RMFER: Semi-supervised Contrastive Learning for Facial Expression Recognition with Reaction Mashup Video

Yunseong Cho[1,2 †] Chanwoo Kim[1] Hoseong Cho[1] Yunhoe Ku[1] Eunseo Kim[1]
Muhammadjon Boboev[1] Joonseok Lee[3] Seungryul Baek[1]

[1]UNIST [2]SNOW Corp. [3]Seoul National University

Abstract

Facial expression recognition (FER) has greatly benefited from deep learning but still faces challenges in dataset collection due to the nuanced nature of facial expressions. In this study, we present a novel unlabeled dataset and semi-supervised contrastive learning framework that utilizes Reaction Mashup (RM) videos, a video that includes multiple individuals reacting to the same film. We created a Reaction Mashup dataset (RMset) from these videos. Our framework integrates three distinct modules: A classification module for supervised facial expression categorization, an attention module for inter-sample attention learning, and a contrastive module for attention-based contrastive learning using RMset. We utilize both the classification and attention modules for the initial training, subsequently incorporating the contrastive module to enhance the learning process. Our experiments demonstrate that our method improves feature learning and outperforms state-of-the-art models on three benchmark FER datasets. Codes are available at https://github.com/yunseongcho/RMFER.

1. Introduction

Facial expression plays an essential role in non-verbal communication [9,20]. Facial expression recognition (FER) is a task to classify facial expressions presented in an input image or a video into a predefined set of categories, *e.g.*, neutral, happiness, sadness, surprise, fear, disgust, anger, and contempt [11, 31]. Lately, the FER has attracted much attention due to its applications in marketing, education, affective computing, and other HCI applications.

With recent advances in deep learning [25] and the emergence of large-scale datasets such as AffectNet [33], RAF-DB [26], and FERPlus [1], FER approaches [27, 34, 47, 50] have advanced significantly, overcoming long-standing challenges in in-the-wild situations such as various poses, illumination, and occlusion. Despite significant progress, FER continues encountering unresolved challenges, such as mislabeling due to annotator subjectivity and the subtlety and complexity of facial expressions. As collecting new data and annotations is time-consuming and labor-intensive, self-supervised approaches, such as contrastive learning, may present a potential solution as they do not rely on fully labeled data.

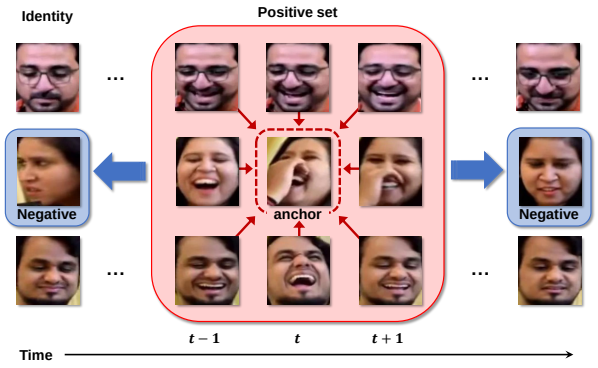


Figure 1. **Overview of RMset-based contrastive learning**. An anchor face is randomly chosen from a video, with surrounding faces forming the positive set. Faces that share the anchor's identity but are temporally distant form the negative set. During sampling, attention to the anchor improves positive and negative sets.

However, most contrastive learning approaches in FER [18, 19, 29, 32, 36] require labeled data, which can be difficult and expensive. Liu et al. [29] and Meng and Liu [32] achieved identity-invariant FER through contrastive learning with labeled datasets. Kim and Song [18] performed contrastive learning on feature transformation, but this approach cannot be easily applied to unlabeled datasets. The same authors performed contrastive learning between weak and strong emotions [19], requiring a labeled dataset for valence and arousal labels. In addition, unsupervised learning without labeled data typically performs worse than supervised learning [35, 40, 48].

Consequently, semi-supervised learning, which lever-

[†]Conducted at UNIST.

ages labeled data to its fullest extent and incorporates unlabeled data, is gaining increased attention. Recently, Li et al. [24] proposed a semi-supervised algorithm where pseudo labels were obtained for the unlabeled data and divided into two subsets based on their confidence score. Contrastive learning was applied to the subset with low confidence scores. However, their confidence score is not directly correlated to the objective of contrastive learning and is suboptimal. We proposed contrastive learning, embedding the attention learning mechanism within it; thus, our attention is better suited for contrastive learning.

To fully exploit the advantages of contrastive learning, we construct the RMset from reaction mashup (RM) videos and introduce a novel semi-supervised algorithm optimized for this dataset. The RM video records multiple viewers reacting toward a common film, as shown in Fig. 2. We assume that when viewing a specific video, people's reactions would be similar, while each person's expression in the video would change according to the time. In other words, based on an arbitrary anchor face, the facial expressions of people with different identities in nearby frames are likely to be similar. Our main assumption of contrastive learning on the RMset is illustrated in Fig. 1. However, faces with the same identity in distant frames may differ. One may think we can simply apply contrastive learning by using the above scenarios as positives and negatives. However, the naïve application of contrastive learning may not work well for the following reasons: Being in a nearby frame does not necessarily guarantee the expression similarity to the anchor face, and faces located far away from the anchor face do not always exhibit different expression to that of the anchor face as well. Therefore, we need a method to measure the expression similarity between the anchor face and positive and negative candidates. We propose inter-sample attention, which measures the expression similarity. Afterward, we sample the improved positive and negative sets from the initial sets based on the measure. Finally, we apply contrastive learning to these improved sets. In order to apply it in our scenario, the existing NT-Xent loss [43] needs to be expanded to have multiple positives.

Our contribution is summarized as follows:

- We present the RMset, which can be effectively utilized for semi-supervised contrastive learning in FER. To the best of our knowledge, this is the first large-scale unlabeled dataset for this task.
- We propose a novel semi-supervised contrastive learning framework, RMFER, that learns the inter-sample attention for contrastive learning. Based on this, contrastive learning on the unlabeled dataset (i.e., RMset) could be effectively achieved.
- From comprehensive evaluations, we verify that RMFER achieves state-of-the-art performance on three FER datasets (i.e., AffectNet, RAF-DB, and FERPlus). We also confirm that our scheme leads to better feature distribution via MDS plots.

