Weakly-Supervised Text-driven Contrastive Learning for Facial Behavior Understanding

Xiang Zhang[1]   Taoyue Wang[1]   Xiaotian Li[1]   Huiyuan Yang[2]   Lijun Yin[1]

[1]State University of New York at Binghamton   [2]Rice University

{zxiang4, twang61, xli210, lyin}@binghamton.edu   hy48@rice.edu

Abstract

Contrastive learning has shown promising potential for learning robust representations by utilizing unlabeled data. However, constructing effective positive-negative pairs for contrastive learning on facial behavior datasets remains challenging. This is because such pairs inevitably encode the subject-ID information, and the randomly constructed pairs may push similar facial images away due to the limited number of subjects in facial behavior datasets. To address this issue, we propose to utilize activity descriptions, coarse-grained information provided in some datasets, which can provide high-level semantic information about the image sequences but is often neglected in previous studies. More specifically, we introduce a two-stage Contrastive Learning with Text-Embeded framework for Facial behavior understanding (CLEF). The first stage is a weakly-supervised contrastive learning method that learns representations from positive-negative pairs constructed using coarse-grained activity information. The second stage aims to train the recognition of facial expressions or facial action units by maximizing the similarity between the image and the corresponding text label names. The proposed CLEF achieves state-of-the-art performance on three in-the-lab datasets for AU recognition and three in-the-wild datasets for facial expression recognition.

1. Introduction

Facial expression is one of the most natural signals to analyze human emotion and behavior. Ekman [13] has indicated that facial expressions of emotion are universal across human cultures and categorized them, apart from neutral expression, into six categories: anger, disgust, fear, happiness, sadness, and surprise. Then contempt was added as another basic emotion, according to the work [35]. Furthermore, facial expressions are coded by specific facial muscle movements, called Action Units (AUs) in Facial Action Coding System (FACS) [14]. Automatic Facial Expression Recognition (FER) and Action Unit recognition (AUR) have been
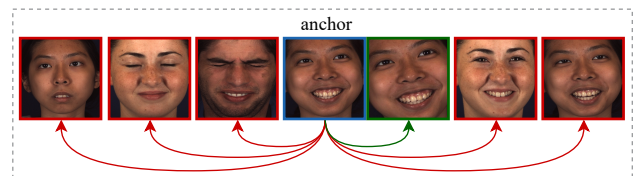
(a) Self-supervised contrastive learning pairs
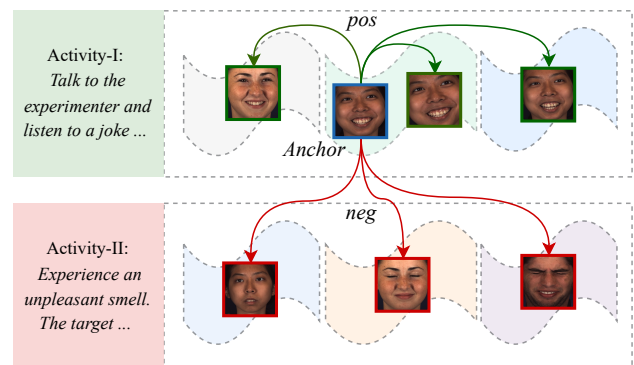
(b) Activity-based weakly-supervised contrastive learning pairs

Figure 1: (a) shows the self-supervised contrastive learning paring, where green represents positive pairs and red represents negative pairs. In a batch, the only positive samples for an anchor are its augmentations, while all others are negative. Even if the last image is similar (same person and same expression) to the anchor, it will be pushed away from the anchor as a negative sample. (b) is the illustration of the proposed weakly-supervised contrastive learning method: samples from the same activity in a batch are selected as positive and the remaining are negative. The textual activity descriptions are used as coarse-grained information to guide contrastive learning, for example, "talk to the experimenter and listen to a joke ..."

core problems in facial analysis, attracting significant interest in the computer vision community.

Recently, many deep learning-based approaches [29, 54, 39, 42, 56, 18, 4] have been proposed and achieved state-of-

the-art performance in FER and AUR. A variety of methods [54, 52, 26, 62] aimed to disentangle the expression or AU features from various disturbing factors, such as identity, ethnic background, pose, etc. Along with the development of Self-Supervised Learning (SSL), unlabeled data is utilized for learning good representations to improve recognition performance. Chang et al. [4] proposed a rule that divides the face into eight regions, which are then fed in a contrastive learning component. Shu et al. [43] explored three core strategies in self-supervised contrastive learning to enforce expression-specific representations and minimize interference from other facial attributes. FaRL [64] proposed a vision-language pre-training model with a large number of facial image-text pairs to learn facial representation. To build an appropriate self-supervised learning task, fine-grained auxiliary information, such as landmarks and image captions, is typically required, which in turn requires more data processing.

On the other hand, several works have investigated the different relations between AU pairs and their applications. SRERL [22] was developed to learn the appearance representation of the semantic relationships between AUs by a graph convolutional network. Yang et al. [56] proposed a cross-modal attention module to enhance the image representations by including AU semantic descriptions. However, due to the low consistency between the data structure of image and text, attention-based integration may not fully exploit the potential of textual data. Some works also modeled the AUs' relationships with the expressions to improve the FER performance. Cui et al. [10] employed a Bayesian Network(BN) to capture the generic knowledge on relationships among AUs and expression. In our work, we are interested in learning the direct relationships between expressions and between AUs in a simpler way. Moreover, previous studies on relationship learning have rarely explored the representation of ground truth labels, instead focusing on fitting the model with numerical labels, thus sparking our interest in investigating label representation.

In order to overcome the above limitations, it is necessary to investigate the following two issues: *i) whether there is any coarse-grained information, which can be **easily obtained and simple to use** without compromising the performance*; *ii) whether there is any approach to **enrich the relationship information** of the label representation*.

