Facial Action Unit Recognition Guided by Labeling Rules

Abstract

Existing facial action unit (AU) recognition studies either ignore AU correlations, or disregard important facial cues for AU judgment. To address these limitations, we design a novel AU recognition framework guided by AU labeling rules. Specifically, we first summarize AU labeling rules from the Facial Action Coding System (FACS) to separate facial judgment areas and define the explicit correspondence between AUs and the judgment areas. We design a region feature extraction component to extract representations of judgment areas. Then, AU-specific representations are mapped from their corresponding judgment areas. AU correlations are further encoded to enhance the AU representation learning. After that, we introduce a region relation learning component to encode the correlations among judgment areas to further guide the region representation learning. Finally, the encoded AU and region patterns are jointly fed into the AU predicting component to perform AU recognition, fully leveraging both AU correlations and facial cues. Experimental results on two benchmark databases demonstrate the effectiveness of the proposed method compared with that of current state-of-the-art methods.

1 Introduction

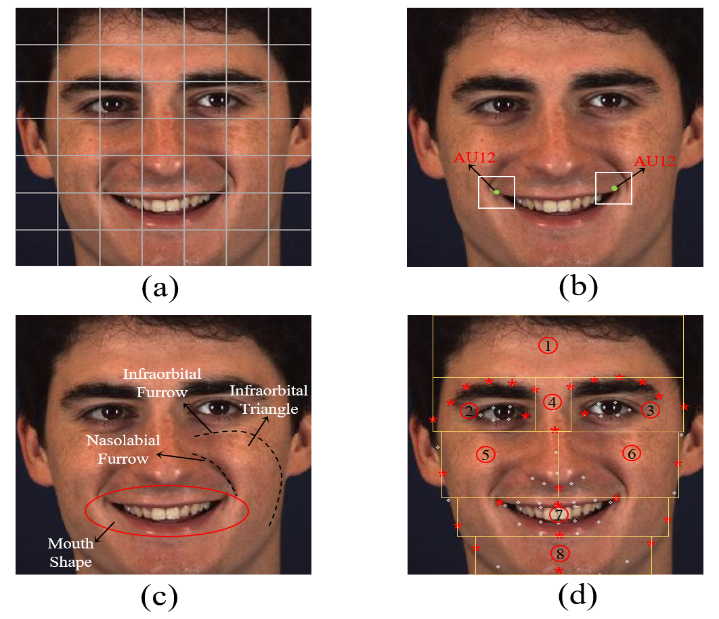
Facial action units (AUs), defined by the Facial Action Coding System (FACS) [1], describe the contraction or relaxation of one or more facial muscles. Automatic facial AU recognition has recently become a popular research topic owing to its wide-ranging applications, such as describing facial expressions and monitoring health.

Figure 1: (a) An example of separated patches. (b) The centers of AU12; The rectangles are regions of interest. (c) The main facial cues for judging AU12. (d) The proposed eight facial judgment regions.

With the development of deep learning, several works [2, 3] extract deep global facial representations and treat facial AU recognition as a multi-label classification problem. However, AUs are usually related to local facial regions. These aforementioned methods, which ignore the local property, demonstrate limited performance. The majority of recent works aim to leverage the local property to enhance AU recognition. These methods either focus on improving local facial representation learning or explicitly extracting AU-specific patterns by defining AU centers. The former approach [4, 5] separates facial regions into multiple local patches. The relationship among local patches is exploited for local representation learning. Figure [1](#fig:example)a shows an example of local patches. Facial local patterns are fully considered for AU prediction by jointly considering multiple patch representations. However, the method used to separate the patches disregards the correspondence between patches and AUs. Therefore, these works cannot directly extract AU-specific patterns. Thus, they cannot include correlations among to enhance AU representation learning. The latter [6, 8] aims to directly extract AU-specific local patterns by defining AU centers. Figure [1](#fig:example)b shows the centers of AU12 (lip corner puller), where the two rectangles are the regions of interest (ROI). AU-specific representations are directly extracted from the related ROI. The AU correlations are exploited using a graph neural network or the self-attention mechanism to enhance AU recognition. However, these methods tend to ignore important facial cues for AU judgment because of the limited areas included near the AU centers. Figure [1](#fig:example)c shows the main facial cues for judging AU12 according to FACS. In addition to considering the limited regions around the AU centers, AU experts consider the lip corners rising obliquely, nasolabial furrow deepening, infraorbital triangle rising, and infraorbital furrow to judge AU12. Overall, previous works either ignore the correlations between AU representations or disregard important facial cues because of the AU center limitations.