2. Related Work

2.1. Deep Facial Expression Recognition

Various Deep-FER methods have been proposed for solving problems such as region attention [17,52,54], noisy annotation [5, 8, 30, 56, 57, 60, 61], and uncertain expression [39, 49, 59]. Although multiple issues exist in FER, we introduce identity-invariant FER and feature learning for FER, which are directly related to our study.

Identity-invariant FER. In the real-world scenario, performing FER regardless of identity is one of the important issues. Liu et al. [29] regarded face images with different facial expressions of the same identity as hard negatives and those with the same facial expressions of different identities as positives and performed deep metric learning. Yang et al. [55] employed a conditional GAN to produce a neutral image. A fully connected layer was added to the intermediate layer to fit FER, as the GAN model was fixed after training and assumed to have learned expression removal. To enable identity-invariant embedding, [58] extracts an identity feature from a fixed identity model and a face feature from a face model with the same structure. Then, they use the deviation between the two features as an expression feature. Following that, the FEC dataset [45] is used to compare contrastively to discover the expression distribution.

In our method's attention-based contrastive learning, we assume that when the anchor perceives the same expressions across different identities, they are considered positive; however, other expressions from the same identity are regarded as negative. The anchor and positives also learn to pull, while negatives learn to push. The same expression is mapped to be the same regardless of identity; hence, this is an assumption for identity-invariant FER.

Feature Learning for FER. Feature learning in FER aims to maximize intra-class similarity and inter-class separation. [4], [28] and [12] all suggest a variation of the center loss [51]. They present island loss, regularized center loss, and DDL-loss, respectively. Their function is not only to make each cluster agglomerate but also to make the distance between the clusters increase, all the same. Another center loss variant, the DACL, is presented by Farzaneh and Qi [13]. By assigning distinct weights for the center loss for each feature, DACL decreases the influence of irrelevant features on the model compared to the center loss [51], which reduces the distance for all features equally. Siqueira et al. [42] demonstrated that ESR [41] could be included in large-scale FER and that the ensemble approach of CNN allows for greater feature learning. According to Ruan et al. [37], the expression feature consists of shared and par-

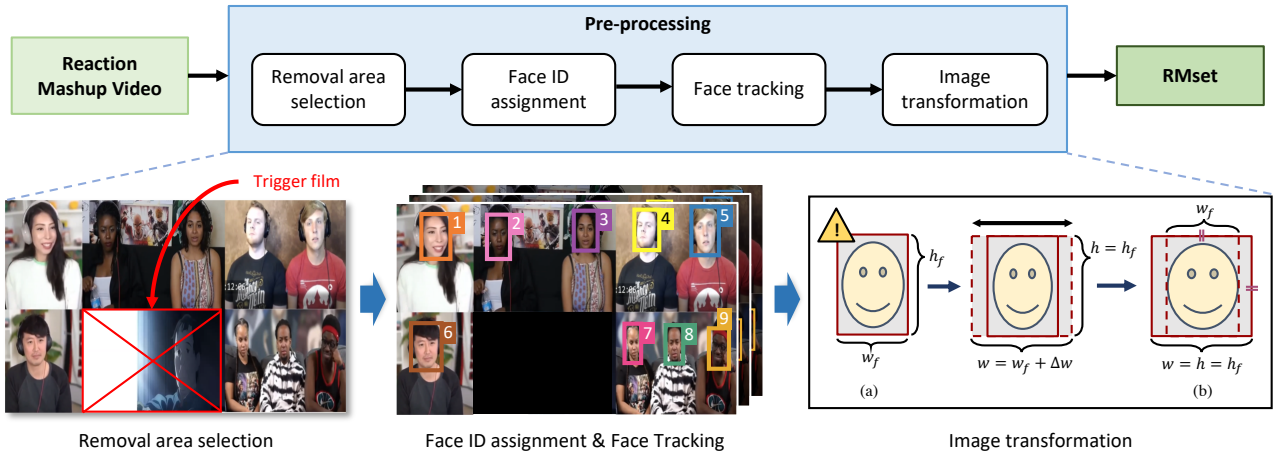


Figure 2. **The process of creating the RMset**. In the "Removal Area Selection" step, the trigger film is seen in each RM video. The trigger film can disturb the face detection algorithm (e.g. when faces appear in the film); thus, we remove pixels there before proceeding to the "Face ID assignment" step. Then, face ID is assigned to each face detected in the first frame and tracked for later frames. In the "Image transformation" step, the distorted face (a) is corrected to a normal face (b) by adjusting the bounding box size involving backgrounds.

ticular information in each expression category. They suggested the FDRL that acquires information while decomposing an expression feature into a latent feature and its reconstruction.

In inter-sample attention learning and attention-based contrastive learning, inter-class separation and intra-class similarity between features increase without explicit supervision, similar to center loss and its variations. While learning inter-sample attention, the separation between classes becomes apparent, and cohesion within the class is strengthened through contrastive learning.

2.2. Contrastive Learning

SimCLR [6] presents contrastive learning using augmentation. Under the premise that augmentation does not change semantic information, they trained the model by making positive samples, using augmentation for anchor, and negative samples, which are different. In SimCLR, the anchor and embedding of positive and negative samples were passed through the projection head, and then contrastive learning was performed using NT-Xent loss [43]. MoCo [16] views contrastive learning in terms of building dynamic dictionaries. The key sampled from the data is expressed through the encoder, and the corresponding query should be similar to the matching key and dissimilar to the others. Here, MoCo is proposed to satisfy two characteristics that a dictionary should have: large scale and consistency. Specifically, MoCo can have an extensive dictionary by storing only the key values of features, not images, using a queue, and achieving consistency using a slowly progressing encoder by momentum update. BYOL [15] proposes a method for contrastive learning without negative pairs. They presented two networks, an online network, and a target network, and let the online network predict the representation of the target network of different augmentations of the same image. This enabled contrastive learning without negative pairs while avoiding the collapse problem. Simsiam [7] revealed that the stop gradient is more decisive for avoiding the collapse problem than the momentum update of BYOL [15] and succeeded in contrastive learning only with positive pairs. Contrastive learning has been extended to video representation learning [21, 22].

