities are largely depending on the quality of AU annotations. Subsequently, self-supervised learning strategies [1, 28, 28, 32] recently have been frequently introduced to AU recognition. In particular, MAE-Face [32] first learns a high-capacity model from a large amount of face images without any data annotations, then after being fine-tuned on downstream task including AU detection and AU intensity estimation, which exhibits convincing performance.

2.2. Graph-based AU recognition approaches

Considering that relationships between AUs (i.e., AU co-occurrence pattern) play a significant role in AU recognition, some researchers utilize graph neural networks (GNNs) to model the underlying relationship. Li et al. [25] is the first attempt that employs the GNN for AU relationships modeling. Song et al. [46] propose an uncertain graph neural network to capture the importance of the dependencies among AUs for each input and estimate the prediction uncertainties. More recently, Luo et al. [31, 45] propose to learn multi-dimensional edge feature-based AU relational graph, where the relationship between each pair of AUs can be explicitly modelled by a task-specific multi-dimensional edge feature. Song et al. [47] construct a co-occurrence knowledge graph and a spatio-temporal Transformer module to capture the temporal and spatial relations of AUs. Nguyen et al. [36] use a facial graph to capture the association among action units for the multi-task learning challenge and they ranked 4th in multi-task challenge at the ABAW competition 2022. These works exhibit the effectiveness of modeling AU relationships.

3. Methodology

Given T consecutive facial frames $S = \{f^1, ..., f^t, ..., f^T\}$, our goal is to predict AUs' occurrences for each frame. Since there are multiple AUs defined for each facial display, our approach aims to jointly predict multiple AUs for all frame, which is denoted as $P_t = \{p_1, p_2, ..., p_N\}$, where N represents the number of predicted AUs, t denotes the t_{th} frame and $p \in \{0, 1\}$ can be either activated (1) or inactivated (0). The pipeline of our approach is illustrated in Fig. 1, which consists of a Facial Representation Encoder (FRE) described in Sec. 3.1,

