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

Zihan Wang\textsuperscript{1,2}, Siyang Song\textsuperscript{3}, Cheng Luo\textsuperscript{1,2}, Yuzhi Zhou\textsuperscript{1,2}, Shiling Wu\textsuperscript{1,2}, Weicheng Xie\textsuperscript{1,2} and Linlin Shen\textsuperscript{1,2} \* \\
\textsuperscript{1}Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University \textsuperscript{2}National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University \textsuperscript{3}School of Computing and Mathematical Sciences, University of Leicester
\{2018112013, lucheng2020, 2200271004, 2110276070\}@email.szu.edu.cn, \{wcxie, l1shen\}@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 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 multimodal features including video and audio to provide multidimensional 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 multidimensional edge features shows strong capability in ex-
plicitly describing the relationship between each pair of AUs. To overcome the overfitting problem, Ma et al. [32] introduce 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 $P = \{p_1, p_2, ..., p_N\}$, 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

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 $S = \{f^1, ..., f^i, ..., f^T\} \in \mathbb{R}^{T \times C \times H \times W}$, 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 $X = \{x^1, ..., x^i, ..., x^T\} \in \mathbb{R}^{T \times m \times d}$ can be generated, where each $x_i \in X$ 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 by 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 spacial dimension as:

$$v_{i}^{t} = \sigma[v_{i}^{t} + g(v_{j}^{t}, \sum_{j=1}^{N} r(v_{j}^{t}, e_{i,j}^{t}))]$$  \hspace{1cm} (1)

where $\sigma$ 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 of which consists of $T$ nodes. For the $i_{th}$ AU sequence $V_i = \{v_{i,1}, ..., v_{i,t}, ..., v_{i,T}\} \in \mathbb{R}^{T \times d}$, 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:

$$\hat{V}_i = Z + \text{FFN}((\text{LayerNorm}(Z)))$$

$$Z = V_i + \text{Att}(Q, K, V)$$

$$Q = V_i W_Q, \quad K = V_i W_K, \quad V = V_i W_V$$  \hspace{1cm} (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:

$$P_i^{t} = \frac{\sigma(\hat{v}_i^{t})^T \sigma(s_i)}{||\sigma(\hat{v}_i^{t})||_2 \sigma ||\sigma(s_i)||_2}$$  \hspace{1cm} (3)

where $\sigma$ is the activation function.
<table>
<thead>
<tr>
<th>AU1</th>
<th>AU2</th>
<th>AU4</th>
<th>AU6</th>
<th>AU7</th>
<th>AU10</th>
<th>AU12</th>
<th>AU15</th>
<th>AU23</th>
<th>AU24</th>
<th>AU25</th>
<th>AU26</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline [20]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>39.0</td>
</tr>
<tr>
<td>ME-Graph [31]</td>
<td>56.1</td>
<td>44.9</td>
<td>52.1</td>
<td>61.7</td>
<td>74.8</td>
<td><strong>75.4</strong></td>
<td>73.4</td>
<td>28.4</td>
<td>18.4</td>
<td>12.4</td>
<td>84.5</td>
<td>32.2</td>
</tr>
<tr>
<td>Netease [57]</td>
<td>55.3</td>
<td><strong>51.4</strong></td>
<td><strong>56.7</strong></td>
<td><strong>67.3</strong></td>
<td><strong>75.8</strong></td>
<td>75.1</td>
<td><strong>75.8</strong></td>
<td>31.2</td>
<td>17.3</td>
<td><strong>33.8</strong></td>
<td>83.9</td>
<td><strong>42.3</strong></td>
</tr>
<tr>
<td>Ours</td>
<td><strong>57.8</strong></td>
<td>48.0</td>
<td>55.9</td>
<td>61.9</td>
<td>75.5</td>
<td>74.6</td>
<td>72.0</td>
<td><strong>35.6</strong></td>
<td>21.6</td>
<td><strong>23.7</strong></td>
<td><strong>86.0</strong></td>
<td>38.3</td>
</tr>
</tbody>
</table>

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

| (i) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| (ii) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| (iii) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| (iv) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| (v) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| (vi) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| (vii) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| (viii) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| (ix) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

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 \( L_{\text{pre}} \) is formulated as:

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

(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 major 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 \( L_{\text{au}} \) is denoted as:

\[
L_{\text{au}} = -\sum_{i=1}^{N}\sum_{t=1}^{T}[y_{it} \log(p_{i}^{t}) + p_{i}^{t}(1 - y_{it}) \log(1 - p_{i}^{t})]
\]  

(5)

where \( p_{i}^{t}, y_{it} \) are the prediction and ground truth; \( N \) and \( T \) are the numbers of AUs and frames of the input face sequence, respectively. The first \( \log(p_{i}^{t}) \) 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,
we compare our results with those of ME-Graph [31], and
ing the average F1-score from 39.0 to 54.3. Moreover,
method achieves notable improvement over the baseline, in-

4.2. Experimental settings

**Details for MAE pre-training:** We first leverage Reti-
naFace [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 ran-
domly select a video clip of 16 frames as input to our model.
For validating and testing, we split each video data into seg-
ments, 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 num-
ber $K$ 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 per-
formance of methods by the average F1-score across all
AUs. This metric is defined as:

$$F_{1}^{AU} = \frac{\sum_{i=1}^{N} F_{1}^{AU,i}}{N} \quad (6)$$

where $N$ denotes the number of AUs, and $F_{1}^{AU,i}$ for
individual AU class is computed as:

$$F_{1}^{AU,i} = 2 \cdot \frac{P_{AU,i} \cdot R_{AU,i}}{P_{AU,i} + R_{AU,i}} \quad (7)$$

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

**Table 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, in-
creasing 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 demon-
strate the effectiveness of our approach in detecting AUs.

4.4. Results on test set

The final results of Action Unit Detection Challenge on
test set are presented in the **Table 3**. Specifically, we achieved
an average F1-score of 51.3, which places us in 4th posi-
tion 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 simi-
lar approach to ours by employing a pre-trained model to
extract facial features. Additionally, Netease Fuxi Virtual
Human leveraged the multi-modal and temporal informa-
tion from the videos and implemented a transformer-based
framework to fuse the multi-modal features. SituTech also
incorporated audio information and employed several en-
semble strategies. Meanwhile, the third-place team focused
on extracting facial local region features related to AU de-
tection 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 ob-
serve that the proposed method of incorporating spatial AU
graph learning, as shown in (ii), (v), and (viii), leads to sig-
ificant 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 mod-
els without it, as demonstrated in (ii), (v), and (viii), re-
spectively. These findings highlight the crucial role of mod-
eling 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


[9] Kaiming He, Xunlei 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