3. Reaction Mashup Dataset (RMset)

In this section, we describe the methodology utilized for acquiring and generating the reaction mashup dataset (RMset), a collection of reaction videos of multiple viewers of a common film they are watching. We illustrate some examples in Fig. 2. When multiple persons are watching the same film, their expressions in the same frame are often similar. We call the film a "trigger film" as it triggers the reaction of viewers.

We compiled reaction mashup (RM) videos featuring seven basic facial expressions of individuals sourced from YouTube. These facial expressions include happiness, sadness, surprise, fear, disgust, anger, and contempt. Given that the neutral expression was consistently present across all videos, we did not collect it separately. To search videos, we used keywords like "sad reaction mashup," "try not to laugh," or "try not to be scared." Only videos with a resolution of 1080p or higher were collected to reduce image noise. In total, we collected 216 videos and contained seven expressions as uniformly as possible to avoid the data imbalance problem. In each video, about 10 to 20 people appear. In aggregate, we collected 3,141,787 frames, 3,485 people, and 45,677,989 facial images. For details on licensing, keywords used for collection, statistics, and more for RMsets, refer to the supplemental Sec. 2. We index each

face by film name, frame index, and face ID. Each person in the RM video watches a specific film, which we use to discriminate the videos. Each video comprises multiple frames; in each frame, we have multiple people's faces, with a unique ID assigned to each person.

Our pre-processing steps are as follows:

Step 1: Removal Area Selection. Each RM video contains a trigger film and the faces of people who react to it. We must remove the area where the trigger film appears since face samples are often mistakenly obtained from the region. This can be achieved by manually reading the region coordinates and removing the pixels.

Step 2: Face ID Assignment. In contrast to the film name and frame index, which are available in a video, face IDs are not trivially available. A pre-trained face detector [10] detects Human faces in the first frame. If any faces are missed, we manually label bounding boxes for them. Then, we assign an ordered index to all faces in the first frame. Faces in the later frames are tracked from the first frame's faces, as discussed in the next step.

Step 3: Face Tracking. We apply the same face detector to the later frames. As the position of the faces does not change significantly in later frames, the identity of each face can be distinguished through the position of each face. The IOU score between the current frame's detected bounding boxes and the first frame's bounding boxes are compared; based on this, the face IDs are assigned. For exceptional cases where the face detector misses any faces, or new persons appear in the middle frames, those faces cannot be reflected in our framework; we ignore them as it is rare, and there are many samples other than them.

Step 4: Image Transformation. RM videos often involve faces whose width/height ratios are different. This difference could lower the FER performance. Thus, we resize the face image patches, as illustrated on the right side of Fig. 2. Let w_f and h_f be the width and height of the face area, respectively, and the width and height of the entire image are denoted as w and h. While making $h = h_f$, we adjust w_f towards $w = w_f + \Delta w$ where Δw is $0.31 \times w_f$. As a result, the ratio of h_f/w_f becomes approximately 1.31, similar to the AffectNet dataset after the process. Upon completion of the process, we normalize the pixel values of images. The RMset will be publicly released upon acceptance.

4. The proposed method: RMFER

In FER, we aim to map an RGB image $\mathbf{x} \in X \subset \mathbb{R}^{260\times260\times3}$ to the corresponding facial expression $\mathbf{y} \in Y \subset \mathbb{R}^{E\times1}$, where E denotes the number of pre-defined expression categories of the benchmark dataset. Our RMFER framework is developed to train the feature extractor $f^{\text{Feat}} : X \rightarrow F$ using three modules, the classification module, attention module, and contrastive module, on the benchmark dataset with annotations $\{\mathbf{x}_i^{\text{b}}, \mathbf{y}_i^{\text{b}}\}_{i=1}^{N}$ as well as the collected RMset $\{\mathbf{x}_i^{\text{RM}}\}_{i=1}^{M}$ without annotation. M and N denote the number of samples in the RMset and benchmark dataset, respectively, and generally $M \gg N$. The classification network $f^{\text{FER}} \circ f^{\text{Feat}}$ is composed of the feature extraction network $f^{\text{Feat}} : X \rightarrow F$ that first maps the input image $\mathbf{x} \in X$ into the feature vector $\mathbf{f} \in F \subset \mathbb{R}^{n_{\text{dim}}\times1}$ and the fully connected layers $f^{\text{FER}} : F \rightarrow Y$ that again maps it towards the expression output $\mathbf{y} \in Y$, where n_{dim} denote the number of feature dimensions.

In the first few epochs, we enforce the feature extractor f^{Feat} to learn the mapping from the input $\mathbf{x}$ to the corresponding expression feature $\mathbf{f}$ suitable for classifier f^{FER} as well as inter-sample feature $\mathbf{f}$'s similarity using classifier f^{IAL} based on the benchmark dataset. After that, the feature space F of f^{Feat} is enriched by exploiting the RMset to learn further the feature similarities among the samples using the contrastive module. The overall operation is summarized in Fig. 3, and we describe each component in detail in the remainder of this section. Additionally, refer to supplemental Sec. 4 for the criteria for the first few epochs.

4.1. Inter-sample Attention Learning (IAL)

In this step, we involved the benchmark dataset with discretized annotations $\{\mathbf{x}_i^b, \mathbf{y}_i^b\}_{i=1}^{N}$ to learn the mapping between the input image $\mathbf{x}$ and the expression label $\mathbf{y}$, at the same time we learn the feature vector $\mathbf{f}$'s pairwise similarity $\{\mathbf{a}_{jk}\}_{j=1,k=1}^{j=B,k=B}$ within batch samples, where B denotes the number of samples in each batch.

