IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 July 2022, accepted 8 August 2022, date of publication 16 August 2022, date of current version 19 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3198658

RESEARCH ARTICLE

Learning Local Attention With Guidance Map for Pose Robust Facial Expression Recognition

SUNYOUNG CHO AND JWAJIN LEE
Defense Artificial Intelligence Center, Agency for Defense Development, Daejeon 34186, Republic of Korea
Corresponding author: Sunyoung Cho (sycho22@add.re.kr)

ABSTRACT Facial expression recognition (FER) is an extremely challenging task under unconstrained conditions. Especially, variant head poses degrade the performance dramatically due to the large variations in appearance of facial expressions. To address this problem, we propose a local attention network (LAN), which adaptively captures the important facial regions according to pose variations. The LAN emphasizes on more attentive regions while suppressing the regions not differentiated between classes. To find out attentive regions, we propose a simple yet efficient coarse-level attention guidance map annotation method in an unsupervised manner. The guidance map includes attention values for regions based on whether features are deformed by facial poses. Further, the attentive regional features obtained by our LAN and original global features are combined for pose-invariant FER. We validate our method on a controlled multiview dataset, KDEF, three popular in-the-wild datasets, RAF-DB, FERPlus, and AffectNet, and their subsets that contain images under pose variation conditions. Extensive experiments show that our LAN largely improves the performance of FER under pose variations. Our method also performs favorably against the previous methods.

INDEX TERMS Facial expression recognition, pose robust, local attention, guidance map.

I. INTRODUCTION

Facial expression is important for human-human communication as it naturally conveys emotional states and intentions. Automatic facial expression recognition (FER) is crucial in its applications across various fields such as service robots, intelligent educational systems, patient monitoring, and driver fatigue awareness. Recent significant progress on FER has been achieved with deep neural networks (DNN) and large-scale datasets in the wild.

The FER datasets in the wild have several challenges such as pose variation, illumination variation, occlusions, and motion blur. Pose variation often occurs in real-world because of either head movements or variable camera position. Therefore, it is one of the major obstacles in FER because it leads to significant changes in facial appearance. As shown in Figure 1, different expressions under variant head poses may also result in problems such as small intra-class similarities and large inter-class similarities [37], [41]. Facial expressions from the same class may have considerable differences while expressions from different classes may have similar appearances. These problems occur more often on low-quality images or micro-expressions, which are common in the wild facial images. Generally, training with these ambiguous samples may lead to over-fitting or divergence of the model, resulting in poor FER performance.

Earlier works primarily address these problems by training a model to learn features from multiple views [5], [29], or by extracting features robust to pose variations [28], [44], [36], [43]. Region-based facial features have been successfully employed to handle pose variations [36], [43], [44]. Features are extracted from grid regions [44], regions around landmark points [43], or fixed positions such as top-left, top-right, and center-down [36]. Few attempts [36], [44] have been made to improve the region-based approach by concentrating on more important regions with an attention mechanism. However, it is not easy to capture pose-robust facial regions in real-world scenes.

In this work, we propose the Local Attention Network (LAN) to adaptively capture the important facial regions

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.

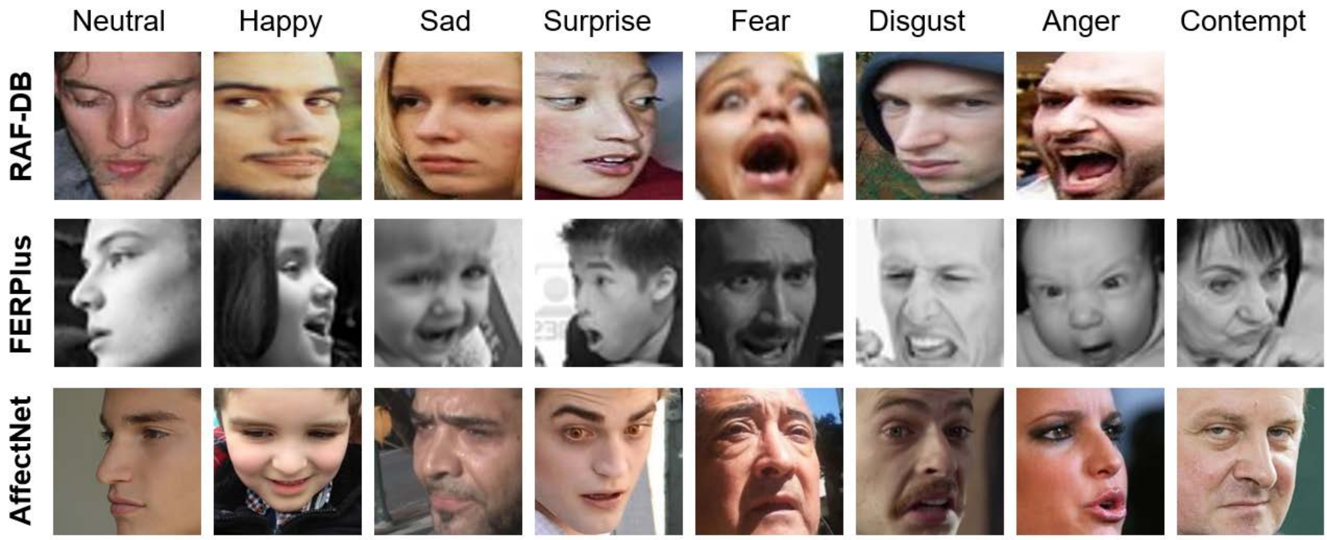

FIGURE 1. **Sample images from RAF-DB, FERPlus and AffectNet. Pose variation is one of key issues for FER in the wild.**