To address the above two issues, we propose a text-driven contrastive learning method, called CLEF, to utilize both the coarse-grained information and text-embedded labels. The proposed method comprises two stages, both using a unified vision-text architecture known as CLIP [38]. In pre-training, for each anchor in a batch, we consider positive samples from the same activity and negative samples from different activities. The activity descriptions are used as coarse-grained labels to guide the weakly-supervised

Table 1: Activity description samples in BP4D. See more descriptions in the Supplementary Material.

	Activity Description
A1	Talk to the experimenter and listen to a joke (Interview). The target emotion is happiness or amusement
A2	Watch and listen to a recorded documentary and discuss their reactions. The target emotion is sadness
A3	Experience sudden, unexpected burst of sound. The target emotion is surprise or startle
A4	Play a game in which they improvise a silly song. The target emotion is embarrassment

contrastive learning model that aims to minimize the intra-activity differences in representations. Table 1 shows some samples of activity descriptions of BP4D [58]. Figure 1b shows how we leverage the activity descriptions to create positive-negative pairs. Each activity contains multiple expressions, but our pairing construction can increase the possibility of grouping images with the same expression into positive ones. The distance between images belonging to different activities increases, even if the images have the same identities, which encourages the encoder to focus on the activity features rather than the identity features. Meanwhile, the activity text description does not contain any identity information, allowing the text encoder to avoid encoding identity features. Cross-modal contrastive learning is therefore designed to push image features close to such textual features. Performing on these pairs can enhance the learning of better representations, which in turn improves the performance of FER or AUR in downstream tasks.

In fine-tuning, we apply vision-text contrastive learning directly to classification tasks. Supervised contrastive learning adapts the image representation to be close to its corresponding label name feature, while self-supervised contrastive loss encourages the feature of label names and descriptions to be similar, enriching the semantic information of the label representation. Therefore, we believe such label representation is more powerful than the numerical label. The recognition prediction is based on finding the most similar label names of the testing image, following the method used in CLIP [38]. The main contributions of this paper are summarized in three aspects:

1. We proposed a weakly-supervised contrastive learning method that effectively leverages coarse-grained activity information. It not only requires less data processing but also learns better representations.

2. We explore the use of text-driven contrastive learning on FER and AUR tasks, where the performance is improved by incorporating textual information.

3. Extensive experiments have been conducted on 3 in-the-lab datasets and 3 in-the-wild datasets. The proposed method achieves state-of-the-art performance in all 6 datasets, demonstrating the effectiveness of the proposed method.

2. Related works

2.1. Facial Expression Recognition

In order to improve the performance of facial expression recognition, various deep neural networks are designed with different insights on FER to obtain powerful representations. Researchers have conducted a series of studies [54, 30, 39, 40] aiming to decompose different attributes from facial behavioral representations and learn robust expression-related features. Another line of methods has aimed to enhance intra-class compactness and reduce inter-class compactness in feature extractions [25, 3]. Additionally, several works explore the attention mechanism on FER to obtain the discriminative features [52, 23, 27]. Furthermore, multi-task learning has been employed in various approaches, including involving facial landmark learning [11], AU recognition [20, 10], and others. Recently, due to the successful recognition performance on laboratory databases, more researchers attempt to perform FER model on in-the-wild databases, which typically contain significant label noises. As a result, addressing such noisy label issues has become a popular topic in the recent research community, as evidenced by several works [51, 42, 59, 60].

2.2. Facial Action Uniti Recognition

In recent years, deep learning has been applied to facial action unit recognition, leading to significant improvements in performance. Some works have focused on learning better facial features by emphasizing important local regions, also known as regions of interest (ROI) [63, 26, 41]. Considering the interdependency between different AUs, several works have applied graph neural networks (GCN) to model these relations [44, 45, 22, 33]. Recent works involved multiple techniques to improve the recognition accuracy, including transformer methods [18], self-supervised methods [4], and semi-supervised methods [47]. Focus on the input data, some recent work [55, 57, 28] utilized multi-modal learning methods with other modalities, such as depth images, and thermal images.

SEV-Net [56] is the first work that exploited semantic text-embedding of AU description on AUR, where the AU relationships are learned by these descriptions. The cross-modal attention mechanism between semantic embeddings and image features is used to enhance the discriminative features. **Instead of** cross-modal attention, we employ text-driven contrastive learning to enhance image-text features, which then improves performance on both FER and AUR.

2.3. Contrastive Learning

Recently, We have witnessed the potential of contrastive learning in representation learning. The principle of contrast learning is to make positive sample pairs consistent and negative sample pairs exclusive. It has been widely applied to unsupervised learning works [6, 17, 7] with outstanding success in representation learning. SupCon [19] extends contrastive learning to a fully supervised setting, named Supervised contrastive learning. In this work, data belonging to the same class are selected as positive samples, and data from different classes as negative.

Text-driven Recognition. Text-driven recognition has become an active area in both Natural Language Processing (NLP) and Computer Vision (CV). In this area, common tasks include visual question answering [1], image captioning [50], and image-text retrieval [8]. Pioneering work CLIP [38] not only demonstrates that image-text contrastive learning achieves promising performance for visual representation learning but also brings textual supervision into the classic recognition tasks in CV. Researchers have extended this vision-language model to other areas, such as object detection [16], image segmentation [21], and video action recognition [53]. Recent FaRL [64] explores this vision-language model on facial representation learning by pre-training on a variety of facial image-text pairs. However, only the image encoder was evaluated on several downstream tasks. **In contrast**, CLEF utilizes the text encoder in downstream facial behavior analysis tasks, resulting in better performance than using only the image encoder.
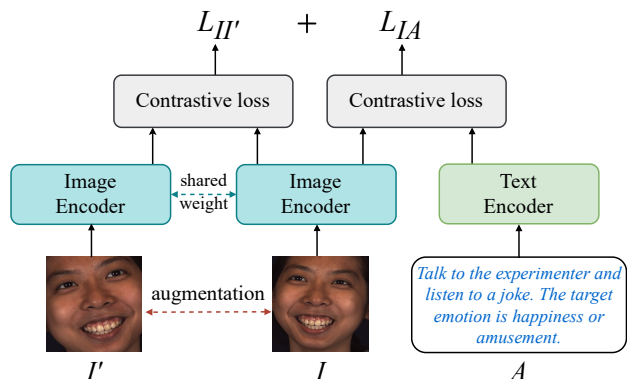

Figure 2: An overview of the architecture of the proposed CLEF in pre-training. $\mathcal{L}_{II'}$ and $\mathcal{L}_{IA}$ indicate supervised contrastive loss between images and between images and activity descriptions, respectively.