Batch-wise Cosine Similarity-based Processing. For B samples in the same batch, we make the $B \times B$-dimensional cosine similarity matrix $\mathbf{S}$ on transformed feature vectors $\mathbf{z}_i = H(\mathbf{f}_i)$ whose entries $\mathbf{s}_{ij}$ using the cosine similarity measures. The projection head H is involved in order to project the original feature vector into the same dimensional different space similar to [6]. In our approach, self-masking is employed to compel the model to focus on inter-sample attention by setting the self-attention values to zero. Specifically, as the softmax function is applied row-wise to the cosine similarity matrix $\mathbf{S}$, we replace the diagonal elements of $\mathbf{S}$ with 10^{-6}, ensuring the self-attention, subsequent to the softmax operation, approaches a value close to zero:

$$\mathbf{s}_{ij} = \begin{cases} \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|}, & \text{if } i \neq j \\ 10^{-6}, & \text{otherwise} \end{cases} \tag{1}$$

To generate the attention feature as a summation of features excluding itself within the batch, we transform the similarity matrix $\mathbf{S}$ into matrix $\mathbf{A}$ by utilizing a softmax operation. After dividing $\mathbf{s}_{ij}$ by the scale value τ, we apply the softmax as follows:

$$\mathbf{a}_{ij} = \frac{\exp\left(\mathbf{s}_{ij}/\tau\right)}{\sum_j \exp\left(\mathbf{s}_{ij}/\tau\right)} \tag{2}$$

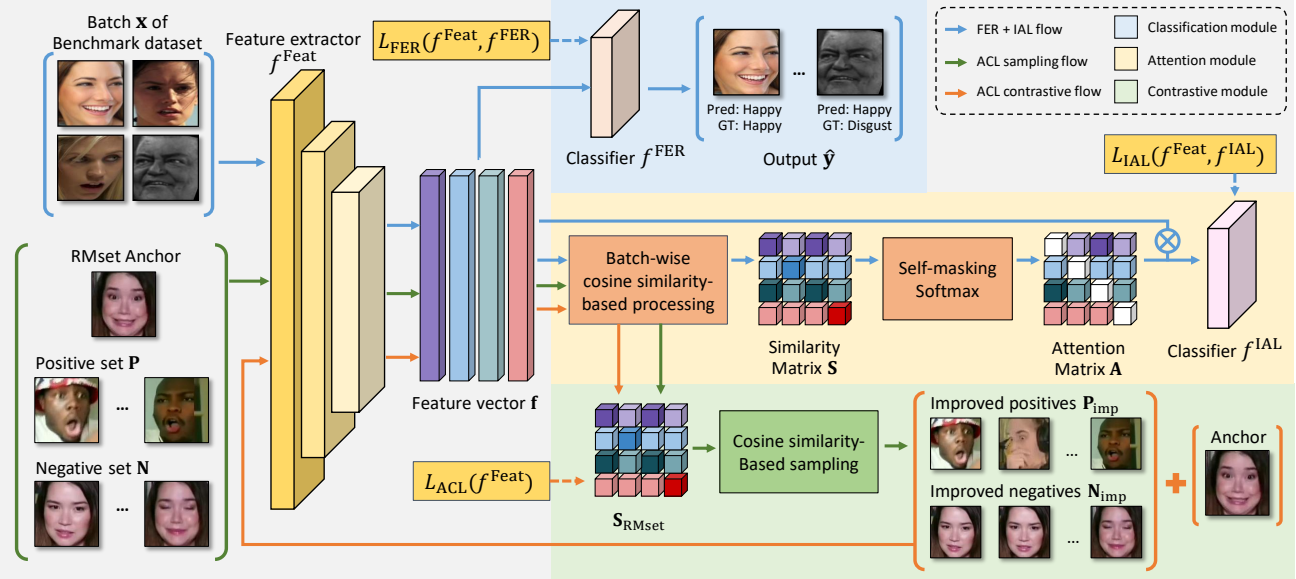


Figure 3. **Schematic diagram of our proposed framework, RMFER**. Our framework consists of three modules that share one backbone: classification module, attention module, and contrastive module. The classification module receives the RGB image $\mathbf{x}$ of the benchmark dataset as input and outputs the expression output $\mathbf{y}$. The attention module receives a batch of RGB image $\mathbf{x}$ as input and creates an attention matrix $\mathbf{A}$ through batch-wise cosine similarity-based processing and self-masking softmax. An attention feature vector $\mathbf{v}$ is created by matrix multiplication of attention matrix $\mathbf{A}$ and feature vector $\mathbf{f}$, and inter-sample attention is learned by $\mathbf{v}$ passing through f^{IAL}. The classification and attention modules are trained for the first few epochs. After that, the contrastive module is added and receives a batch of the RMset, anchor, positive set $\mathbf{P}$, and negative set $\mathbf{N}$ as inputs, which $\mathbf{P}$ and $\mathbf{N}$ are improved through attention-based sampling. The improved positives $\mathbf{P}_{\text{imp}}$, negatives $\mathbf{N}_{\text{imp}}$, and the anchor passes through f^{Feat} again to be applied to L_{ACL} (Eq. (10)).

The attention feature vector for the i-th sample $\mathbf{v}_i$ is obtained by weight-summing the feature similarity of i-th and j-th sample with the feature vector j-th sample's feature vector $\mathbf{f}_j$ as follows:

$$\mathbf{v}_i = \sum_{j=1}^{B} \mathbf{a}_{ij} \times \mathbf{f}_j \tag{3}$$

Afterward, we map the attention feature vector $\mathbf{v}_i$ towards the expression $\mathbf{y}$ via the fully connected layer $f^{\text{IAL}} : V \rightarrow Y$, which has the same architecture as f^{FER}.

Then, the attention value between i-th and j-th sample $\mathbf{a}_{ij}$ is learned in a supervised way on the benchmark dataset. Here, we assume that if i-th and j-th samples are similar, j-th sample's feature vector can be exploited to represent the attention i-th feature vector $\mathbf{v}_i$ and the attention value $\mathbf{a}_{ij}$ between i-th and j-th samples is learned higher than others. Similarly, we also assume that the attention value between i-th and j-th sample is learned lower for the dissimilar case.