according to the pose variations. Our method emphasizes on more attentive regions while suppressing the regions not differentiated between different classes. Such a method maximizes intra-class similarity and minimizes inter-class similarity. Given a batch of images, a backbone convolutional neural networks (CNN) extracts the global facial features. Then the LAN learns its attention map comprising importance weights for facial regions in the input feature map. It captures the important facial regions, especially for pose-robust FER. To determine the region importance, we propose an unsupervised coarse-level attention guidance map annotation method. We apply a clustering algorithm in the neural feature space and determine the attentive cluster based on the sample distributions. Further, the global facial feature and re-weighted attentive local feature are combined for classification of facial expressions.

The main contributions of this paper can be summarized as follows:

- We propose the Local Attention Network (LAN) to learn the importance of facial regions for pose-invariant FER. LAN produces an attention map, which consists of attention weights at the pixel-level of a feature map, highlighting more attentive regions while suppressing the common regions to differentiate classes. The attention map is then used to generate the attentive regional features by combining global features.
- We present a simple yet efficient method for generating the attention guidance map annotation. Our method learns whether features are common or attentive by applying a clustering algorithm in the neural feature space. This coarse-level annotation enables a simple but efficient method to discover attentive regions for facial images under pose variation conditions.
- We extensively validate our method on a controlled multiview FER, real-world FER, and pose variation datasets. Our method improves the previous methods on KDEF, RAF-DB, FERPlus, and AffectNet datasets.

The rest of the paper is organized as follows: Section II discusses the related work. Section III introduces the proposed FER model based on LAN, and then describes our attention guidance map annotation method and loss function. Section IV describes the experiments on FER and pose variation datasets. Finally, Section V concludes the paper.

II. RELATED WORK

In this section, we mainly present methods that are related to FER, FER under variant poses, and attention mechanism.

A. FACIAL EXPRESSION RECOGNITION

Generally, a FER system primarily comprises three stages: face detection, feature extraction, and expression recognition. In the first stage, faces are located and further aligned in complex scenes using face detectors like MTCNN [42] and RetinaFace [4]. Features are then extracted from the facial images to capture the facial appearance and geometry changes caused by facial expressions. Earlier works have mainly used the texture-based local features such as SIFT [44], HOG [8], Histograms of LBP [45], Gabor wavelet coefficients [19], and geometry-based global features based on the landmark points around the eyes, mouth, and noses [29], [30]. Recent works mainly use the features that are learned from DNN. Tang [32] and Kahou *et al.* [15] won the ICML2013 FER and EmotiW2013 challenge, respectively, using deep CNN for feature extraction. Liu *et al.* [20] learn hierarchical features by constructing a deep CNN architecture based on multiple facial action units. The extracted features are fed into a supervised classifier such as support vector machines (SVMs), softmax, and logistic regression to train those expression categories. Although the recent works based on CNN show the progress on FER, they still struggle under pose variations.

B. FER UNDER REAL-WORLD VARIANT POSES

Unlike FER based on frontal view facial images, analyzing non-frontal facial images is challenging because there are several issues involving inaccurate non-frontal face alignment based on inaccurate facial landmarks, face occlusions, and facial appearance changes. Several attempts [13] have been made to address this challenging issue by using pose-robust features [5], [7], [22], [28], [43], [44], pose normalization [29], [33], [40], or pose-specific classification [14], [23]. Rudovic *et al.* [29] propose the Coupled Scaled Gaussian Process Regression (CSGPR) model for geometric head pose normalization for head-pose invariant FER. Eleftheriadis *et al.* [5] design a discriminative shared Gaussian process latent variable model (DS-GPLVM) for learning discriminative shared manifolds of facial expressions from multiple views. Zheng [44] adopts a region-based approach for facial feature extraction, and describes the relationship between the facial features of different facial views, and synthesizes the features of multiple facial views through kernel reduced-rank regression model for pose aware FER.