3. Methodology

Our proposed framework consists of two stages, and each is built with an image-text encoder, the same as
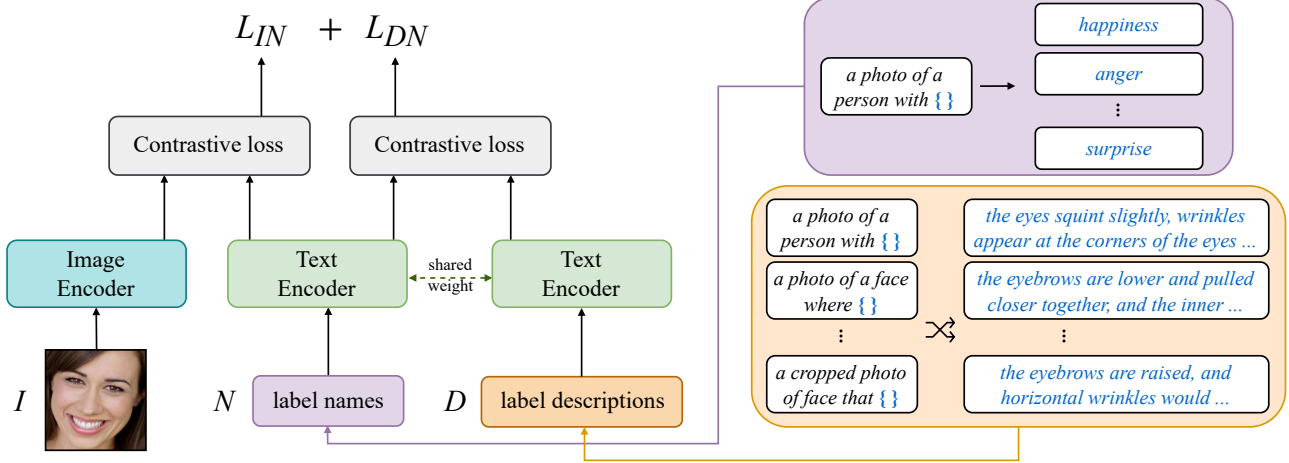
Figure 3: An overview of the architecture of the proposed CLEF in the downstream FER task. It is based on the CLIP model which consists of an image encoder and a text encoder. $\mathcal{L}_{IN}$ indicates supervised contrastive loss between images and label names. The self-supervised contrastive loss $\mathcal{L}_{DN}$ between label descriptions and label names is jointly adapted.

CLIP [38]. Figure 2 and Figure 3 show the overview of architectures in pre-training and fine-tuning respectively.

3.1. Pre-training

In pre-training, we aim to learn robust deep representation and alleviate the influence of identity variation. Therefore, we designed a contrastive learning task that pulls together features from the same activity and pushes them away from features of other activities. Sets of images, activities, and activity labels are defined by I, A, and Y^A. Given n samples in a mini-batch, we generate two augmentations, X^I, and $\tilde{X}^I$. The extracted feature representations by image encoder $g(\cdot)$ and text encoder $h(\cdot)$ are $z^I = g(x^I)$, $\tilde{z}^I = g(\tilde{x}^I)$, and $z^A = h(x^A)$ where $x^I \in X^I$, $\tilde{x}^I \in \tilde{X}^I$, and $x^A \in A$. Meanwhile, the labels are duplicated to $2n$ as $\tilde{Y}^A$. Inspired by the work [19], we consider images and their corresponding textual activity descriptions under the same activity as positive samples. We propose the cross-modal supervised contrastive loss and leverage the coarse-grained activity label to guide contrastive learning in pre-training.

Cross-modal Supervised Contrastive Loss: The contrastive loss, in the scenario of z, y pairs, at temperature ϵ is defined as:

$$\mathcal{L}_{\alpha\beta}^{\text{sup}} = -\sum_{i=1}^{n} \frac{1}{2N_i} \sum_{j \in J} \log \frac{\exp(\mathbf{z}_i^\alpha \cdot \mathbf{z}_j^{\alpha+\beta}/\epsilon)}{\sum_{k \in K} \exp(\mathbf{z}_i^\alpha \cdot \mathbf{z}_k^{\alpha+\beta}/\epsilon)} \quad (1)$$

where the symbol $(\cdot)$ denotes the inner (dot) product, $J \equiv 2N(y_i^A), j \neq i$; $K \equiv \{i\}_{i=1}^{2n}, k \neq i$, and α, β are from the extracted multi-modal feature sets. $N_i = \{j \in \{i\}_{i=1}^n : \tilde{y}_j^A = \tilde{y}_i^A\}$ contains a set of indices of positive samples with label y_i^A.

Given the $z^{I\tilde{I}} \in \{Z^I, \tilde{Z}^I\}$, $z^{IA} \in \{Z^I, Z^A\}$, the final loss in pre-training is:

$$\mathcal{L}_{pre} = \mathcal{L}_{I\tilde{I}}^{\text{sup}} + \mathcal{L}_{IA}^{\text{sup}} \quad (2)$$

$\mathcal{L}_{I\tilde{I}}^{\text{sup}}$ encourages similar representations for images from the same activity, while $\mathcal{L}_{IA}^{\text{sup}}$ encourages similar representations between images and their corresponding texts. Given that similar facial behaviors are more likely to appear within the same activity, the encoder tends to focus on capturing facial behavior features while avoiding personal attributes features, such as identity, gender, and ethnicity.