Loss. The total loss of the IAL is as follows:

$$L_{\text{pre}} = L_{\text{FER}}(f^{\text{Feat}}, f^{\text{FER}}) + \lambda_1 L_{\text{IAL}}(f^{\text{Feat}}, f^{\text{IAL}}) \tag{4}$$

where λ_1 are used to balance with L_{FER} and the loss L_{FER} is defined for closing the distance between the prediction of our FER classifier $f^{\text{FER}}(\mathbf{x})$ and ground truth $\mathbf{y}$ using the cross-entropy loss as follows:

$$L_{\text{FER}}(f^{\text{Feat}}, f^{\text{FC}}) = -\sum_{i=1}^{N} \mathbf{y}_i \cdot \log\left(f^{\text{FC}}\left(f^{\text{Feat}}(\mathbf{x}_i)\right)\right) \tag{5}$$

and the loss L_{IAL} is also the cross-entropy loss defined on f^{Feat} and f^{IAL} as follows:

$$L_{\text{IAL}}(f^{\text{Feat}}, f^{\text{IAL}}) = -\sum_{i=1}^{N} \mathbf{y}_i \cdot \log\left(f^{\text{IAL}}(\hat{\mathbf{v}}_i)\right) \tag{6}$$

where $\hat{\mathbf{v}}_i$ is the differentiably outputted value from $\mathbf{x}$ using Eq. (1), (2), (3) and $\mathbf{f}_i = f^{\text{Feat}}(\mathbf{x}_i)$.

4.2. Attention-based contrastive learning (ACL)

The goal of the attention-based contrastive learning is further to learn the output space of feature extraction network f^{Feat} based on the RMset $\{\mathbf{x}_i^{\text{RM}}\}_{i=1}^{M}$. There could be a couple of ways to exploit the RMset to improve the feature distribution; we proposed to use the contrastive learning framework for that by fully exploiting priors inherent in the RMset.

Priors inherent in the RMset. Our RM videos are composed of multi-persons' faces watching the same film and the RMset is made from it. Let $\mathbf{x}_t^l$ be the face obtained from the l-th person in the t-th frame. Depending on the content

of the film multiple persons are watching, the expression value of faces in the same frame, e.g., $\mathbf{x}_t^l$ and $\mathbf{x}_t^{l'}$ might be similar to each other; while the expression value of same person's faces that exist in distance frames, e.g., $\mathbf{x}_t^l$ and $\mathbf{x}_{t'}^l$ might become dissimilar. To exploit the prior inherent in the RMset, we first construct the positive set $\mathbf{P}$ and negative set $\mathbf{N}$ for the anchor face $\mathbf{x}_{t=T}^{l=L}$ using the information:

$$\mathbf{P} = \{\mathbf{x}_t^l | l \in [1, L'], l \neq L \text{ and } t \in [T-d, T+d]\}, \quad (7)$$

$$\mathbf{N} = \{\mathbf{x}_t^{l=L} | t < T-d' \text{ or } t > T+d'\} \quad (8)$$

where L' denotes the total number of persons in the video and $d = 5$, $d' = 30$ are the hyper-parameters for deciding the nearby frames and distant frames, respectively.

Improving Positive/Negative Sets using Attention. We empirically found that our initial assumption on positive set $\mathbf{P}$ and negative set $\mathbf{N}$ is roughly valid (see Fig. 1.), while some cases violate our initial assumption: For multiple faces with different identities watching the same film frame, their expression could be different. Also, the same person's face in distant frames could have similar expressions.

To relieve this, we proposed to further sample improved positives $\mathbf{P}_{\text{imp}}$ and improved negatives $\mathbf{N}_{\text{imp}}$ from positive set $\mathbf{P}$ and negative set $\mathbf{N}$, respectively, using the cosine similarity with the anchor. The self-masking softmax that computes inter-sample attention from cosine similarity is a monotonically increasing function. Therefore, if the model is IAL trained to have high inter-sample attention between similar facial expressions and low inter-sample attention between dissimilar facial expressions, the cosine similarity between faces will have the same property. (see Sec. 4.1). Among the samples in the initial positive set $\mathbf{P}$, we select the samples whose cosine similarity to the anchor face is high. Similarly, among the samples in the initial negative set $\mathbf{N}$, we select the samples whose cosine similarity to the anchor face is low. Then, we choose upper γ ratio and lower γ ratio of samples for improved positive/negative sets $\mathbf{P}_{\text{imp}}$ and $\mathbf{N}_{\text{imp}}$ out of the initial positive/negative sets $\mathbf{P}$ and $\mathbf{N}$. We set γ as 0.1 from the 10-fold cross-validation and denote the testing accuracy for this in the Supplemental Sec. 6.

Loss. The attention-based contrastive learning loss L_{ACL} is applied to enrich f^{Feat} network based on improved positive set $\mathbf{P}_{\text{imp}}$ and improved negative set $\mathbf{N}_{\text{imp}}$ generated from the RMset. The loss L_{pre} based on the benchmark dataset is additionally involved to prevent forgetting the information. Thus, the total loss of the RMFER is composed as follows:

$$L_{\text{total}} = L_{\text{pre}} + \lambda_2 L_{\text{ACL}}(f^{\text{Feat}}) \quad (9)$$

where λ_2 are used to balance each term.

By extending the NT-Xent loss used in [6], which involves only one positive sample for the anchor, we use the loss function L_{ACL} that can involve multiple samples in both positive and negative sets as follows:

$$L_{\text{ACL}}(f^{\text{Feat}}) = -\log \frac{\sum\limits_{i \in \mathbf{P}_{\text{imp}}} \exp\left(\frac{\mathbf{s}_i}{\tau}\right)}{\sum\limits_{i \in \mathbf{P}_{\text{imp}}} \exp\left(\frac{\mathbf{s}_i}{\tau}\right) + \sum\limits_{j \in \mathbf{N}_{\text{imp}}} \exp\left(\frac{\mathbf{s}_j}{\tau}\right)} \quad (10)$$

where τ is equal to Eq. (2)'s and $\mathbf{s}$ denotes the cosine similarity of i, j between the anchors.

4.3. Testing

Testing is performed using the feature extractor f^{Feat} and classifier f^{FER}. Without adding any other modules for the testing, we use only enriched feature extractor f^{Feat} by the IAL and ACL, which are semi-supervised learning.

5. Experiment

In this section, we brief our experimental settings and current against state-of-the-art methods qualitatively and quantitatively, alongside ablation studies on various hyper-parameters.