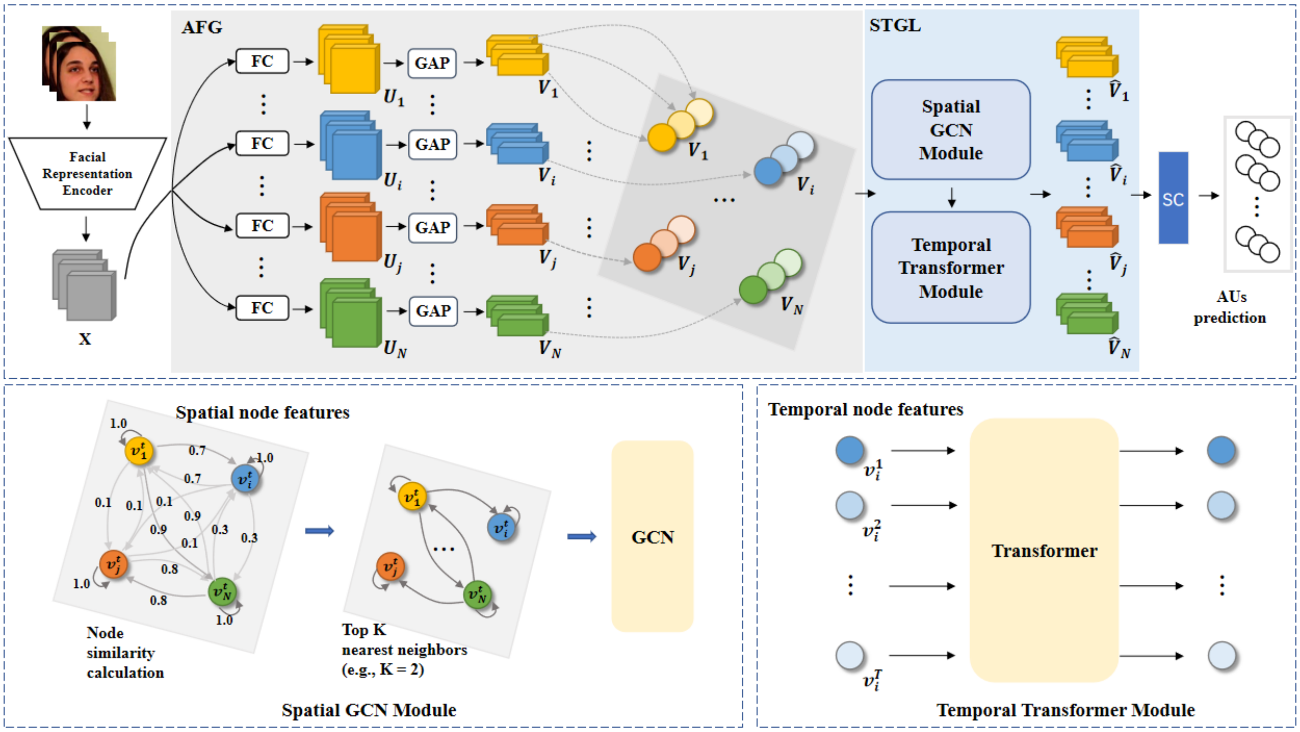

Figure 1. The pipeline of the proposed Spatio-Temporal AU Relational Graph Representation Learning approach

an AU-specific Feature Generator (AFG) described in Sec. 3.2 and a Spatio-Temporal Graph Learning (STGL) module described in Sec. 3.3. Specifically, the FRE first extracts global facial features from each image of the input face sequence. Then, the AFG individually learns a representation for each AU of each facial frame based on its global representation. This way, a spatio-temporal facial graph representation can be constructed. Finally, the STGL module takes the spatio-temporal facial graph as the input and jointly predicts all AUs' occurrence for all frames, where both spatial facial displays and temporal facial dynamics are considered.

3.1. Facial Representation Encoder

Masked autoencoder (MAE) [9] is a self-supervised learned model which reconstructs original images from a set of masked images. It is made up of a linear projection layer, a 12-layer encoder and a 4-layer decoder that were defined by the Vision Transformer [7]. The well-trained MAE can be fine-tuned for various downstream tasks.

Since MAE has a strong representation learning capability and scalability, we propose to first pre-train a MAE model using a large amount of face images from CASIA-WebFace [52], AffectNet [34], IMDB-WIKI [38] and CelebA [30], making the pre-trained MAE to be able to generate strong facial representations from previous unseen face images. This pre-training strategy would not only help the model to alleviate the data imbalance problem in target AU database, but also improve the generalization ability of network in uncontrolled environments. Fig. 2 illustrates the MAE model's pre-training, where randomly masked face images are fed to the encoder to generate latent features, and then the decoder reconstructs the original image from these latent features.

Subsequently, the linear projection layer and the encoder of the pre-trained MAE can generate a strong representation for each input face image, which are employed as the encoder for our facial AU recognition pipeline. To generate AU predictions for the input facial image sequence $S = \{f^1, ..., f^t, ..., f^T\} \in \mathbb{R}^{T \times C \times H \times W}$, the linear projection layer first encodes each frame to a set of patches, which are treated as a set of tokens to be fed into the pre-trained transformer encoder without masking operation. As a result, a set of facial representations $X = \{x^1, ..., x^t, ..., x^T\} \in \mathbb{R}^{T \times m \times d}$ can be generated, where each $x_t \in X$ represents a global facial representation of a face image; m is the number of patches and d denotes the dimension of each patch.