To address these drawbacks, we propose a novel AU recognition method inspired by AU labeling rules to thoroughly leverage AU correlations and AU-related local facial patterns. Specifically, AU labeling rules are first summarized according to FACS. The facial area is divided into eight judgment regions according to the location of AU-related facial judgment cues, as shown in Figure [1](#fig:example)d. Certain correspondences between AUs and the judgment areas are also summarized. For each AU, the facial cues from the corresponding facial judgment regions are jointly considered to learn AU-specific patterns. Then, we introduce a region feature extraction component to extract the local features of eight judgment regions by using an RoIAlign layer. We leverage the correspondences between judgment areas and AUs to learn AU-specific representations by a mapping layer. AU correlations are encoded to enhance AU-specific representation learning based on the self-attention mechanism. Moreover, the patterns of different judgment regions are related with AU activation. We introduce a region relation learning component to encode the relationships among judgment areas to facilitate local representation learning based on the self-attention mechanism. Finally, both the encoded AU and region representations are fused to perform AU recognition using the AU prediction component. Thus, both AU correlations and facial cues from judgment areas are fully exploited to enhance AU recognition.

The contributions of the paper are as follows. We propose a facial AU recognition method guided by FACS labeling rules. We divide facial judgment regions and summarize explicit correspondences between AUs and facial regions according to the labeling rules. AU-specific representations are learned by jointly considering the corresponding judgment areas, and AU correlations are further encoded to enhance AU representation learning. The relationships between judgment areas are also encoded to enhance region representation learning. Both encoded AU and region representations are fused to enhance AU recognition. AU correlations and facial cues are fully leveraged in this way. Experimental results on two benchmark databases show that the proposed method outperforms the state-of-the-art methods.

2 Related work

A comprehensive survey on facial AU recognition is provided in [9]. This section briefly reviews recent advances in deep AU recognition. The deep methods can be divided into two categories: AU recognition from global facial representations and AU recognition from local facial patterns.

Several works leveraged powerful deep neural networks to extract global facial representations and treated AU recognition as a multi-label classification problem. For example, Chu et al. [2] proposed a hybrid network architecture to jointly capture spatial and temporal representations using the convolutional neural network (CNN) and the long short-term memory neural network (LSTM), respectively. Han et al. [3] proposed an optimized filter-size CNN for AU recognition, thereby estimating optimal kernel size for varying image resolutions. Song et al. [10, 11] generated AU features by predicting the probability of one specific AU from global facial representations. The graph neural network was used to exploit AU dependencies. These approaches ignore the local property and thus demonstrate limited performance because AUs are usually related to local facial areas. Hence, improved AU recognition results are expected by leveraging the local property.