5.1. Experimental Settings

Datasets. To validate our method, we evaluate on three FER benchmarks: RAF-DB [26], AffectNet [33], and FERPlus [1]. **AffectNet** [33] is the largest database of affect that provides eight categorical basic emotions (seven basic emotions plus *contempt*). About $280,000$ images are manually annotated with eight basic emotions and used as a training set, and a total of $4,000$ images, 500 for each emotion, are used as a validation set. The testing set is currently unpublished; thus, the validation set is used for the evaluation. In the validation set of AffectNet, the number of samples on each label is balanced. **RAF-DB** [26] consists of a total of $29,672$ images whose annotation is performed by crowdsourcing. RAF-DB provides seven basic emotions (surprised, fearful, disgusted, happy, sad, angry, and neutral). Specifically, there are $15,339$ basic emotion images, which are divided into $12,271$ training sets and $3,068$ testing sets. **FERPlus** [1] is an extension of the FER2013 [14] dataset, where ten new annotators vote for the labels. It consists of a grayscale image with a resolution of 48×48 and has eight emotion categories. The train, validation, and test sets have $28,389$, $3,553$, and $3,546$ images, respectively.

Evaluation Metrics. We used overall accuracy and average accuracy in our evaluation. Overall accuracy is a metric for the entire test set without considering class-specific performance. If the test set is imbalanced, this is an inappropriate metric to evaluate the average performance of each class. Therefore, we additionally used average accuracy, a metric that shows the average performance of each class. The average of the diagonal values of the confusion matrix represents average accuracy. For the AffectNet, due to

Method	AffectNet		RAF-DB		FERPlus	
	7 emotion	8 emotion	overall	average	overall	average
PSR [46]	63.77	60.68	88.98	80.78	89.75	69.63
ESR [42]	-	59.30	-	-	87.15	69.26
DDA loss [12]	62.34	-	86.90	79.71	-	-
LDL-ALSG [5]	59.35	-	85.53	-	-	-
SCN [49]	-	60.23	88.14	-	89.35	-
RAN [50]	-	59.50	86.90	-	89.16	-
KTN [23]	63.97	-	88.07	81.38	**90.49**	74.31
DAN [52]	65.69	62.09	89.70	85.32	-	-
FER-VT [17]	-	-	88.26	80.63	90.04	73.24
DACL [13]	65.20	-	87.78	80.44	-	-
RUL [59]	-	61.43	88.98	-	88.75	-
EfficientFace [61]	63.70	60.23	88.36	-	-	-
DMUE [39]	-	63.11	89.42	-	89.51	-
EfficientNet-b2 [38]	66.34	63.03	-	-	-	-
EAC [60]	65.32	-	89.99	-	89.64	-
SOFT [30]	66.13	62.69	90.42	-	88.60	-
Ours w/o ACL, IAL	66.13	63.32	90.58	84.03	86.62	76.30
Ours w/o ACL	66.33	63.54	90.81	84.55	86.57	76.71
Ours (full)	**66.85**	**63.82**	**91.33**	**85.59**	86.48	**77.37**

Table 1. Quantitative comparisons with state-of-the-art methods in AffectNet, RAF-DB, FERPlus datasets.

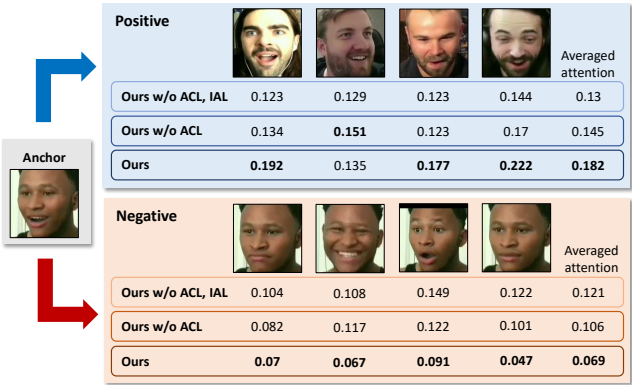


Figure 4. **Visualization of attention for the samples of the RMset**. As we involve the ACL and IAL losses, the attention becomes better: Attention needs to be higher for similar expressions while becoming less for different expressions with the same identity.

the lack of *contempt* expression in the natural world, some studies [5, 12, 13, 23, 54, 60] use only seven classes excluding *contempt* expression, while others [39,42,49,50,59] use eight classes including *contempt* expression. We evaluated ours using both to compare with all methods for AffectNet. As the validation set of AffectNet is balanced, there is no need to measure the average accuracy; while in RAF-DB and FERPlus, the test set is imbalanced, so we additionally report average accuracy in RAF-DB and FERPlus.

5.2. Results and Discussion

Table 1 presents our quantitative comparisons to state-of-the-art methods in three datasets (i.e., AffectNet, RAF-DB, and FERPlus). 'Ours w/o ACL, IAL' is a network trained only using $L_{\text{FER}}(f^{\text{Feat}}, f^{\text{FER}})$, and it is the same setting as [38]. However, we obtain slightly better performance compared to [38]. We think the reason is that we additionally consider the label imbalance in a batch during the training for IAL. 'Ours w/o ACL' is trained additionally using the loss L_{pre} (Eq. (4)) so that the feature learns inter-sample attention. Even though only inter-sample attention is added to the baseline, performance improvement occurs in almost all datasets. This means that inter-sample attention learning acts positively on expression classification. 'Ours' is trained using the total losses L_{total} (Eq. (9)). This shows the best performance in all measures in all datasets except for the 'overall accuracy' of FERPlus. This is due to the extremely unbalanced test set of FERPlus. In an unbalanced test set, overall accuracy is strongly influenced by the class with the most samples. Refer to Sec. 5 in the Supplemental for an in-depth discussion of the underlying reasons behind this phenomenon; additionally, Sec. 3 provides more experimental results on the effectiveness of the RMset and the RMFER framework, time efficiency, and hyper-parameter analysis.