3.2. AU-specific Feature Generator

Since each AU's activation only appears in a specific local facial region but will be reflected bu other facial regions, we propose a AU-specific Feature Generator(AFG) to extract unique feature for each AU from the global facial representation X. In particular, the AFG consists of N

branches, where each is made up of a fully connected layers (FC) followed by a global average pooling (GAP) layer. The i_{th} FC layer of first projects the X to an AU-specific feature map $U_i \in \mathbb{R}^{T\times m\times d}$, and then GAP layer yields T vectors consisting $V_i = \{v_i^1, ..., v_i^t, ..., v_i^T\} \in \mathbb{R}^{T\times d}$, where v_i^t denotes the representation of the i_{th} AU in the t_{th} frame.

3.3. Spatial-Temporal Graph Learning

As discussed before, AUs in each facial display are related to each other. Meanwhile, since human facial behaviours are continuous and smooth, AU activation status in adjacent frames are also temporally correlated. In this sense, our method jointly learns both the spatial relationship among AUs within each face frame as well as the their temporal relationship among face frames. Specifically, the Spatial-Temporal Graph Learning (STGL) module consists of a spacial GCN module for spatial AU relationship modelling and a temporal transformer module for temporal AU relationship modelling.

3.3.1 Spacial GCN module

Firstly, we employ the Facial Graph Generator (FGG) proposed by [31, 45] to learn a spatial AU graph representation for each face frame, which consists of N nodes describing features of the N target AUs. Then, the connectivity (edge presence) between each pair of nodes is defined according to the similarity of their features, i.e., each node connects with its K nearest neighbour nodes with highest similarity scores. This way, the topology of the generated graph representation would have adapted topology for different facial displays. After that, a GCN layer is adopted to update node features for the obtained facial graph representation. The new representation of the i_{th} AU in the t_{th} frame can be calculated by its neighbours in the spatial dimension as:

$$v_i^t = \sigma[v_i^t + g(v_i^t, \sum_{j=1}^{N} r(v_j^t, e_{i,j}^t))] \tag{1}$$

where σ is the activation function; g and r denotes differentiable functions of the GCN layer, and $e_{i,j}^t \in \{0, 1\}$denotes the connectivity between v_i^t and v_j^t. Specifically, the above operations will be conducting in each frame of the input sequence.

3.3.2 Temporal transformer module

Since transformer [48] is a superior model to learn long and short-range temporal dependencies, we then propose to utilize the transformer to update AU representations by considering temporal dynamics among facial frames. In temporal dimension, the graph nodes output from the GCN layer are considered as N sequences for N AUs, each

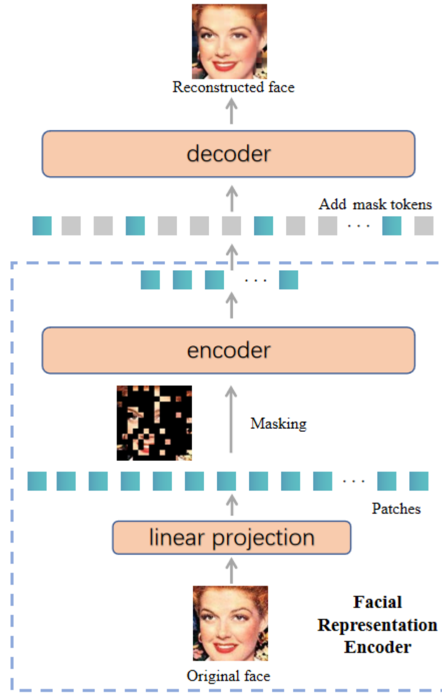

Figure 2. Illustration of the masked autoencoder model

of which consists of T nodes. For the i_{th} AU sequence $V_i = \{v_i^1, \cdots, v_i^t, \cdots, v_i^T\} \in \mathbb{R}^{T\times d}$, the T nodes will be taken as T tokens which are then fed to the transformer. In particular, each AU node sequence V_i is then individually updated as $\hat{V}_i$ as follows:

$$\begin{aligned} \hat{V}_i &= Z + \text{FFN}(\text{LayerNorm}(Z)) \\ Z &= V_i + \text{Att}(Q, K, V) \\ Q &= V_i W_Q, \;\; K = V_i W_K, \;\; V = V_i W_V \end{aligned} \tag{2}$$

where FFN is the feed forward network in transformer; Att denotes the self-attention function; and W_Q, W_K and W_V are trainable weight matrices. We perform these operations on each AU sequence in the temporal dimension.