3.2. Fine-tuning

Unlike the previous work [64] which only utilized the image encoder in downstream facial analysis tasks, our approach is in the scenarios of both image and text, as we believe that text contains useful information for facial behavior analysis. Given a set of images, label names, and label descriptions. i.e., I, N, D. Similar to the pre-training, the extracted features representations are $\{z^I, z^N, z^D\}$ by image encoder $g(\cdot)$ and text encoder $h(\cdot)$. Our loss functions include self-supervised contrastive loss and supervised contrastive loss for name-description and image-label pairs respectively. The self-supervised contrastive loss, in the scenario of name-description pairs, is given as

$$\mathcal{L}_{DN} = -\frac{1}{C} \sum_{i=1}^{C} \log \frac{\exp(\mathbf{z}_i^D \cdot \mathbf{z}_j^N/\tau)}{\sum_{j=1}^{C} \exp(\mathbf{z}_i^D \cdot \mathbf{z}_j^N/\tau)} \quad (3)$$

where C is the class number, e.g., 12 AUs, and 8 expressions. τ is a learnable parameter of the temperature to scale the logits.

The supervised-contrastive learning for image-text pairs is jointly trained. We design different supervised contrastive losses based on cross-entropy and binary cross-entropy loss, as FER is a multi-class classification problem and AUR is a multi-label problem. In the FER task, the loss is defined as,

$$\mathcal{L}_{IN}^{fe} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} w_c \log \frac{\exp(\mathbf{z}_i^I \cdot \mathbf{z}_c^N/\tau)y_i^c}{\sum_{j=1}^{C}\exp(\mathbf{z}_i^I \cdot \mathbf{z}_j^N/\tau)} \quad (4)$$

where B is the batch size, w is the weight, y is the target of ground truth.

The loss in AUR is formulated as,

$$\mathcal{L}_{IN}^{au} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C}(w_c y_i^c log(\sigma(\mathbf{z}_i^I \cdot \mathbf{z}_c^N/\tau))+$$
$$(1-y_i^c)log(1-\sigma(\mathbf{z}_i^I \cdot \mathbf{z}_c^N/\tau))) \quad (5)$$

where σ is the activation function $sigmoid(\cdot)$.

Consequently, the total loss function in fine-tuning is defined as:

$$\mathcal{L}_{fine} = (\lambda \mathcal{L}_{IN}^{\alpha} + \mathcal{L}_{DN})/2.0 \quad (6)$$

where the $\alpha \in \{fe, au\}$, λ is a hyperparameter.

The loss function $\mathcal{L}_{IN}$ forces the image features to be close to the target textual features of label names. The self-supervised contrastive loss $\mathcal{L}_{DN}$ leverages the inter-class difference of semantic information to enhance the feature extracted from the text encoder. By jointly training both self-supervised and supervised contrastive components, our method learns not only the inter-class relations but also features in shared latent space across modalities, where the textual feature is unbiased to every subject identity. The algorithms' pseudocodes in PyTorch-style are shown in the Supplementary Material.

3.3. Text Prompting

Following CLIP [38], we also use prompt templates to augment the original label in our method. We only use one prompt template "a photo of a person with {label name}." for label names, e.g., "a photo of a person with *happiness*.", "a photo of a person with *inner brow raiser*.". For label descriptions, we prepare multiple prompt templates on them, e.g., "a photo shows a person that {label description}.", "a cropped photo of face that {label description}.". Considering the limited number of label descriptions in databases, instead of ensembling all prompt templates by their mean textual representation, we randomly select one prompt template in training. Similarly, activity descriptions are also randomly applied with prompt templates, e.g., "a photo of an activity that {activity description}.", "a photo of a person from an activity that {activity description}." The detail of prompting is in the Supplementary Material.

4. Experiments

The proposed CLEF is compared with the state-of-the-art methods on six popular databases for FER and AUR tasks. Furthermore, we conduct ablation studies to verify the component-wise contribution of our method.

4.1. Databases

4.1.1 AU Databases

BP4D [58] contains 41 subjects captured in laboratory environments. There are 8 activities designed to elicit different spontaneous emotions, resulting in 41×8 video clips. Expert coders select the most expressive 20 seconds of each video clip for AU coding, producing 140,000 labeled frames. Following the work [26], we split all labeled frames into subject-exclusive 3-fold with 12 AUs for both two stages.

BP4D+ [61] consists of 140 subjects with a total of 1.5 M frames in the same laboratory environments. For each subject, 20 seconds from 4 activities are annotated, resulting in 192,000 labeled frames. First, the 140 subjects are split into four-fold, following the same setting in [57]. In pre-training, we equally sample 480,000 frames from all 1.5 M frames by 10 activity categories. In fine-tuning, 12 AUs, the same as in BP4D, are selected for AU recognition.

DISFA [36] contains videos from the left view and right view of 27 subjects. In the same manner as [56], we choose 8 of 12 AUs with AU intensities higher or equal to 2 as positive samples. The model trained on BP4D is then fine-tuned to the DISFA dataset, which is following the setting in [26, 22]. F1-score is reported based on subject-exclusive 3-fold cross-validation.

4.1.2 FE Databases

AffectNet [37] is currently the largest FER dataset, including 440,000 images with manual annotation of 8 basic expressions. AffectNet-7 refers to a manually annotated set without contempt class, resulting in 283,901 and 3,500 images for training and testing respectively. AffectNet-8 includes all expression images with 287,568 training samples and 4,000 testing samples.

RAF-DB [24] is labeled by 15,000 facial images with 7 expressions, i.e., neutral, happiness, surprise, sadness, anger, disgust, and fear. Following the previous work setting [42], we choose 12,271 images for training and the remaining 3,068 for testing.

FERPlus [2] is an extended version of FER2013 [15], where 8 emotions (with contempt) are annotated. It contains 28,709 training images, 3,589 validation images, and the remaining 3,589 testing images. For a fair comparison, we report the accuracy on the test set with the same setting from [52].