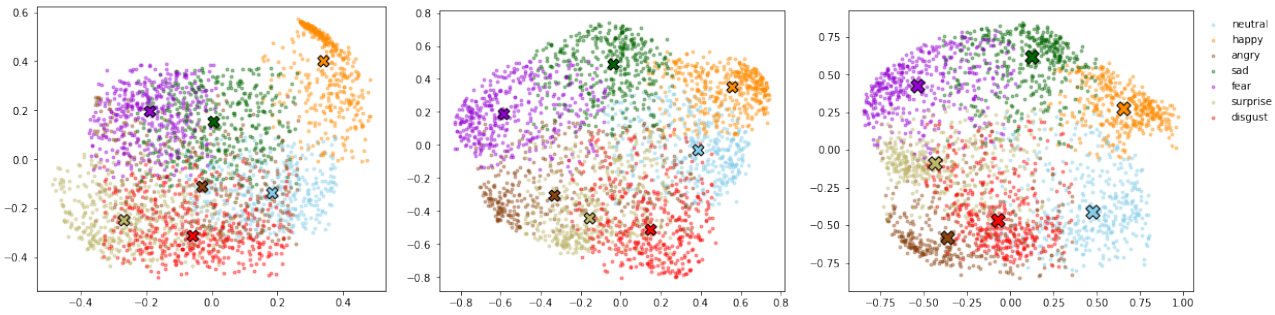
Our setting using a labeled dataset combined with unsupervised video via contrastive learning could be considered semi-supervised learning. Thus, we compare our method to several state-of-the-art semi-supervised learning methods, as in Table 2. In the experiment, labeled data come from the RAF-DB selecting only $4,000$ images: This is the same setting as [24]. Our method that involves the RMset using contrastive learning clearly outperforms Li et al. [24], which also performs contrastive learning, as mentioned in Sec. 1. We confirm the superiority of involving attention-learning mechanisms within contrastive learning.

Qualitative Results. The value of attention for each model in real-world images can be found in Fig. 4. It visualizes attention with anchors by selecting an anchor, positives, and negatives from the RMset that is not used during training. Overall, the attention given to positives is highest in 'Ours' and lowest in 'Ours w/o ACL, IAL.' Also, the attention given to negatives is highest in 'Ours w/o ACL, IAL' and lowest in 'Ours.' In other words, even if negatives and the anchor share the same identity, inter-sample attention is trained to react to the expression similarity rather than the similarity of the identity in the IAL. The ACL reinforces that effect.

Additionally, the MDS plot based on the cosine distance is provided to examine the impact of inter-sample attention learning and contrastive learning on feature learning. In Fig. 5, the classes are mixed at 'Ours w/o ACL, IAL,' but it can be observed that a boundary is established for each class when attention is learned. Moreover, each class starts to converge after contrastive learning is conducted. The mean distance from the center is gradually reduced from 0.329, 0.325, and 0.291 for (a), (b), and (c). It further supports the finding.

5.3. Ablation Study

Here, we evaluated the effectiveness of the 'self-masking softmax' we proposed in the IAL over the standard softmax



(a) Ours w/o ACL, IAL (b) Ours w/o ACL (c) Ours

Figure 5. MDS plot of (a) Ours w/o ACL, IAL, (b) Ours w/o ACL, (c) Ours in AffectNet-7. An X mark represents the center of each expression sample in the MDS plot. The mean distance from the center in the MDS plot is 0.329, 0.325, and 0.291 for (a), (b), and (c), respectively. More analysis is given in the supplemental Sec. 7.

Method	Overall (%)
MixMatch [3]	83.57
UDA [53]	83.56
ReMixMatch [2]	83.51
FixMatch [44]	83.31
Ada-CM [24]	84.42
Ours	**87.13**

Table 2. Comparison to semi-supervised learning methods

Acc(%)	RAF-DB		AffectNet	
	overall	average	7 emotion	8 emotion
Ours w/o ACL, IAL	90.58	84.03	66.13	63.32
Ours w/o ACL, SM	90.42	83.95	66.13	63.37
Ours w/o ACL	90.81	84.55	66.33	63.54
Ours w/o SM	90.94	83.97	66.53	63.59
Ours (full)	91.33	85.59	66.85	63.82

Table 3. Ablation study of self-masking softmax. The SM denotes 'self-masking'

Size	AffectNet		RAF-DB	
	7 emo	8 emo	overall	average
0%	66.33	63.54	90.81	84.55
50%	66.53	63.67	90.94	84.4
100%	**66.85**	**63.82**	**91.33**	**85.59**

Table 4. Ablation study of the size of the RMset

and the size of the RMset. In Sec. 6 of the Supplemental, we provides ablation studies on the effects of γ.

Self-masking Softmax. In Table 3, 'Ours' denotes our full model, and 'Ours w/o SM' denotes the model without the 'self-masking softmax' in the attention module. 'Ours w/o ACL, SM' is a model that does not perform the ACL, nor does it use 'self-masking softmax' in the attention module. The accuracy consistently becomes accurate as we involve the 'self-masking softmax' scheme in the attention module in both the IAL and ACL. For more results on self-masking softmax, refer to Sec. 6 of the Supplemental.

Size of the RMset. To further validate the effectiveness of the RMset, we assessed the accuracy based on the amount of the RMset used. We involved 0%, 50%, and 100% of the overall RMset during the ACL and evaluated the performance on RAF-DB and AffectNet datasets. The results are presented in Table 4. The results show that the accuracy consistently increases in all testing cases as more of the RMset is involved. As the dataset expands, the performance is expected to improve further, and it is easy to expand, as demonstrated in Supplemental Sec. 2.

6. Conclusion

We overcome the difficulties of FER, the limitation of collecting datasets due to the subjectivity of the annotator, and subtle expressions via semi-supervised contrastive learning. Specifically, we made the RMset from the RM videos and proposed a framework with inter-sample attention learning (IAL) and attention-based contrastive learning (ACL) learning that utilize the RMset. The attention module performs supervised learning on a benchmark dataset and learns the inter-sample attention. The contrastive module achieves learning the backbone using unlabeled data via contrastive learning using the learned attention. We obtained state-of-the-art results in three FER datasets.

Limitations and Future works. Our work proposed the semi-supervised learning framework exploiting large-scale unlabeled video data, RMFER. The limitation lies in the efficiency of the unlabeled dataset. Even though we proposed several modules to filter out data based on the quality, the data efficiency is still limited compared to the labeled data. The future work may need to improve it.