In this paper, three Spatio-Temporal Graph Learning (STGL) modules are stacked to produce spatio-temporal AU graph representations. Then, the similarity calculating (SC) strategy [31] is employed to predict the probability. For the i_{th} AU in the t_{th} frame, a trainable vector s_i which has the same dimension as $\hat{v}_i^t$ is shared across all frames, and the prediction can be denoted as:

$$P_i^t = \frac{\sigma(\hat{v}_i^t)^T \sigma(s_i)}{||\sigma(\hat{v}_i^t)||_2 ||\sigma(s_i)||_2} \tag{3}$$

where σ is the activation function.

	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU15	AU23	AU24	AU25	AU26	Average
Baseline [20]	-	-	-	-	-	-	-	-	-	-	-	-	39.0
ME-Graph [31]	56.1	44.9	52.1	61.7	74.8	**75.4**	73.4	28.4	18.4	12.4	84.5	32.2	51.2
Netease [57]	55.3	**51.4**	**56.7**	**67.3**	**75.8**	75.1	**75.8**	31.2	17.3	**33.8**	83.9	**42.3**	**55.9**
Ours	**57.8**	48.0	55.9	61.9	75.5	74.6	72.0	**35.6**	**21.6**	23.7	**86.0**	38.3	54.3

Table 1. F1 score (in %) results achieved for 12 AUs on validation set.The highest scores are indicated in bold.

	Swin-B (ImageNet)	MAE (ImageNet)	MAE (hybrid dataset)	Spatial	Temporal	Average F1-score
(i)	✓					49.4
(ii)	✓			✓		50.7
(iii)	✓			✓	✓	52.6
(iv)		✓				48.2
(v)		✓		✓		49.5
(vi)		✓		✓	✓	50.0
(vii)			✓			52.4
(viii)			✓	✓		52.7
(ix)			✓	✓	✓	54.3

Table 2. Ablation study on validation set.

3.4. Loss Function

We propose a two-stage training strategy to train our AU detection model. At the first stage, we first pre-train the transformer-based facial representation encoder by reconstructing the masked patches of the original face images. Here, we employ Mean Square Error (MSE) loss to constrain the difference between the reconstructed patches and the original patches at the pixel-level. Suppose that M patches are masked at the beginning, the pre-training loss L_{pre} is formulated as:

$$L_{\text{pre}} = \sum_{m=1}^{M} (\hat{p}_m - p_m)^2 \tag{4}$$

where $\hat{p}_m$ denotes the ground truth pixels and p_m denotes the reconstructed pixels.

Since AU detection is a multi-label binary classification problem, and most AUs are inactivated for the marjority of face frames, we use an asymmetric loss to optimize the network at the second training stage, which enforce the whole framework to jointly output multiple AUs occurrence prediction. The L_{au} is denoted as:

$$L_{au} = -\sum_{i=1}^{N}\sum_{t=1}^{T} [y_i^t \log(p_i^t) + p_i^t(1-y_i^t)\log(1-p_i^t)] \tag{5}$$

where p_i^t, y_i^t are the prediction and ground truth; N and T are the numbers of AUs and frames of the input face sequence, respectively. The first p_i^t in $p_i^t(1 - y_i^t)\log(1 - p_i^t)$ can be considered as the weight of negative samples(inactivated AUs), which down the loss values caused by inactivated AUs that are easy to detect,enforcing the training process to focus on activated AUs and inactivated AUs that are hard to be correctly recognized.