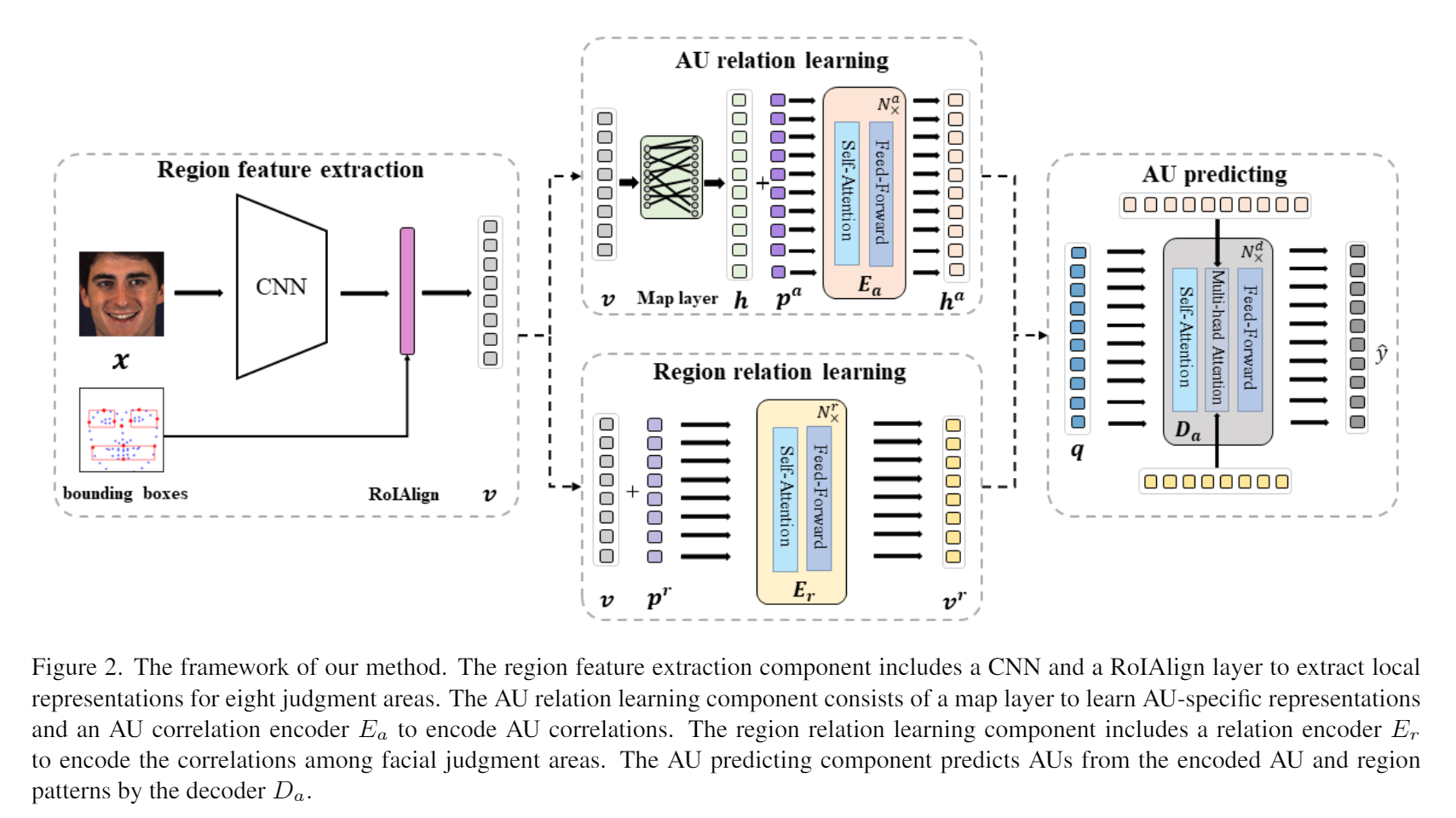
The majority of recent works, which leverage local facial patterns for AU recognition, include improving local facial representation learning and explicitly extracting AU-specific patterns by defining AU centers. Several works separated facial regions into multiple local patches and the correlations between patches were leveraged to enhance local representation learning. Patterns from different patches were jointly considered to perform AU recognition. For example, Zhao et al. [12] proposed deep region and multi-label learning to address region learning and multi-label learning simultaneously. The region layer was introduced to capture local appearance changes for different facial regions. Corneanu et al. [13] proposed a deep structure inference network for AU recognition. Global and local patch information was jointly considered to predict AUs. Ertugrul et al. [4] proposed a patch-attentive network for AU recognition, where each patch was encoded individually. An attention mechanism was used to weight the contribution of each patch. Niu et al. [5] used local information and the correlations of individual local face regions to enhance AU recognition. A local relationship learning module was proposed to explore the underlying relationships among local facial regions. None of these methods explicitly consider the correspondence between AUs and local patches. AU-specific representations cannot be learned directly from the local patches. Furthermore, the correlations between AU-specific representations are ignored.

To leverage AU correlations, several works directly extracted AU-specific local patterns by defining AU centers. The correlations among AU representations were involved to enhance AU recognition. For example, Li et al. [6] proposed a framework that integrated region of interest adaption and optimal LSTM-based temporal fusing for AU recognition. AU-specific representations were extracted by the proposed cropping layers. Shao et al. [14] jointly performed AU detection and face alignment, whereby the adaptive attention learning module was used to locate the ROIs of the AUs for better local feature learning. Li et al. [7] proposed semantic relationship-guided representation learning for AU recognition. AU correlations were embedded into the representation learning based on a knowledge-graph. Jacob et al. [8] performed AU recognition based on the transformer encoder. The regions of interest for each AU were used to extract AU-specific representations. Tang et al. [15] proposed a pixel-interested learning method to learn pixel-level attention and thereby improve AU detection. The predefined pixel interest maps are defined for each AU in the first stage of the algorithm. All of these methods extract AU-specific representations by defining AU centers. The regions around these centers are limited and often disregard important facial cues for judging AUs, limiting their overall performance. In this paper, we learn AU-specific representations according to labeling rules. Facial cues are fully leveraged to enhance AU recognition by jointly considering the encoded AU and region patterns.