Acknowledgements. This work was supported by IITP grants (No. 2020-0-01336 Artificial intelligence graduate school program (UNIST) 10%; No. 2021-0-02068 AI innovation hub 10%; No. 2022-0-00264 Comprehensive video understanding and generation with knowledge-based deep logic neural network 10%) and the NRF grant (No. RS-2023-00252630 10%, 2021H1D3A2A03038607 10%, 2022R1C1C1010627 10%), all funded by the Korean government (MSIT). This work was also supported by Korea Institute of Marine Science & Technology Promotion (KIMST) funded by Ministry of Oceans and Fisheries (RS-2022-KS221674) 20% and received support from AI Center, CJ Corporation (20%).

References

[1] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM MI*, 2016.

[2] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *ICLR*, 2020.

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NIPS*, 2019.

[4] 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 *IEEE FG*, 2018.

[5] Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, and Yong Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *CVPR*, 2020.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[8] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *ICCV*, 2021.

[9] Jeffrey F Cohn and Paul Ekman. Measuring facial action. the new handbook of methods in nonverbal behavior research. *The new handbook of methods in nonverbal behavior research*, 2005.

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020.

[11] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 1971.

[12] Amir Hossein Farzaneh and Xiaojun Qi. Discriminant distribution-agnostic loss for facial expression recognition in the wild. In *CVPRW*, 2020.

[13] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *WACV*, 2021.

[14] 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 *NIPS*, 2013.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NIPS*, 2020.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[17] Qionghao Huang, Changqin Huang, Xizhe Wang, and Fan Jiang. Facial expression recognition with grid-wise attention and visual transformer. *Information Sciences*, 2021.

[18] Daeha Kim and Byung Cheol Song. Emotion-aware multi-view contrastive learning for facial emotion recognition. In *ECCV*, 2022.

[19] Dae Ha Kim and Byung Cheol Song. Contrastive adversarial learning for person independent facial emotion recognition. In *AAAI*, 2021.

[20] Irene Kotsia, Stefanos Zafeiriou, and Spiros Fotopoulos. Affective gaming: A comprehensive survey. In *CVPRW*, 2013.

[21] Hyodong Lee, Joonseok Lee, Joe Yue-Hei Ng, and Paul Natsev. Large scale video representation learning via relational graph clustering. In *CVPR*, 2020.

[22] Joonseok Lee, Sami Abu-El-Haija, Balakrishnan Varadarajan, and Apostol Natsev. Collaborative deep metric learning for video understanding. In *ACM SIGKDD*, 2018.

[23] Hangyu Li, Nannan Wang, Xinpeng Ding, Xi Yang, and Xinbo Gao. Adaptively learning facial expression representation via cf labels and distillation. *TIP*, 2021.

[24] Hangyu Li, Nannan Wang, Xi Yang, Xiaoyu Wang, and Xinbo Gao. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In *CVPR*, 2022.

[25] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective computing*, 2020.

[26] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017.

[27] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *TIP*, 2018.

[28] Zhenghao Li, Song Wu, and Guoqiang Xiao. Facial expression recognition by multi-scale cnn with regularized center loss. In *ICPR*, 2018.

[29] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *CVPRW*, 2017.

[30] Tohar Lukov, Na Zhao, Gim Hee Lee, and Ser-Nam Lim. Teaching with soft label smoothing for mitigating noisy labels in facial expressions. In *ECCV*, 2022.

[31] David Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 1992.

[32] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression recognition. In *IEEE FG*, 2017.

[33] 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*, 2017.

[34] Bowen Pan, Shangfei Wang, and Bin Xia. Occluded facial expression recognition enhanced through privileged information. In *ACM MM*, 2019.

[35] Xiangshuai Pan, Weifeng Liu, Yanjiang Wang, Xiaoping Lu, and Baodi Liu. Msl-fer: Mirrored self-supervised learning for facial expression recognition. In *ICIP*, 2022.

[36] Shuvendu Roy and Ali Etemad. Self-supervised contrastive learning of multi-view facial expressions. In *Proceedings of the International Conference on Multimodal Interaction*, 2021.

[37] Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, and Hanzi Wang. Feature decomposition and reconstruction learning for effective facial expression recognition. In *CVPR*, 2021.

[38] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 2022.

[39] 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 *CVPR*, 2021.

[40] Yuxuan Shu, Xiao Gu, Guang-Zhong Yang, and Benny P L Lo. Revisiting self-supervised contrastive learning for facial expression recognition. In *BMVC*, 2022.

[41] Henrique Siqueira, Pablo Barros, Sven Magg, and Stefan Wermter. An ensemble with shared representations based on convolutional networks for continually learning facial expressions. In *IROS*, 2018.

[42] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient facial feature learning with wide ensemble-based convolutional neural networks. In *AAAI*, 2020.

[43] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.

[44] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NIPS*, 2020.

[45] Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. In *CVPR*, 2019.

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

[47] Can Wang, Shangfei Wang, and Guang Liang. Identity-and pose-robust facial expression recognition through adversarial feature learning. In *ACM MM*, 2019.

[48] Jiahe Wang, Heyan Ding, and Shangfei Wang. Occluded facial expression recognition using self-supervised learning. In *ACCV*, 2022.

[49] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *CVPR*, 2020.

[50] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *TIP*, 2020.

[51] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

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

[53] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *NIPS*, 2020.

[54] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *ICCV*, 2021.

[55] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *CVPR*, 2018.

[56] Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, and Bo Tang. Face2exp: Combating data biases for facial expression recognition. In *CVPR*, 2022.

[57] Siwei Zhang, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Facial emotion recognition with noisy multi-task annotations. In *WACV*, 2021.

[58] Wei Zhang, Xianpeng Ji, Keyu Chen, Yu Ding, and Changjie Fan. Learning a facial expression embedding disentangled from identity. In *CVPR*, 2021.

[59] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. In *NIPS*, 2021.

[60] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: erasing attention consistency for noisy label facial expression recognition. In *ECCV*, 2022.

[61] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *AAAI*, 2021.