4. Experimental results

4.1. Dataset

Dataset for MAE: Self-supervised pre-training helps neural networks learn effectively discriminative representations, however, it will bring limited gains for downstream tasks if the size of training data is limited. For this reason, we collect a hybrid face dataset from different in-the-wild datasets, following the collecting method in the prior work [32]. This hybrid dataset involves four subsets from CASIA-WebFace [52], AffectNet [34], IMDB-WIKI [38] and CelebA [30], respectively. These source datasets are well-known and widely used in fields spanning from face recognition to expression recognition. However, we found some low-quality image data included in these datasets. In order to make the training more effective, we remove all images with blur and incomplete faces. Finally, we can obtain the hybrid dataset of around 1,920,000 face images without annotations for our pre-training.

Dataset for AU detection: The AU Detection Challenge at 5th ABAW Competition [20] provides 541 video sequences from Aff-Wild2 dataset. Each frame of a video sequence in this dataset is manually or automatically annotated with labels of 12 AUs, namely AU1, AU2, AU4,

AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, and AU26. Totally, this dataset contains 2,627,632 frames, with 438 subjects, 268 of which are males and 170 are females. Meanwhile, this video dataset is split into a training set of 295 sequences, a validation set of 105 sequences, and a testing set of 141 sequences in a subject-independent manner.

4.2. Experimental settings

Details for MAE pre-training: We first leverage RetinaFace [5] to perform face detection and alignment for each image from the hybrid dataset and crop it to 256×256. The encoder and decoder are initialized with the weights pre-trained on ImageNet-1k [4] dataset. When reconstructing each masked image, we applied random cropping augmentation and chose a mask ratio of 75%. During training, a AdamW optimizer with the learning rate of $1.5e^{-4}$ is used, with batch size of 512, and weight decay of 0.05. Totally, we pre-train our network for 300 epochs, 40 of which are warm-up epochs.

Details for AU detection training: At this training stage, we follow the cropped-and-aligned version of the Aff-Wild2 dataset. Each subject from the training set is recorded with one video sequence. During training, we randomly select a video clip of 16 frames as input to our model. For validating and testing, we split each video data into segments, each of which contains 16 frames. If the number of frames of the segment is less than 16, we supplement it with blank frames. During the training process, we employ an AdamW optimizer with a weight decay of $5e^{-4}$. The number K for choosing the nearest neighbors is set to 4. The learning rate is set to $1e^{-4}$ and adjusted by a cosine decay learning rate scheduler.

Evaluation metrics: We evaluate the AU detection performance of methods by the average F1-score across all AUs. This metric is defined as:

$$F_1^{AU} = \frac{\sum_{i=1}^{N} F_1^{AU,i}}{N} \tag{6}$$

where N denotes the number of AUs, and F1-score $F_1^{AU,i}$for individual AU class is computed as:

$$F_1^{AU,i} = 2 \cdot \frac{P^{AU,i} \cdot R^{AU,i}}{P^{AU,i} + R^{AU,i}} \tag{7}$$

where $P^{AU,i}$ is the calculated precision for the i_{th} AU and $R^{AU,i}$ is the recall rate for it.

4.3. Results on validation set

Tab. 1 presents the evaluation results of AU detection on the validation set, reporting the F1-score for each AU. Our method achieves notable improvement over the baseline, increasing the average F1-score from 39.0 to 54.3. Moreover, we compare our results with those of ME-Graph [31], and our method outperforms theirs by an average F1-score of 3.1. While our overall result is slightly lower than that of the first-place team Netease, we achieved higher F1-score than them in some AU categories. These results demonstrate the effectiveness of our approach in detecting AUs.