In this paper, we leverage AU labeling rules to separate facial judgment areas and define certain correspondences between AUs and the judgment areas. AU-specific representations can be learned from the related judgment areas. AU correlations are further encoded to enhance AU-specific representation learning.

3 Problem statement

Let denote the training samples, where represents a facial image, represents the AU label, is the index of the AU label, and is the number of training samples. The goal is to learn a region feature extraction network to extract representations of AU-related judgment areas, and to leverage the learned representations to perform AU recognition. Both AU correlations and facial cues are fully employed to enhance performance. Let indicate the set of unlabeled testing images. Given an unseen facial image , we can apply the learned network to predict the AU labels.



4 Methodology

Figure [2](#fig:framework) illustrates the framework of the proposed method, which includes region feature extraction, AU relation learning, region relation learning, and AU prediction. The region feature extraction component extracts representations of AU-related judgment areas according to AU labeling rules. The AU relation learning component learns AU-specific representations from the judgment regions, and the AU relations are encoded to improve the learned representations. The region relation learning component encodes the correlations among different facial regions to enhance the region representation learning. Finally, the AU prediction component fuses the encoded AU and region representations to perform AU recognition. Thus, the proposed method boosts AU recognition by thoroughly leveraging facial cues and AU correlations.

### 4.1 AU labeling rules

We first summarize AU labeling rules according to FACS. Table [1](#table:au_areas) shows the main appearance changes guiding manual AU labels. For example, the main appearance changes for labeling AU4 (Brow Lowerer) are the eyebrows lowering and pulling together. Wrinkling or muscles bulging between the eyebrows are also involved. For AU17 (Chin Raiser), the main components are pushing up the lower lip and wrinkling the chin boss. According to the locations of appearance changes, we can divide the facial area into eight judgment regions, as shown in Figure [1](#fig:example)d. Several facial landmarks labeled with a red asterisk indicate the regions. Table 1 lists the judgment areas for each AU. For example, areas 2, 3, and 4 are used to judge AU4; while the judgment areas for AU17 are 7 and 8.

We can leverage the summarized rules in two ways. Firstly, we can learn AU-specific representations by jointly considering the appearance changes of corresponding judgment regions. AU correlations are further leveraged to enhance AU representation learning. Secondly, there are correlations between local judgment areas owing to the activation of AUs. We can leverage the region correlations to improve the region representation learning. Jointly considering both aspects ensures that facial cues and AU correlations are thoroughly leveraged for AU recognition guided by labeling rules.

### 4.2 Region feature extraction

The region feature extraction component extracts representations for eight facial judgment regions. First, a mini-batch of N images, , is randomly sampled from . According to Figure [1](#fig:example)d, we leverage 68 landmarks to generate bounding boxes for eight judgment regions, . A CNN module extracts global facial representations. Then, the global features and bounding boxes are fed into an RoIAlign layer to extract 8N representations of judgment areas, .

In contrast to existing approaches, we divide facial judgment areas according to labeling rules. There are explicit correspondences between judgment areas and AUs. The extracted representations are leveraged to learn AU-specific patterns. AU correlations are further encoded to enhance AU representation learning.

### 4.3 AU relation learning

For each AU, there are several related judgment regions. We first introduce a map layer to learn AU-specific representations from the region representations . Equation 1 defines the mapping correlations as:

, (1)

where is a convolution of the k-th AU, and denotes concatenating all of the representations of the AU-related judgment areas according to Table [1](#table:au_areas). The AU-specific representations are thus learned from the related facial judgment areas.

We introduce the self-attention mechanism to encode AU correlations to enhance AU representation learning. The representations are added to a position embedding component, . The sum is fed into a transformer encoder to encode the correlations among AUs. The transformer encoder is composed of a stack of identical layers as in [16, 17]. Each layer comprises a multi-head self-attention component and a feed-forward network. The encoded representations are denoted as . To guide the representation learning, a set of classifiers are trained to predict AUs from the encoded representations. Equation 2 defines the loss function, where denotes the predicted AU labels and denotes cross-entropy loss.

. (2)

The AU relation learning component learns AU-specific representations from the corresponding facial regions. The approach preserves the facial cues for AU judgment and further encodes the AU correlations to enhance AU representation learning.