Table 2: F1 scores in terms of 12 AUs on BP4D. Bold numbers indicate the best performance; bracketed numbers indicate the second best.

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg
EAC [26]	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
DSIN[9]	51.7	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	62.9	38.8	41.6	58.9
JAA-Net[41]	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
HMP-PS[46]	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	49.9	54.5	63.4
SEV-Net[56]	**58.2**	**50.4**	58.3	**81.9**	73.9	**87.8**	87.5	61.6	52.6	62.2	44.6	47.6	63.9
FAUT[18]	51.7	49.3	[61.0]	77.8	79.5	82.9	86.3	[67.6]	51.9	63.0	43.7	[56.3]	64.2
PIAP[47]	55.0	[50.3]	51.2	[80.0]	79.7	84.7	**90.1**	65.6	51.4	[63.8]	[50.5]	50.9	64.4
KSRL[4]	53.3	47.4	56.2	79.4	**80.7**	85.1	89.0	67.4	**55.9**	61.9	48.5	49.0	64.5
ANFL [33]	52.7	44.3	60.9	79.9	[80.1]	[85.3]	[89.2]	**69.4**	[55.4]	**64.4**	49.8	55.1	[65.5]
CLEF	[55.8]	46.8	**63.3**	79.5	77.6	83.6	87.8	67.3	55.2	63.5	**53.0**	**57.8**	**65.9**

Table 3: F1 scores in terms of 8 AUs on DISFA. Bold numbers indicate the best performance; bracketed number indicate the second best.

Methods	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	Avg
EAC [26]	41.5	26.4	66.4	50.7	**80.5**	**89.3**	88.9	15.6	48.5
DSIN [9]	42.4	39.0	68.4	28.6	46.8	70.8	90.4	42.2	53.6
JAA-Net[41]	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
HMP-PS[46]	38.0	45.9	65.2	50.9	50.8	76.0	93.3	**67.6**	61.0
SEV-Net[56]	55.3	53.1	61.5	[53.6]	38.2	71.6	**95.7**	41.5	58.8
FAUT[18]	46.1	48.6	**72.8**	**56.7**	50.0	72.1	90.8	55.4	61.5
PIAP[47]	50.2	51.8	[71.9]	50.6	54.5	[79.7]	[94.1]	57.2	63.8
KSRL[4]	[60.4]	[59.2]	67.5	52.7	51.5	76.1	91.3	[57.7]	[64.5]
ANFL [33]	54.6	47.1	[72.9]	**54.0**	**55.7**	76.7	91.1	53.0	63.1
CLEF	**64.3**	**61.8**	68.4	49.0	[55.2]	72.9	89.9	57.0	**64.8**

4.2. Implementation Details

Model Architecture. The proposed model consists of a text encoder $h(\cdot)$ of transformer [49] model, and an image encoder $g(\cdot)$ of ViT [12] model to learn textual features and visual features respectively. Specifically, the image encoder is ViT-B/16 with 12-layer and 768-width, resulting in 87M parameters with the input of $3 \times 224 \times 224$. The input image is first split into 14×14 patches, and then 14×14 patch embeddings are obtained by linear projection. A learnable cls token is inserted at the beginning of these embeddings, and then we can get 197 embeddings by adding position embeddings. The text encoder is a 12-layer, 512-width, and 8-head Transformer with 63M parameters. The length of the input text token is 77, and truncation or padding is performed if the input length does not match. We project features from both the image cls token and the text eos token to 512 widths as the output logits. Finally, we calculate the contrastive losses by the normalized output logits.

Pre-training setup. BP4D and BP4D+ contain the activity descriptions for our weakly-supervised contrastive learning in the first stage. Model parameters are loaded from FaRL [64] during this stage. mage augmentation techniques such as random cropping, horizontal flipping, and random rotation are used. We set the batch size by 64 and choose Adamw [32] optimizer with 0.01 weight decay. The model has been trained 5 epochs with 1 epoch warmup, followed by cosine decay [31] with a minimal learning rate of 1.e-6. The fixed temperature ϵ is set at 0.25.

Downstream tasks setup. In downstream fine-tuning, lr of 2×10^{-4} is set in BP4D, AffectNet, RAF-DB and 10^{-4} in DISFA and FER+. The model is trained with 64 batch-size and an Adamw optimizer. The evaluation metric for AUR is the averaged F1-score over all AUs, and for FER it is accuracy. Hyperparameter λ is set to 2 and its investigation is in the Supplementary Material. Other implementation details can also be found in the Supplementary Material.

4.3. Comparison with the State of the Art

4.3.1 Facial Action Unit recognition

We compare our method with several state-of-the-art works, namely EAC [26], DSIN [9], JAA-Net [41], HMP-PS [46], SEV-Net [56], FAUT [18], PIAP [47], KSRL [4] and

Table 4: F1 scores in terms of 12 AUs on BP4D+. Bold numbers indicate the best performance; bracketed numbers indicate the second best.

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg
ViT [12]	45.6	38.2	35.5	85.9	88.3	90.3	[89.0]	81.9	45.8	48.8	57.2	34.6	61.6
CLIP [38]	**49.4**	[39.7]	[38.9]	85.7	87.6	[90.6]	[89.0]	80.6	44.9	50.3	56.1	32.8	62.1
EAC [26]	43.7	39.0	14.0	85.6	87.2	90.5	88.7	**88.4**	45.7	49.0	[57.3]	**43.6**	61.1
JAA [41]	46.0	**41.3**	36.0	86.5	[88.5]	90.5	89.6	81.1	43.4	51.0	56.0	32.6	61.9
SEV-Net [56]	47.9	40.8	31.2	**86.9**	87.5	89.7	88.9	[82.6]	39.9	**55.6**	**59.4**	27.1	61.5
MFT [57]	[48.4]	37.1	34.4	85.6	**88.6**	**90.7**	88.8	81.0	**47.6**	[51.5]	55.6	36.9	[62.2]
CLEF	47.5	39.6	**40.2**	[86.5]	87.3	90.5	**89.9**	81.6	[47.0]	46.6	54.3	[41.5]	**63.1**

ANFL [33] on BP4D and DISFA datasets. Table 2 shows the comparison result on BP4D in terms of the F1-score of 12 AUs. Overall, CLEF achieves outstanding performance on the widely used database and outperforms the state-of-the-art methods in 3 AUs, namely AU4, AU23, and AU24. In addition, the quantitative results on the DISFA database are reported in Table 3, where CLEF achieves the best performance on average F1-score in terms of 8 AUs.