Teams	F1-score(in %)
Netease Fuxi Virtual Human [57]	55.5
SituTech	54.2
USTC-IAT-united [54]	51.4
SZFaceU (Ours)	**51.3**
PRL [49]	51.0
CtyunAI [60]	48.9
HSE-NN-SberAI [39]	48.8
USTC-AC [51]	48.1
HFUT-MAC [59]	47.5
SCLAB-CNU [35]	45.6
USC-IHP [53]	42.9
Baseline [20]	36.5

Table 3. Action Unit Detection Challenge Results on test set

4.4. Results on test set

The final results of Action Unit Detection Challenge on test set are presented in the Tab. 3. Specifically,we achieved an average F1-score of 51.3, which places us in 4th position with only a slight difference from the third-place team's score of 51.4. Netease Fuxi Virtual Human and SituTech, who took first and second place respectively, used a similar approach to ours by employing a pre-trained model to extract facial features. Additionally, Netease Fuxi Virtual Human leveraged the multi-modal and temporal information from the videos and implemented a transformer-based framework to fuse the multi-modal features. SituTech also incorporated audio information and employed several ensemble strategies. Meanwhile, the third-place team focused on extracting facial local region features related to AU detection and also utilized a graph neural network to model the relationship between AUs.

4.5. Ablation study

Tab. 2 presents the results of our ablation studies.We choose swin transformer [29] as our baseline. We can observe that the proposed method of incorporating spatial AU graph learning, as shown in (ii), (v), and (viii), leads to significant improvements over the models that lack this feature, as indicated in (i), (iv), and (vii), respectively. Similarly, the inclusion of temporal graph learning, as demonstrated in (iii), (vi), and (ix), yields substantial gains over the models without it, as demonstrated in (ii), (v), and (viii), respectively. These findings highlight the crucial role of modeling the temporal relationships between successive facial frames and the spatial relationships among different AUs in enhancing the accuracy of AU detection. Furthermore, the model pre-trained on ImageNet using vanilla MAE (vi) is not capable of performing better than the baseline (iii). The possible reason could be that there is a significant domain gap between the dataset for universal object recognition and the dataset for facial tasks. However, when we replace the MAE pre-train dataset with the collected hybrid dataset, the model (ix) shows superior performance (54.3 average F1-score) than the baseline (iii) (52.6 average F1-score).

5. Conclusion

This paper proposes an effective spatio-temporal AU relational graph representation learning method for AU occurrence recognition, where MAE is introduced as the facial representation encoder. Experimental results demonstrate that the proposed approach achieved excellent performance in jointly detecting multiple AUs in face videos, which ranked at the 4th place at the 5th Affective Behavior Analysis in-the-wild (ABAW) Competition.

References

[1] Yanan Chang and Shangfei Wang. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20417–20426, 2022. 2

[2] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017. 2

[3] Didan Deng and Bertram E Shi. Estimating multiple emotion descriptors by separating description and inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2392–2400, 2022. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[5] 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. 6

[6] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017. 1

[7] 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. 3

[8] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 1

[9] 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. 2, 3

[10] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2

[11] Shashank Jaiswal, Siyang Song, and Michel Valstar. Automatic prediction of depression and anxiety from behaviour and personality attributes. In *2019 8th international conference on affective computing and intelligent interaction (acii)*, pages 1–7. IEEE, 2019. 1

[12] Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–8. IEEE, 2016. 2

[13] Euiseok Jeong, Geesung Oh, and Sejoon Lim. Multi-task learning for human affect prediction with auditory-visual synchronized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2438–2445, 2022. 1

[14] 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. 1

[15] Zhao Kaili, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016. 2

[16] 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

[17] 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

[18] 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

[19] 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. 1

[20] 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, 5, 6

[21] 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

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

[23] 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

[24] 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

[25] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8594–8601, 2019. 2

[26] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1841–1850, 2017. 2

[27] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 103–110. IEEE, 2017. 2

[28] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer vision and pattern recognition*, pages 10924–10933, 2019. 2

[29] 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. 6

[30] 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. 3, 5

[31] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022. 1, 2, 4, 5, 6

[32] 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. 2, 5

[33] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 10(3):325–347, 2017. 1

[34] 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. 3, 5

[35] Dang-Khanh Nguyen, Ngoc-Huynh Ho, Sudarshan Pant, and Hyung-Jeong Yang. A transformer-based approach to video frame-level prediction in affective behaviour analysis in-the-wild. *arXiv preprint arXiv:2303.09293*, 2023. 6