Recent DNNs have been also successfully applied in pose-invariant FER. Fasel [7] find that a CNN is robust to face pose and scale variations. Rifai *et al.* [28] proposes a multi-scale contractive convolutional network that learns a hierarchy of features to handle the pose variations. Zheng [44] proposes a group sparse reduced-rank regression (GSRRR) model to describe the relationship between the multi-view facial feature vectors and the corresponding expression class label vectors. The group sparsity of GSRRR automatically selects the optimal sub-regions of a face that contribute most to the expression recognition. Zhang *et al.* [43] propose to use a feature matrix consisting of the feature vectors extracted from a set of landmark points as input data of a DNN for view-invariant FER. Zhang *et al.* [41] propose a learning model for simultaneous facial image synthesis and pose-invariant FER by disentangling the expression and pose based on generative adversarial network (GAN). Wang *et al.* [36] aggregate and embed a varied number of region features extracted by a CNN into a compact representation and capture the importance of facial parts for pose robust FER. Liu *et al.* [23] propose a multi-channel pose-aware CNN to obtain a high-level fusion feature representation for different views and scales in a hierarchical way. PhaNet [22] introduces a pose-adaptive hierarchical attention network that discovers the most relevant regions to the facial expression.

C. ATTENTION MECHANISM

Attention mechanism mimics cognitive attention, enhancing the important parts of the data and ignoring the other parts. It is initially emerged from improvement over the encoder-decoder based neural machine translation system in natural language processing (NLP). Now, it is successfully used in a variety of machine learning models such as machine translation, computer vision, and speech processing. Mnih *et al.* [26] introduce a visual attention model based on recurrent neural network (RNN) for image classification tasks and demonstrate that the model outperforms a CNN. Wang *et al.* [34] propose an anchor-level attention that highlights features from the face region and relieves the false positives for occluded face detection. Yang *et al.* [38] use an attention mechanism to aggregate the features of video frames with a set of content-adaptive weights and produce a compact representation for robust face recognition in the wild. Wang *et al.* [36] propose a model comprising a feature extraction module, a self-attention module, and a relation attention module to capture the important facial regions. Liu *et al.* [22] use an attention mechanism in hierarchical scales to discover the most relevant regions to the facial expression and learn pose-invariant representations.

III. METHODOLOGY

In this section, we first present an overview of the proposed method for pose-invariant FER. We then describe the local attention network module and attention guidance map annotation module in detail. We finally present the loss function.

A. OVERVIEW

The overview of our method is shown in Figure 2. The proposed FER model is consists of two modules. The local attention network module extracts a map employing importance of each region in the feature map. It highlights more attentive facial parts and suppresses the general facial parts not differentiated between classes. The map is obtained based on the attention guidance map through a binary cross-entropy loss (BCE-Loss). The attention guidance map annotation module generates the guidance map for attentive facial regions regarding pose variations. Attentive clusters are calculated by clustering regional features in the feature maps from face images in the training set. Given the attentive clusters, the attention guidance map is easily obtained by assigning the attentive or general cluster to each facial region. Subsequently, the original and attentive features are combined for obtaining global and local attentive features for accurate FER.

B. LOCAL ATTENTION NETWORK MODULE

The inputs to the local attention network (LAN) module are the multichannel features from a CNN backbone network and the outputs are the local attentive features. The LAN module is constructed by four 3 × 3 convolutional layers followed by a sigmoid layer, as shown in Figure 2. We denote input features as $F_o \in [H \times W \times C]$, where H and W are the resolution and the C is the depth. They are passed through LAN module to produce the local attention map $F_m \in [H \times W \times 1]$. In our experiments, H and W are set to 7, and C is set to 512.

The attention map F_m is a pixel-wise map employing attentive information at pixel-level. It modulates the multichannel features F_o of facial images to obtain re-weighted attentive features $F_a \in [H \times W \times C]$. We achieve this by applying