### 4.4 Region relation learning

The activation of AUs spurs specific appearance changes in the related judgment regions. The region changes depend on the particular activated AUs, and correlations between different facial judgment regions are meaningful for AU recognition. We introduce a region relation learning component to encode the correlations among different facial regions to enhance region representation learning. Specifically, after the region feature extraction, the representations, , are fed into the region relation learning component to further learn the correlations between different regions. are first added to position embedding to retain the positional information. The sum is fed into a region transformer encoder to encode region correlations. The structure of is similar to , with layers. The encoded region representations are denoted as .

To guide the encoding of AU-related region representations, we introduce an auxiliary task to supervise the processing according to labeling rules. The aim of the auxiliary task is to predict the regions related to the activated AUs. For image , the label is defined as . When AUs related to the m-th facial region are activated, the m-th bit of is set to 1. A set of classifiers are trained to predict the activated regions from . Equation 3 defines the corresponding loss function:

, (3)

where is cross-entropy loss, and is the predicted label of activated regions. The task predicts AU-related facial regions. The auxiliary task thus encodes the correlations among different judgment regions to enhance the region representation learning.

### 4.5 AU prediction

The AU relation learning component learns AU-specific representations from judgment areas and encodes AU correlations to enhance AU representation learning. The region relation learning component encodes region correlations to improve region representation learning. We introduce the AU prediction component to combine the two components to further improve AU recognition. Specifically, we combine and to form . Then, the combined patterns are fed into transformer decoder to predict the AU labels. The transformer decoder is composed of layers as in [16, 17]. In addition to the self-attention component and the feed-forward network, we use a multi-head attention component over the combination of encoded AU and region patterns. The inputs of are K AU queries, . The AUs are predicted from the outputs of by training a set of classifiers. Equation 4 defines the corresponding loss function, where are the predicted AU labels:

. (4)

The AU prediction component jointly leverages encoded AU and region patterns to predict AUs. Thus, our approach considers both facial cues and AU correlations to enhance AU recognition.

### 4.6 Overall learning

The overall learning loss is defined in Equation 5, where , , and are weighted coefficients:

. (5)

We design the overall framework according to AU labeling rules. First, the AU relation learning component leverages the correspondences between AUs and judgment areas to learn AU-specific representations from judgment areas. The correlations among different AUs are encoded to enhance AU representation learning. Second, the auxiliary task enables the region relation learning component to encode the region correlations, thereby improving local representation learning of judgment regions. Finally, the AU prediction component fuses the encoded AU representations and region patterns to predict AUs. Both facial cues and AU correlations are fully leveraged to enhance AU recognition. Note that we obtain two AU predictions, one from the AU relation learning component and the other from the AU prediction component. The output from the AU prediction component forms the final prediction during inference.

5 Experiments

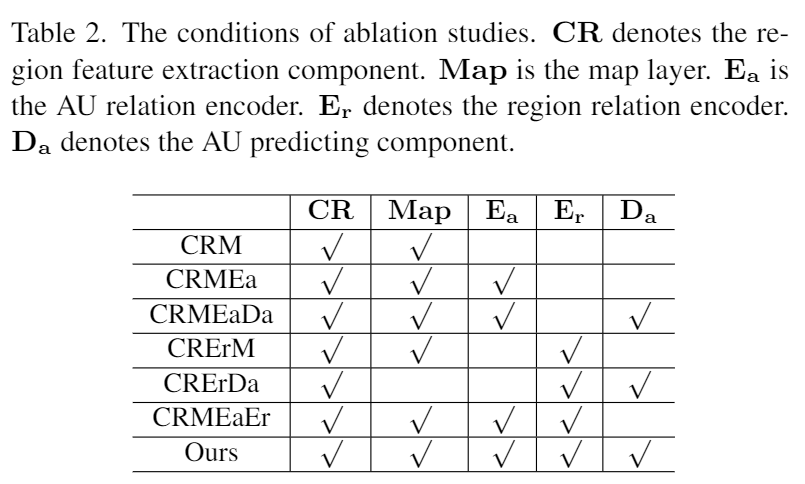
### 5.1 Experimental conditions

We evaluate our approach on two benchmark databases, including the BP4D database [18] and the Denver Intensity of Spontaneous Facial Action (DISFA) database [19].

The BP4D database includes 2D and 3D spontaneous facial videos of eight tasks recorded from 41 subjects. In total, 328 two-dimensional videos are coded with 12 AUs: 1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, and 24. We use all of the approximately 140,000 available image samples.