Table 4 shows the comparison results of our proposed method CLEF with ViT [12], CLIP [38], EAC [26], JAA [41], SEV-Net [56], and MFT [57] on the BP4D+ database. ViT and CLIP are used as the baseline methods, while the results of EAC and JAA are reported in the work of MFT. Our method performs better than the state-of-the-art methods in terms of 12 AUs, with an overall improvement of 1.4%.

4.3.2 Facial Expression Recognition

To demonstrate the generalization ability of CLEF, we also conduct experiments on the facial expression recognition task. The performance of CLEF is evaluated on the facial expression recognition task, and the results are shown in Table 5 on three commonly used in-the-wild FER databases. The state-of-the-art works are including RAN [52], SCN [51], RUL [59], DMUE [42], VTFF [34] and the most recent EAC [60]. The model is fine-tuned from the pre-trained CLEF on BP4D+. Our method achieves the best performance than other state-of-the-art methods on AffectNet-7, RAF-DB, and FER+, while slightly lower than DMUE under AffectNet-8.

4.4. Zero-shot Evaluation

We evaluate our model using zero-shot settings, where training a model with Neutral, Happiness, and Fear on AffectNet and test it by Sadness, Surprise, Disgust, and Anger on RAF-DB and FER+. See the results in the left part of Table 6. Additionally, we also evaluated FER on all expressions using a BP4D+ AUR model, shown in the right section of Table 6. Label descriptions are used to infer the model. Since the model is unaware of the unseen label names, label

Table 5: Facial expression recognition accuracies on 3 FER databases. AN-7: AffectNet-7, AN-8: AffectNet-8. Bold numbers indicate the best performance; racketed numbers indicate the second best.

Methods	AN-7	AN-8	RAF-DB	FER+
RAN [52]	59.50	-	86.90	88.55
SCN [51]	63.40	60.23	87.03	88.01
RUL [59]	61.43	-	88.98	88.75
DMUE [42]	-	**62.84**	88.76	88.64
VTFF [34]	64.80	61.85	88.14	88.81
EAC [60]	[65.32]	-	[89.99]	[89.64]
CLEF	**65.66**	[62.77]	**90.09**	**89.74**

descriptions are used in inference. Zero-shot is challenging, but CLEF outperforms the baseline FaRL obviously.

Table 6: Zero-shot results on RAF-DB and FER+

Methods	RAF-DB	FER+	RAF-DB	FER+
FaRL	16.21	25.73	13.10	21.20
CLEF	**29.14**	**34.40**	**29.47**	**24.90**

4.5. Ablation Study

To evaluate the effectiveness of each component in CLEF, we conducted ablation studies on both AUR and FER tasks. We assessed the contributions of each important component in our method, i.e., pre-trained stage with images (PI), pre-trained stage with activity texts (PA), image encoder (I), label names (N), and label description (D). It is worth noting that the text encoder is trainable only when N or D is available. Otherwise, the image feature is followed by a linear projection as the output for supervised learning. N and D are also two modalities that contribute to the contrastive losses in Equations 3, 4, 5. Table 7 shows the performance of various combinations of the components. The original CLIP and FaRL are used as the baseline methods for comparison. The result shows our model effectively learns features in the pre-training stage and leads to an improvement in recognition performance. Specifically,

Table 7: Evaluation of key components on BP4d and RAF-DB. Results indicate F1-score on BP4D, while accuracy on RAF-DB. PA: pre-trained with activity texts. PI: pre-trained with image. I: image encoder. N: label names. D: label descriptions.

Methods	PA	PI	I	N	D	BP4D	RAF-DB
CLIP			✓			63.4	87.88
CLIP			✓	✓		64.0	88.72
CLIP			✓	✓	✓	64.4	89.70
FaRL			✓			63.7	88.31
FaRL			✓	✓		64.1	88.69
FaRL			✓	✓	✓	64.6	88.78
CLEF		✓	✓			65.0	89.67
CLEF	✓	✓	✓			64.2	89.34
CLEF			✓	✓	✓	64.7	88.57
CLEF	✓	✓	✓	✓		64.9	89.57
CLEF			✓	✓	✓	64.8	89.44
CLEF	✓	✓	✓		✓	65.7	89.73
CLEF	✓	✓	✓	✓	✓	**65.9**	**90.09**

using the image encoder alone in pre-training (PI) results in some improvement (65.0 on BP4D), adding the textual activity (PA) and text encoder (ND) further improves the performance (65.9 on BP4D). Additionally, regardless of pre-training, a model with the text encoder using N and D achieves better performance than a single image encoder.

Contrastive learning between names and descriptions not only enhances the text feature from names but also expands the distinction among different descriptions. If names such as 'Disgust', and 'Fear' are isolated points in a high dimensional space, descriptions such as '...eyebrows are pulled down...' and '...eyebrows are pulled up...' are more likely to be surfaces interacted at specific points. Hence, when utilizing contrastive learning, the distance between the corresponding name-description becomes closer, while the distance between inter-descriptions is also further. The best performance is achieved by using both names and descriptions, which demonstrated that there's an optimal balance between 'distinction' and 'similarity'.

Weight-shared Text Encoder We share the weight of the text-encoder to extract the features of label names and label descriptions respectively in fine-tuning. We assume label names and label descriptions are projected in the same features space, where the distance depends on words combinations; Otherwise, the contrastive learning of relationships is limited by cross-spaces. Meanwhile, feeding the names and descriptions into different text encoders could reduce the input diversities, which can lead to performance degradation.