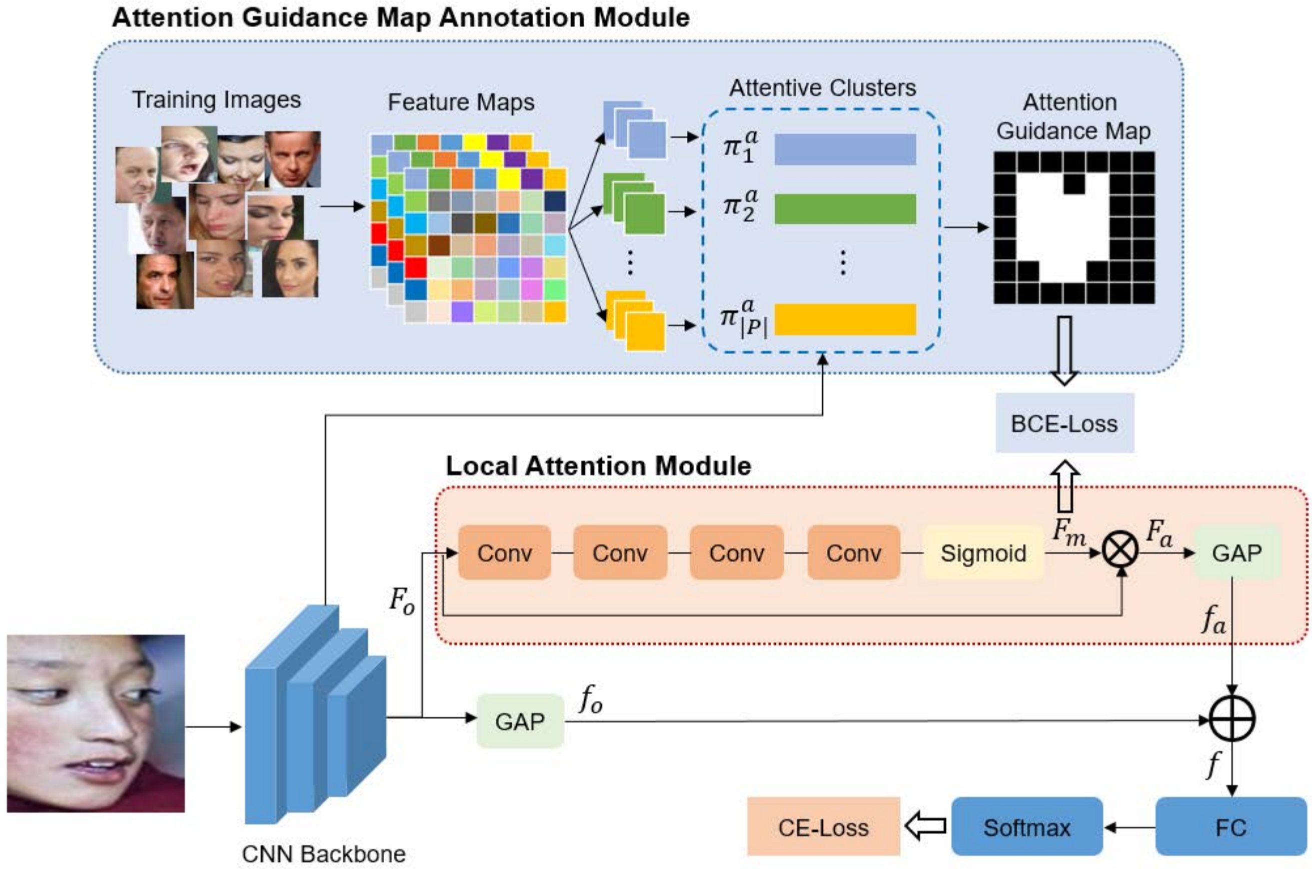

FIGURE 2. The overall architecture of the proposed method. First, input facial image goes through a backbone network for deep feature extraction. Given attentive clusters, the feature is then used to generate the guidance map for attentive regions and extract the attention map. The global CNN feature and local attention feature are combined by concatenation. The classification of expressions is performed through cross-entropy loss (CE-Loss) and binary cross-entropy loss (BCE-Loss). GAP denotes global average pooling, and FC denotes a fully-connected network.

the element-wise multiplication of every feature channel in F_o with F_m to generate F_a as:

$$F_a = F_o \odot F_m \tag{1}$$

F_a is fed into a global average pooling (GAP) layer to obtain a feature vector f_a with size of C. The original multichannel feature F_o is also fed into a GAP layer to obtain a feature vector f_o with size of C. The element-wise summation operation is further used to combine f_o and f_a as:

$$f = f_o \oplus f_a \tag{2}$$

where f represents the final features for the classification of facial expressions. These features contain global information from F_o and local attentive information from F_a simultaneously.

We conduct visualization of the proposed method through class activation mapping (CAM) [46] to compare the performance of our local attention network module with the baseline CNN backbone. Figure 3 shows the visual comparisons for input features F_o, re-weighted attentive features F_a by our LAN, and our final features f. The first row shows that the global facial regions near the mouth and nose are primarily activated by the CNN backbone network. From the middle row, we observe that our LAN captures local facial regions, such as ‘‘lip corner’’ and ‘‘wrinkles’’ compared to F_o. By combining both features, our final features become more discriminative than the input features as shown in the last row.

C. ATTENTION GUIDANCE MAP ANNOTATION MODULE

Training of attention maps for face images requires supervision indicating more important and attentive facial parts for FER in the form of dense pixel-wise segmentation. However, this is a tedious process, and the images are difficult to annotate owing to their inter-class ambiguity, especially for facial images with extreme pose variability. Another approach for learning attention maps is to design loss functions [36] that encourage the network to obtain more weights for attentive regions. However, designing such delicate loss functions is difficult without any constraint on the weights of regions.