[36] Dang-Khanh Nguyen, Sudarshan Pant, Ngoc-Huynh Ho, Guee-Sang Lee, Soo-Hyung Kim, and Hyung-Jeong Yang. Affective behavior analysis using action unit relation graph and multi-task cross attention. In *European Conference on Computer Vision*, pages 132–142. Springer, 2023. 1, 2

[37] Hong-Hai Nguyen, Van-Thong Huynh, and Soo-Hyung Kim. An ensemble approach for facial behavior analysis in-the-wild video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2512–2517, 2022. 1, 2

[38] 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(2-4):144–157, 2018. 3, 5

[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] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention and relation learning. *IEEE transactions on affective computing*, 13(3):1274–1289, 2019. 2

[41] Zhiwen Shao, Lixin Zou, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Spatio-temporal relation and attention learning for facial action unit detection. *arXiv preprint arXiv:2001.01168*, 2020. 2

[42] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 13(2):829–844, 2020. 1

[43] Siyang Song, Enrique Sanchez, Linlin Shen, and Michel Valstar. Self-supervised learning of dynamic representations for static images. In *2020 25th international conference on pattern recognition (icpr)*, pages 1619–1626. IEEE, 2021. 2

[44] Siyang Song, Linlin Shen, and Michel Valstar. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 158–165. IEEE, 2018. 1

[45] Siyang Song, Yuxin Song, Cheng Luo, Zhiyuan Song, Selim Kuzucu, Xi Jia, Zhijiang Guo, Weicheng Xie, Linlin Shen, and Hatice Gunes. Gratis: Deep learning graph representation with task-specific topology and multi-dimensional edge features. *arXiv preprint arXiv:2211.12482*, 2022. 1, 2, 4

[46] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. Uncertain graph neural networks for facial action unit detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 1, 2021. 1, 2

[47] Wenyu Song, Shuze Shi, Yu Dong, and Gaoyun An. Heterogeneous spatio-temporal relation learning network for facial

action unit detection. *Pattern Recognition Letters*, 164:268–275, 2022. 2

[48] 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. 1, 4

[49] Tu Vu, Van Thong Huynh, and Soo Hyung Kim. Vision transformer for action units detection. *arXiv preprint arXiv:2303.09917*, 2023. 6

[50] Lingfeng Wang, Jin Qi, Jian Cheng, and Kenji Suzuki. Action unit detection by exploiting spatial-temporal and label-wise attention with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2470–2475, 2022. 1

[51] Shangfei Wang, Yanan Chang, Yi Wu, Xiangyu Miao, Jiaqiang Wu, Zhouan Zhu, Jiahe Wang, and Yufei Xiao. Facial affective behavior analysis method for 5th abaw competition. *arXiv preprint arXiv:2303.09145*, 2023. 6

[52] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 3, 5

[53] Yufeng Yin, Minh Tran, Di Chang, Xinrui Wang, and Mohammad Soleymani. Multi-modal facial action unit detection with large pre-trained models for the 5th competition on affective behavior analysis in-the-wild. *arXiv preprint arXiv:2303.10590*, 2023. 6

[54] 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. *arXiv preprint arXiv:2303.08545*, 2023. 6

[55] 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

[56] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. *arXiv preprint arXiv:2107.03708*, 2021. 1

[57] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Facial affective analysis based on mae and multi-modal information for 5th abaw competition. *arXiv preprint arXiv:2303.10849*, 2023. 5, 6

[58] 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. 1

[59] Ziyang Zhang, Liuwei An, Zishun Cui, Tengteng Dong, et al. Facial affect recognition based on transformer encoder and audiovisual fusion for the abaw5 challenge. *arXiv preprint arXiv:2303.09158*, 2023. 6

[60] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Continuous emotion recognition based on tcn and transformer. *arXiv preprint arXiv:2303.08356*, 2023. 6