Such an assumption means our model not only reduces the model size but also achieves better performance than the

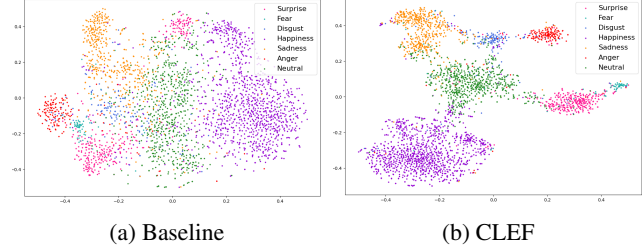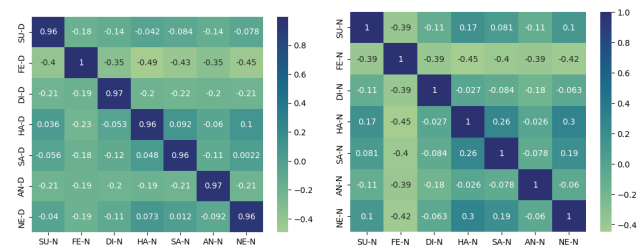(a) Baseline   (b) CLEF

Figure 4: t-SNE visualization of the expression features on RAF-DB.

(a) Cosine similarity matrix between features of label descriptions and label names trained on RAF-DB

(b) Pearson correlation coefficient matrix of label name features trained on RAF-DB

Figure 5: Visualization of similarities on RAF-DB. SU: Surprise, FE: Fear, DI: Disgust, HA: Happiness, SA: Sadness, AN: Anger, NE: Neutral. D: description, N: name

application of two separated text encoders. Hence, we continue to conduct experiments based on two individual text encoders on BP4D, achieving the average F1-score of 64.5, which is worse than using weight-shared text encoders.

4.6. Visualization

Figure 4 shows t-SNE [48] visualization of visual expression features extracted by the baseline method (FaRL) and the proposed CLEF on RAF-DB, respectively. The expression features extracted by the baseline method are not easily distinguishable from different facial expressions, while the proposed CLEF effectively enhances the separability of different classes. In particular, CLEF makes the differences among neutral, disgust, and sadness more pronounced compared to the baseline. We visualized the similarity matrix and correlation coefficient matrix of the text features on RAF-DB, which is shown in Figure 5.

We also visualize the relevancy between the image and corresponding text queries by GAE [5] in Figure 6. The image heatmap is arranged in increasing order of relevance from blue to red, while the text heatmap is arranged by increasing green intensity. Examples of 8 expressions on AffectNet and 8 AUs on BP4D can be seen in Figure 6a and Figure 6b, respectively. The text heatmap shows attention to relevant semantic words in the text, while the image
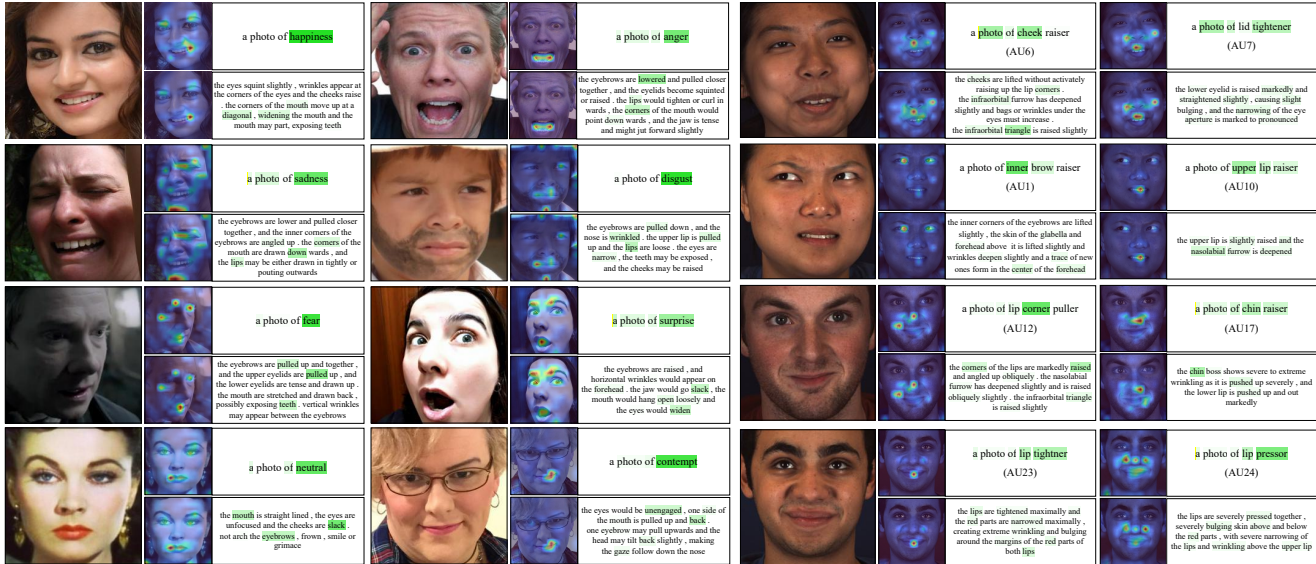
(a) Heatmap samples of 8 expressions on AffectNet

(b) Heatmap samples of 8 AUs on BP4D.

Figure 6: Visualization of the relevancy heatmap between image-name and image-description pairs using GAE [5]. In (b), the (AU id) just indicates which AU it is, but not in the textual label name.

heatmap localizes the corresponding regions of the face by querying the label name or label description. We observe that the same face regions are highlighted when querying for label names and label descriptions, indicating that the text encoder has successfully learned to extract semantic knowledge even from the label names.