Inspired by the observation that features are deformed by pose variations [41], we design a straightforward method to generate the attention guidance map based on facial poses. As it is challenging to infer an accurate facial pose, we instead

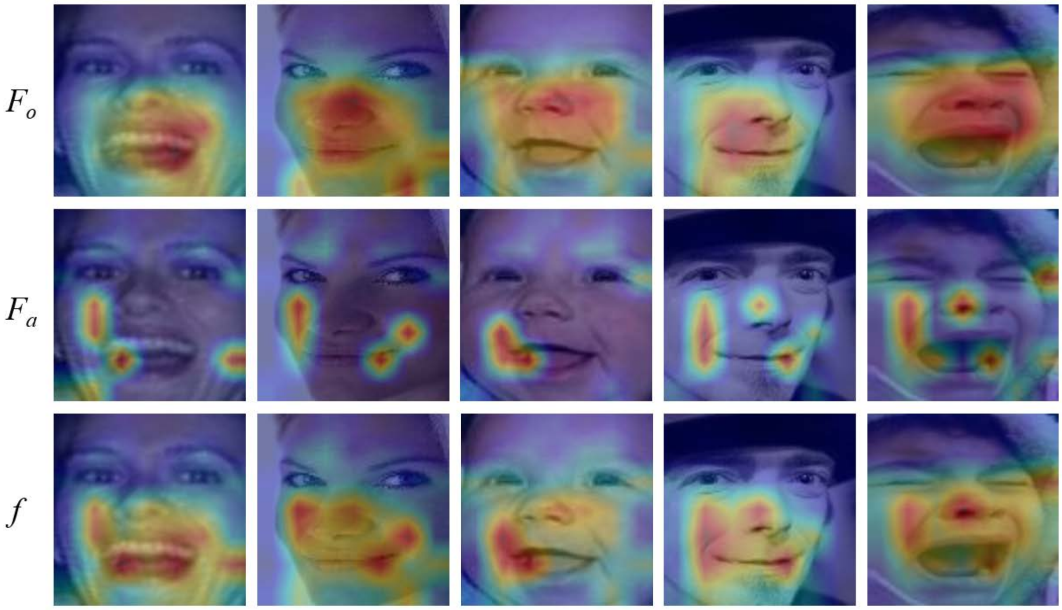

FIGURE 3. Visual comparison of class activation mapping (CAM) among input features F_o, re-weighted attentive features F_a by our local attention module, and our final features f. Note that our local attention module captures local facial regions and our final features become more discriminative by combining global and local features.

extract whether the features are deformed according to the pose in an unsupervised manner. We denote all face images in the training set as $I_1, I_2, \ldots, I_N$. For each face image I_n, we first extract the feature map $F_o^n \in [H \times W \times C]$ using a CNN backbone network. $f_p^n \in R^C$ is the feature vector at position p on the 2D lattice P of the feature map F_o^n. For each position p, we perform a clustering on all feature vectors f_p^1, $f_p^2, \ldots, f_p^N$ of feature maps of training face images. Let K be the number of clusters; π_p^k be the kth ($k = 1, 2, \cdots, K$) cluster at position p of the feature map; μ_p^k be the center of the cluster π_p^k, as defined in Eq. (3).

$$\mu_p^k = \frac{1}{\left|\pi_p^k\right|} \sum_{f_p \in \pi_p^k} f_p \tag{3}$$

where the cluster center μ_p^k at position p is calculated by averaging feature vectors f_p at position p of the feature maps.

We apply K-means clustering in the neural feature space and use 2 as the number of clusters K. In other words, we consider two types of features: one is general features for facial expressions and the other is the features deformed by the facial poses. We assume that CNN features are mostly robust but some features are severely distorted owing to severe pose variations. We determine whether features are deformed or not by the pose variations based on clustering in the neural feature space.

In case of $K = 2$, we have two clusters π_p^1 and π_p^2 at each position p. Between two clusters, we consider the number of samples of each cluster in order to determine the attentive cluster. In other words, we assume that majority of samples have more general features while a few of them have more attentive and discriminative features deformed by pose variations. Therefore, we consider the cluster including fewer samples as the attentive cluster. For two clusters π_p^1 and π_p^2, we count the number of samples and then define an attentive cluster π_p^a as follows:

$$\begin{cases} \pi_p^a = \pi_p^1 & if \left|\pi_p^1\right| < \left|\pi_p^2\right|, \\ \pi_p^a = \pi_p^2 & otherwise, \end{cases} \tag{4}$$

Our annotation method generates the attention guidance map based on the selected attentive cluster. We assign labels for each position p on the 2D lattice P of the output attention map. Given an input feature map, each feature f_p at position p is assigned to its nearest cluster based on the Euclidean distance. If f_p is assigned to the attentive cluster, we set its label as one, otherwise, we set its label as zero.

Note that we call the resulting map as the attention guidance map because we directly do not use the map as F_m. We exploit the map as a guidance to automatically extract the attention map F_m through the model as mentioned in Section III-B. Since the attention guidance map is not a ground-truth for attentive facial regions, we use it as a reference to find which facial regions should be focused for FER. To prevent the excessive consideration for attention guidance map, we reduce its impact by setting the weight in the loss function as mentioned in Section III-D.