The DISFA database provides spontaneous facial videos from 27 subjects. Each AU includes an intensity label ranging from zero to five. In total, there are approximately 130,000 valid samples. We consider the same eight AUs that most related works study, i.e., 1, 2, 4, 6, 9, 12, 25, and 26. An AU with an intensity greater than or equal to two is considered active.

To perform the comparison with recent studies under the same conditions, we conduct a three-fold subject-independent cross validation on the BP4D and DISFA databases. We conduct several ablation experiments. CRM leverages the **r**egion feature extraction component and the **m**ap layer to extract AU-specific representations. The AUs are predicted by training a set of classifiers from the AU representations. CRMEa leverages the **r**egion feature extraction component, the **m**ap layer, and the AU relation encoder to perform AU recognition. CRMEaDa leverages the **r**egion feature extraction and the AU relation learning components to encode AU patterns (thereby including the **m**ap layer and ***Ea***), and uses the AU prediction component (***Da***) to predict AUs from the encoded AU patterns. CRErM leverages the **r**egion feature exaction and region relation learning component (***Er***) to learn region patterns, and learns AU-specific representations using the **m**ap layer, to predict AUs. CRErDa leverages the **r**egion feature exaction and ***Er*** components to learn region patterns, and predicts AUs using ***Da***. CRMEaEr jointly trains the region feature exaction, AU relation learning, and region relation learning components. The AUs are predicted from the outputs of the AU relation learning components. Our method considers all of the components together. Table [2](#table:ablation) shows the conditions of the ablation studies. We use the frame-based F1 score to assess AU recognition.



### 5.2 Implementation details

We first crop and resize the facial images to pixels and three color channels in all experiments. For the region feature extraction component, the CNN is based on a ResNet\_50 network [20]. The output of the conv4\_x layer forms the extracted global representations. The RoIAlign layer is designed according to [21]. We leverage the facial landmark detection method proposed in [22] to detect 68 facial landmarks for each facial image. Both the region and AU encoders include six layers. Each layer combines an 8-heads self-attention layer and feed-forward layer. The AU prediction component has six layers. Each layer combines an 8-heads self-attention layer, 8-heads multi-head attention layer, and feed-forward layer. The classifiers for AU and region prediction are multilayer perceptron (MLP) networks with two hidden layers of sizes 256 and 128. The overall framework is trained end-to-end by minimizing the loss function in Equation [5](#Equ5), where . The framework is implemented using PyTorch [23] and trained with the Adam optimizer [24]. We use an initial learning rate of 0.00001 and batch size of 32.

### 5.3 Ablation studies

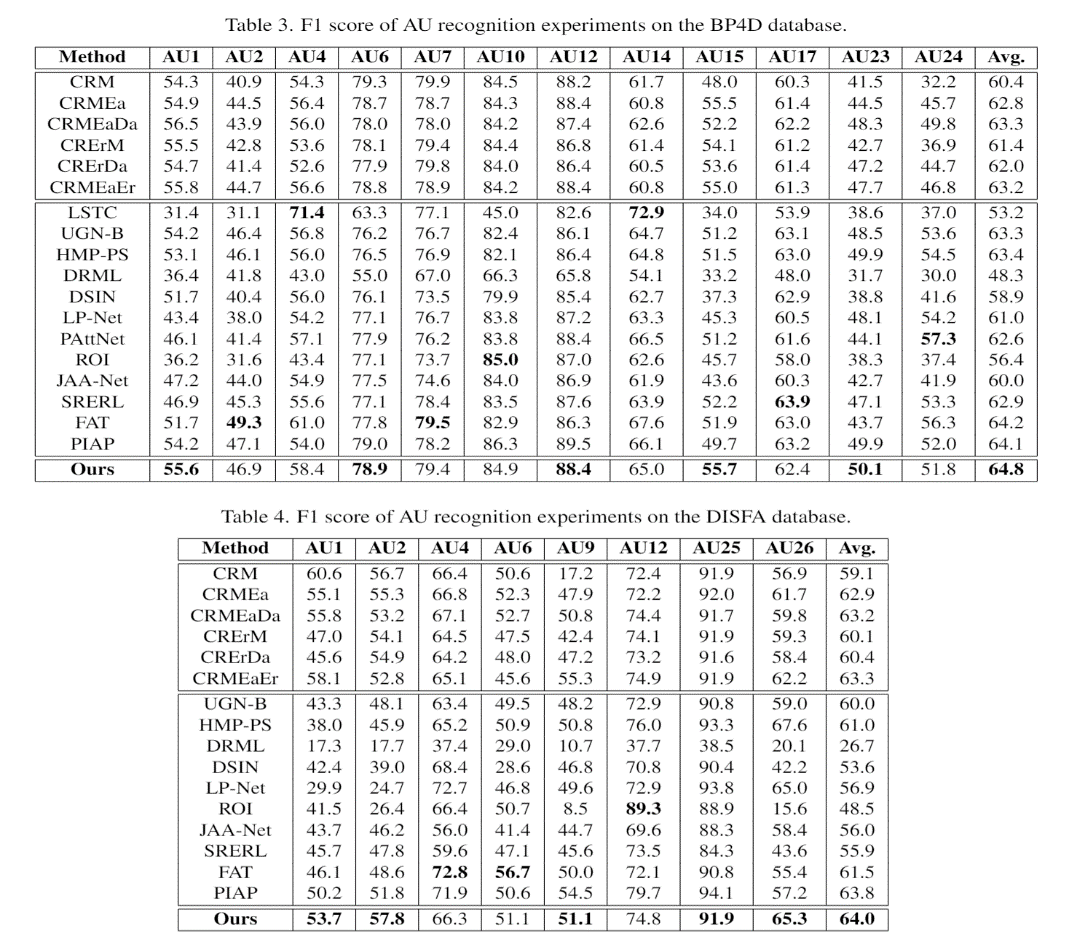
In this section, we evaluate the components of the proposed method. Tables [3](#table:bp4d) and [4](#table:disfa) show the experimental results on the BP4D and DISFA databases. The two tables yield the following observations:

First, CRMEa and CRMEaDa achieve better results than CRM. Specifically, CRMEa improves the average F1 scores by 2.4% and 3.8% over CRM on the BP4D and DISFA databases, respectively. CRMEaDa achieves a 2.9% and 4.1% F1 score improvement over CRM on BP4D and DISFA, respectively. CRM only extracts the facial local representations of judgment areas. Both CRMEa and CRMEaDa encode AU correlations and thus significantly enhance AU recognition.

Second, CRErM and CRErDa also outperform CRM. CRErM improves the F1 scores by 1.0% over CRM on both databases. CRErDa improves the F1 scores by 1.6% and 1.3% over CRM on BP4D and DISFA, respectively. These results indicate that encoding region correlations improves facial region representation learning and boosts AU recognition.

Third, CRMEaEr outperforms CRMEa and CRErM. Specifically, CRMEaEr improves the F1 score by 0.4% over CRMEa on the BP4D and DISFA databases, and 1.8% and 3.2% over CRMEaEr on BP4D and DISFA, respectively. CRMEaEr jointly performs the AU recognition from encoded AU representations and the region prediction from encoded region features. These two tasks form multi-task learning. Both region and AU correlations are leveraged to enhance AU recognition based on the shared region feature extraction component.

Fourth, CRMEaDa outperforms CRMEa. CRMEaDa improves the F1 scores by 0.5% and 0.3% over CRMEa on the BP4D and DISFA databases, respectively. CRErDa improves the F1 scores by 0.6% and 0.3% over CRErM on BP4D and DISFA, respectively. Furthermore, our method outperforms CRMEaEr. CRMEa predicts AUs from the AU representations encoded with AU correlations. CRErM encodes region correlations to enhance facial region representations. CRMEaEr jointly encodes AU and region correlations. Our method, CRMEaDa, and CRErDa, decode AUs from the encoded AU or region representations. The superior performance of these three approaches demonstrates that leveraging the AU prediction component based on the self-attention mechanism further enhances AU recognition.

Finally, our method outperforms CRMEaDa and CRErDa. Our proposed method improves the F1 scores by 1.5% and 0.8% over CRMEaDa on the BP4D and DISFA databases, respectively. CRMEaDa only learns AU-specific representations and encodes AU correlations. Our method leverages encoded AU representations and region patterns to jointly and comprehensively consider facial cues and AU correlations. Our proposed method improves the F1 score by 2.8% and 3.6% over CRErDa on BP4D and DISFA, respectively. It demonstrates that exploiting AU correlations facilitates AU recognition.