5. Discussion

Advanced Paring Method. Unlike widely used object detection databases, which typically contain thousands of categories with distinct identities, facial behavior databases have limitations in terms of both the number of expression categories and identities, rendering traditional paring methods less efficient. Pairing in CLEF is activity-based, where each activity is deliberately designed to elicit a specific expression, resulting in images with expression intensity, ranging from none to onset, peak, and offset. Hence, the probability of grouping similar expressions is higher than self-supervised pairing (only the anchor itself is positive).

Easy Extension. Using texts as label names facilitates easy extension with other information. For example, intensity details can be integrated into label names by including phrases, "with low intensity", or "with high intensity".

Limitation. While our pre-trained CLEF can improve the performance of downstream tasks on various databases, it has certain limitations. 1) Our pre-training approach relies on prior knowledge of coarse-grained textual descriptions, which may not be available in some databases. We plan to address this issue in future updates by generating coarse-grained text descriptions. 2) We use a fixed prompt template for label names, and a random template for label descriptions, where the prompting is not fully explored. 3) Variations in performance across AUs can be caused by semantic descriptive writing. Thus, further investigation into description writing is necessary.

6. Conclusion

This paper has proposed a weakly-supervised text-driven contrastive method that leverages the coarse-grained activity information to learn advanced facial representations. The method minimizes intra-activity feature differences and maximizes inter-activity feature differences while disentangling the effects of subject identity features. By incorporating textual label names and descriptions, the proposed network can directly be applied to FER and AUR tasks. CLEF achieves SOTA results on 3 widely used in-the-lab databases for AUR and 3 in-the-wild databases for FER. Ablation experiments show the effectiveness of weakly-supervised contrast learning in pre-training, as well as the validity of using textual information from activity, label name, and label description. Compared to previous fine-grained pre-training methods, such as detecting landmarks, our coarse-grained approach requires less data processing while still achieving improvements.

7. Acknowledgment

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3

[2] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2016. 5

[3] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018. 3

[4] 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. 1, 2, 3, 6

[5] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, October 2021. 8, 9

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 3

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3

[9] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 309–324. Springer International Publishing, 2018. 6

[10] Zijun Cui, Tengfei Song, Yuru Wang, and Qiang Ji. Knowledge augmented deep neural networks for joint facial expression and action unit recognition. *Advances in Neural Information Processing Systems*, 33:14338–14349, 2020. 2, 3

[11] Terrance Devries, Kumar Biswaranjan, and Graham W Taylor. Multi-task learning of facial landmarks and expression. In *2014 Canadian conference on computer and robot vision*, pages 98–103. IEEE, 2014. 3

[12] 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. In *International Conference on Learning Representations*, 2020. 6, 7

[13] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 1

[14] Paul Ekman and Erika L Rosenberg, editors. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997. 1

[15] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013. 5

[16] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[18] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 1, 3, 6

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 3, 4

[20] 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. 3

[21] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 3

[22] 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, 3, 5

[23] Jing Li, Kan Jin, Dalin Zhou, Naoyuki Kubota, and Zhaojie Ju. Attention mechanism-based cnn for facial expression recognition. *Neurocomputing*, 411:340–350, 2020. 3

[24] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019. 5

[25] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 3

[26] 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 *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2017. 2, 3, 5, 6, 7

[27] Xiaotian Li, Zhihua Li, Huiyuan Yang, Geran Zhao, and Lijun Yin. Your "attention" deserves attention: A self-diversified multi-channel attention for facial action analysis. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 3

[28] Xiaotian Li, Zheng Zhang, Xiang Zhang, Taoyue Wang, Zhihua Li, Huiyuan Yang, Umur Ciftci, Qiang Ji, Jeffrey Cohn, and Lijun Yin. Disagreement matters: Exploring internal diversification for redundant attention in generic facial action analysis. *IEEE Transactions on Affective Computing*, 2023. 3

[29] Peng Liu and Lijun Yin. Spontaneous facial expression analysis based on temperature changes and head motions. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6. IEEE, 2015. 1

[30] Xiaofeng Liu, BVK Vijaya Kumar, Ping Jia, and Jane You. Hard negative generation for identity-disentangled facial expression recognition. *Pattern Recognition*, 88:1–12, 2019. 3

[31] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2016. 6

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[33] 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. 3, 6, 7

[34] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 2021. 7

[35] David Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16(4):363–368, 1992. 1

[36] S. Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 5

[37] 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. 5

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 5, 7

[39] Delian Ruan, Yan Yan, Si Chen, Jing-Hao Xue, and Hanzi Wang. Deep disturbance-disentangled learning for facial expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2833–2841, 2020. 1, 3

[40] Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, and Hanzi Wang. Feature decomposition and reconstruction learning for effective facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7660–7669, 2021. 3

[41] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018. 3, 6, 7

[42] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6248–6257, 2021. 1, 3, 5, 7

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

[44] 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*, volume 35, pages 5993–6001, 2021. 3

[45] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6267–6276, 2021. 3

[46] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6267–6276, June 2021. 6

[47] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12899–12908, 2021. 3, 6

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 6

[50] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 3

[51] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020. 3, 7

[52] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 2, 3, 5, 7

[53] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 3

[54] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2168–2177, 2018. 1, 2, 3

[55] Huiyuan Yang, Taoyue Wang, and Lijun Yin. Adaptive multimodal fusion for facial action units recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2982–2990, 2020. 3

[56] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491, 2021. 1, 2, 3, 5, 6, 7

[57] Xiang Zhang and Lijun Yin. Multi-modal learning for AU detection based on multi-head fused transformers. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021. 3, 5, 7

[58] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 2, 5

[59] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021. 3, 7

[60] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. *arXiv preprint arXiv:2207.10299*, 2022. 3, 7

[61] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Lijun Yin, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3438–3446, 2016. 5

[62] Zheng Zhang, Shuangfei Zhai, Lijun Yin, et al. Identity-based adversarial training of deep cnns for facial action unit recognition. In *BMVC*, page 226. Newcastle, 2018. 2

[63] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016. 3

[64] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 2, 3, 4, 6