D. LOSS FUNCTION

We present our loss function for the proposed network. The overall loss formulation L is:

$$L = L_{CE} + \alpha L_{map} \tag{5}$$

where $L_{CE}(y, y')$ is the CE-Loss between the network output y' and the true class label y. L_{map} is the loss of the proposed local attention module, formulated as a per-pixel BCE-Loss:

$$L_{map} = BCELoss(F_m(x, y), F'_m(x, y)) \tag{6}$$

where $F'_m(x, y)$ are the predictions produced by local attention module and $F_m(x, y)$ represents the attention guidance map obtained by feature clustering as described in Section III-C. As mentioned in Section III-C, we reduce the impact of the attention guidance map to produce the local attention map by setting the weight α less than 1. Therefore, we set $\alpha = 0.5$ by default.

IV. EXPERIMENTS

In this section, we demonstrate the experimental results of our method on the four public datasets and the robustness of our method under pose variations on the pose variation datasets. We then conduct ablation studies to show the effectiveness of our method.

A. DATASETS

To evaluate our method, we conduct extensive experiments on a controlled multiview FER dataset, KDEF, three popular in-the-wild FER datasets, RAF-DB [16], FERPlus [2], AffectNet [27], and three pose variation datasets, Pose-RAF-DB, Pose-FERPlus, Pose-AffectNet [36].

1) KDEF [24]

The Karolinska Directed Emotional Faces (KDEF) dataset is a multi-view facial image dataset that contains 4,900 images from 70 individuals with seven basic facial expressions (*neutral*, *happiness*, *surprise*, *sadness*, *anger*, *disgust*, *fear*). Each expression is captured from 5 different angles (−90°: full left profile, −45°: half left profile, 0°: straight, +45°: half right profile, +90°: full right profile).

2) RAF-DB [16]

The RAF-DB dataset contains 30,000 facial images with basic or compound expressions annotated by 40 trained human coders. Same as the most previous works, seven basic expressions are used together with 12,271 images for training and 3,068 images for testing.

3) FERPLUS [2]

The FERPlus dataset is extended from FER2013 [37] which is introduced in the ICML 2013 Challenges. It is a large-scale dataset collected by the Google search engine. It consists of 28,709 training images, 3,589 validation images and 3,589 test images. *Contempt* is included with seven expressions which makes eight expressions in this dataset.

4) AFFECTNET [27]

The AffectNet dataset contains about 450,000 images that are manually annotated with eight expression labels as FERPlus. It has an imbalanced training set, a balanced validation set, an imbalanced test set. We use the validation set for measurement.

TABLE 1. Comparison on the KDEF dataset.

Method	Accuracy (%)
TLCNN [47]	86.43
PhaNet [22]	86.50
MPCNN [23]	86.90
DML-Net [21]	88.20
RBFNN [25]	88.87
Our	**95.39**

5) POSE VARIATION DATASET [36]

To evaluate the performance of FER models under pose variation conditions, [36] built three subsets, Pose-RAF-DB, Pose-FERPlus, and Pose-AffectNet. Pose-RAF-DB contains 1,248 and 558 images with an angle larger than 30 and 45 degrees respectively, Pose-FERPlus contains 1,171 and 634 images with an angle larger than 30 and 45 degrees respectively, and Pose-AffectNet contains 1,949 and 985 images with an angle larger than 30 and 45 degrees respectively.

B. IMPLEMENTATION DETAILS

In all the experiments, face regions are detected and aligned by RetinaFace [4] and then resized to 224 × 224. For the backbone CNN, we use ResNet-18 [10] that is pre-trained on the MS-Celeb-1M face recognition dataset [9]. The original facial features are extracted from its last pooling layer. The input to our LAN is extracted from the layer before (Conv5_x) the last pooling layer.

The whole network is jointly optimized with CE-Loss and BCE-Loss. The ratio of the two losses is empirically set at 2:1, and its influence is evaluated in the ablation study of experiments. On all datasets, parameters are optimized via SGD optimizer with an initial learning rate of 0.1 and a mini-batch size of 64.

C. EVALUATION ON A CONTROLLED MULTIVIEW FER DATASET

In order to show that the proposed method is robust to the head pose variation, we compare our method on a multiview FER dataset.

1) COMPARISON ON KDEF

We compare the proposed method with previous methods on KDEF dataset in Table 1. The results are achieved by 10-fold cross-validation on the dataset. TLCNN [47] proposes two automatic selection schemes on high-level feature maps of CNN on generic images for FER. RBFNN [25] proposes a method to concatenate spatial pyramid Zernike moments based shape features and Law's texture features. Compared to pose-aware methods such as PhaNet [22], MPCNN [23], and DML-Net [21], our method largely improves FER accuracy. Overall, our method outperforms these previous methods with **95.39%**.

TABLE 2. Comparison on the RAF-DB dataset.

Method	Accuracy (%)
Resnet-18 [10]	74.87
DLP-CNN [16]	80.89
pACNN [18]	83.27
IPA2LT [39]	86.77
Separate-Loss [17]	86.38
gACNN [18]	85.07
RAN [36]	86.90
LDL-ALSG [3]	85.53
DDA-Loss [6]	86.90
SCN [35]	87.03
Our	**87.09**

TABLE 3. Comparison on the FERPlus dataset.