### 5.4 Comparisons to state-of-the-art work

To further demonstrate the effectiveness of our proposed method, we compare it with recent works under the same experimental conditions. We compare our method’s extraction of global facial features for AU recognition with LSTC [2], HMP-PS [11], and UGN-B [10]. Recent work leveraging AU local properties assume one of two approaches: AU recognition based on separating local facial patches, or AU recognition based on locating AUs. For the former, we compare our work with recent methods: DRML [12], DSIN [13], LP-Net [5], and PAttNet [4]. For the latter, we compare our proposed method with: ROI [6], JAA-Net [14], SRERL [7], PIAP [15], and FAT [8]. We compare the method with all available results. Tables [3](#table:bp4d) and [4](#table:disfa) show the results on the BP4D and DISFA databases. For easy comparison, the best results for each AU are in bold font.

First, our proposed method outperforms LSTC. Our method achieves an 11.6% improvement of the F1 score over LSTC on the BP4D database. Our method achieves 1.5% and 4.0% F1 score improvements over UGN-B on the BP4D and DISFA databases, respectively. Our method achieves 1.4% and 3.0% F1 score improvements over HMP-PS on BP4D and DISFA, respectively. The previous methods ignore the local property of AUs. Our method leverages AU labeling rules to separate facial judgment areas and learn AU-specific patterns. Thus, using the AU local property improves AU recognition.

Second, our proposed method outperforms AU recognition approaches based on separating local facial patches. On the BP4D database, our method improves the F1 score by 16.5%, 5.9%, 3.8%, and 2.2% over DRML, DSIN, LP-Net, and PAttNet, respectively. On the DISFA database, our method improves the F1 score by 37.3%, 10.4%, and 7.1% over DRML, DSIN, and LP-Net, respectively. These methods separate facial regions into multiple local patches, and leverage the correlations among patches to enhance local representation learning. However, they do not define the correspondence between AUs and local patches. Thus, AU correlations are not involved to further enhance AU recognition. Our method is guided by AU labeling rules and considers the correspondences between facial judgment areas and AUs. AU correlations are encoded to demonstrably enhance AU recognition.

Finally, our method outperforms recent methods locating AUs. On the BP4D database, our method achieves F1 score improvements of 8.4%, 4.8%, 1.9%, 0.6%, and 0.7% over ROI, JAA-Net, SRERL, FAT, and PIAP, respectively. On the DISFA database, the proposed method improves the F1 score by 15.5%, 8.0%, 8.1%, 2.5%, and 0.2% over ROI, JAA-Net, SRERL, FAT, and PIAP, respectively. The existing methods directly extract AU-specific local patterns by defining AU centers and enhance AU recognition by leveraging AU correlations. However, the regions near AU centers are limited and disregard important facial cues. Our method separates facial regions into multiple judgment areas guided by AU labeling rules. For each AU, the facial cues from its related judgment areas are jointly considered to learn AU-specific representations. In the AU prediction component, both the encoded AU and region representations guide AU recognition, fully considering cues from facial judgment areas.

Overall, our method fully leverages AU correlations and facial cues to enhance AU recognition compared with existing methods.

### Examples of the attention mechanism in the last layer of AU encoder E_a and region encoder E_r.5.5 Discussion of AU and region correlations

Figure 3: Examples of the attention mechanism in the last layer of AU encoder and region encoder .

To further demonstrate the effectiveness of the proposed method, we visualize the last self-attention layers of the AU correlation encoder and region relation encoder . Figure [3](#fig:visualize) shows the examples. First, encodes the correlations between AUs. For example, in the first column, both AU6 and AU7 are activated. For the feature encoding of AU6, the two co-occurrence AUs correspond to high weights. In the second column, the related AU14 and AU24 have high weights for the feature encoding of AU24. Second, encodes the region correlations. The related judgment regions of AU6, areas 2, 3, 5, and 6, show high weights for the region feature encoding of area 2. The region feature encoding of area 8 is highly related to area 7. Thus, our method jointly encodes AU and region correlations.

6 Conclusion

In this paper, we propose a novel AU recognition method guided by labeling rules. Specifically, previous works either ignore AU correlations, or miss important facial cues for AU prediction. We summarize AU labeling rules to separate facial judgment areas and define the correspondence between judgment areas and AUs. We leverage a region feature extraction component to extract local representations for judgment areas. An AU relation learning component learns AU-specific representations from its related judgment areas and encodes AU correlations. We also leverage a region relation learning component to enhance the representation learning of judgment areas. Finally, both the encoded AU and region patterns guide AU prediction. Experimental results on two benchmark databases demonstrate that the proposed method fully leverages AU correlations and facial cues to enhance AU recognition compared with previous supervised methods.