Method	Accuracy (%)
Resnet-18 [10]	83.52
PLD [2]	85.10
ResNet+VGG [12]	87.40
RAN [36]	88.55
SeNet50 [1]	**88.80**
SCN [35]	88.01
Our	88.45

TABLE 4. Comparison on the AffectNet dataset. $^+$ Oversampling is used since affectnet is imbalanced.

Method	Accuracy (%)
Resnet-18 [10]	54.37
Upsample [27]	47.00
DLP-CNN [16]	54.47
pACNN [18]	55.33
Weighted loss [27]	58.00
gACNN [18]	58.78
IPA2LT [39]	55.71
Separate-Loss [17]	58.89
RAN$^+$ [36]	59.50
ESR-9 [31]	59.30
SCN [35]	60.23
Our	**60.88**

TABLE 5. Comparison on the pose variation dataset.

Datasets	Methods	Accuracy (%)
Pose-RAF-DB-30	Resnet-18 [10]	84.04
	RAN [36]	86.74
	SCN [35]	86.85
	Our	**87.33**
Pose-RAF-DB-45	Resnet-18 [10]	83.15
	RAN [36]	85.20
	SCN [35]	85.30
	Our	**85.66**
Pose-FERPlus-30	Resnet-18 [10]	78.11
	RAN [36]	82.23
	SCN [35]	85.73
	Our	**87.18**
Pose-FERPlus-45	Resnet-18 [10]	75.50
	RAN [36]	80.40
	SCN [35]	82.78
	Our	**85.31**
Pose-AffectNet-30	Resnet-18 [10]	50.10
	RAN [36]	53.90
	SCN [35]	56.17
	Our	**58.44**
Pose-AffectNet-45	Resnet-18 [10]	48.50
	RAN [36]	53.19
	SCN [35]	53.81
	Our	**58.09**

D. EVALUATION ON IN-THE-WILD FER DATASETS

In this section, we compare the proposed method to several previous methods on RAF-DB, FERPlus, and AffectNet datasets.

1) COMPARISON ON RAF-DB

We compare our method to several methods on the RAF-DB dataset in Table 2. DLP-CNN [16] improves the discriminability of features by maximizing the inter-class scatters and retaining the locality closeness. IPA2LT [39] introduces the latent ground-truth for learning with inconsistent annotations across different FER datasets. Separate-Loss [17] proposes the separate loss that consists of intra-class loss and inter-class loss to learn discriminative features. pACNN and gACNN [18] leverages patch-based and global-local-based networks. RAN [36] proposes a region attention network that learns attention weights for each facial region. LDL-ALSG [3] proposes the label distribution learning on auxiliary label space graphs to address the annotation inconsistency and improve the FER performance. DDA-Loss [6] presents discriminant distribution-agnostic loss to increase feature discriminability in the embedding space and solve the class imbalance problem. SCN [35] proposes self-attention and relabeling mechanisms to suppress the uncertainties of FER annotations. As shown in Table 2, our method outperforms these previous methods with **87.09%** on RAF-DB dataset.

2) COMPARISON ON FERPLUS

For FERPlus dataset, we also compare our method to several methods, as shown in Table 3. PLD [2] uses probabilistic label drawing to obtain label distribution and train a FER model. ResNet+VGG [12] uses linear SVM with concatenated features of ResNet-18 and VGG-16 models. SeNet50 [1] improves the performance using squeeze-and-excitation architecture [11] based on ResNet-50 for FER. Compared with our method, we achieve a comparable result of 88.45% using a network shallower than SeNet50.

3) COMPARISON ON AFFECTNET

Table 4 shows the comparison on AffectNet dataset. As the training set of AffectNet dataset is imbalanced, [27] applies up-sampling and weighted-loss. ESR-9 [31] ensembles with shared representations based on CNN to reduce the residual generalization error. Our method outperforms these recent methods with **60.88%**.

E. EVALUATION UNDER POSE VARIATION

To evaluate the proposed method under pose variation conditions, we conduct experiments on Pose-RAF-DB,

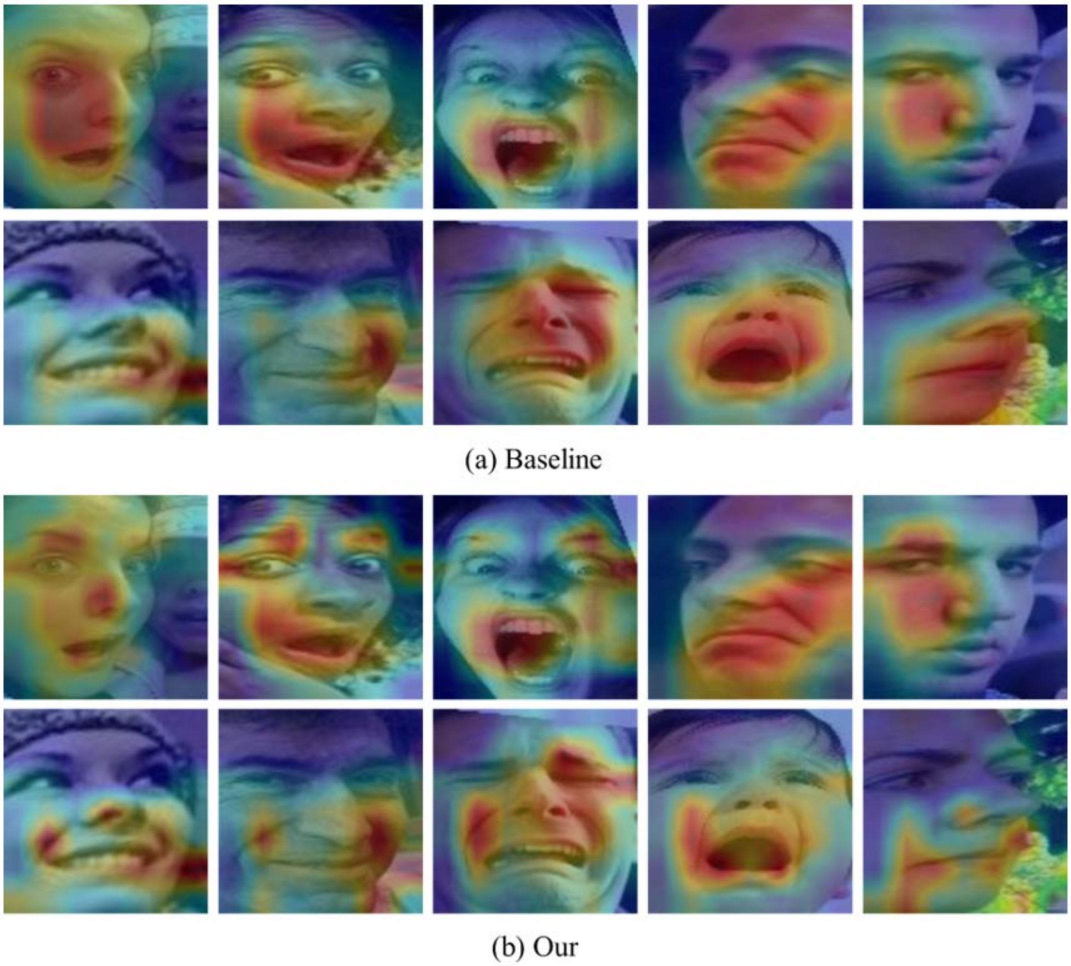

FIGURE 4. **The comparison of class activation mapping (CAM) between the baseline and our method. The images are from the Pose-RAF-DB dataset.**

Pose-FERPlus, and Pose-AffectNet. As shown in Table 5, the performance of our method is superior on both datasets comprising images with angles larger than 30 and 45 degrees. On Pose-FERPlus and Pose-AffectNet, our method outperforms the baselines with a large gap. Compared to the RAN method, the gains are 4.95% and 4.54% with pose larger than 30 degree, and 4.91% and 4.90% with pose larger than 45 degree. Overall, these results demonstrate the robustness of our method on variant pose FER datasets.

Figure 4 shows the visualization results of CAM between the baseline and our method. We fine-tune ResNet-18 on the FER datasets as a baseline. The images are obtained from the Pose-RAF-DB dataset. We can observe that the baseline mainly captures regions near the mouth and cheek. However, several expressions such as *Surprise*, *Anger*, and *Sadness* share a similar mouth. Cheek regions also cause confusion because some samples from expressions such as *Happiness*, *Sadness*, *Anger* have similar cheek wrinkles. Compared to the baseline, our method captures more important local regions to distinguish expressions from different classes. As shown in Figure 4, our method exploits facial regions near lip corner, eye corner, eyebrows, and facial wrinkles, which are more critical clues to distinguish expressions.

We present the confusion matrices of our method on pose variation datasets, as shown in Figure 5. On RAF-DB dataset, we find that *Fear* is often misclassified to *Surprise* because the two classes appear to have similar facial features, such as ''opening a mouth'' or ''bigger eyes.'' On FERPlus dataset, *Contempt* is mostly misclassified to *Neutral* or *Sadness*. This is because some samples of *Contempt* contain moderate emotions or similar facial expressions to *Sadness*. *Sadness* and *Anger* are also often misclassified to *Neutral* because restrained expressions of *Sadness* or *Anger* are confusing to *Neutral*. Similar to RAF-DB dataset, *Fear* is often misclassified to *Surprise* on FERPlus dataset. On AffectNet dataset, we find that *Neutral* is often misclassified to various classes such as *Sadness*, *Surprise*, *Anger*, or *Contempt* because some retrained or micro expressions of those classes cause confusion. Overall, the above misclassification cases also occur in AffectNet dataset.

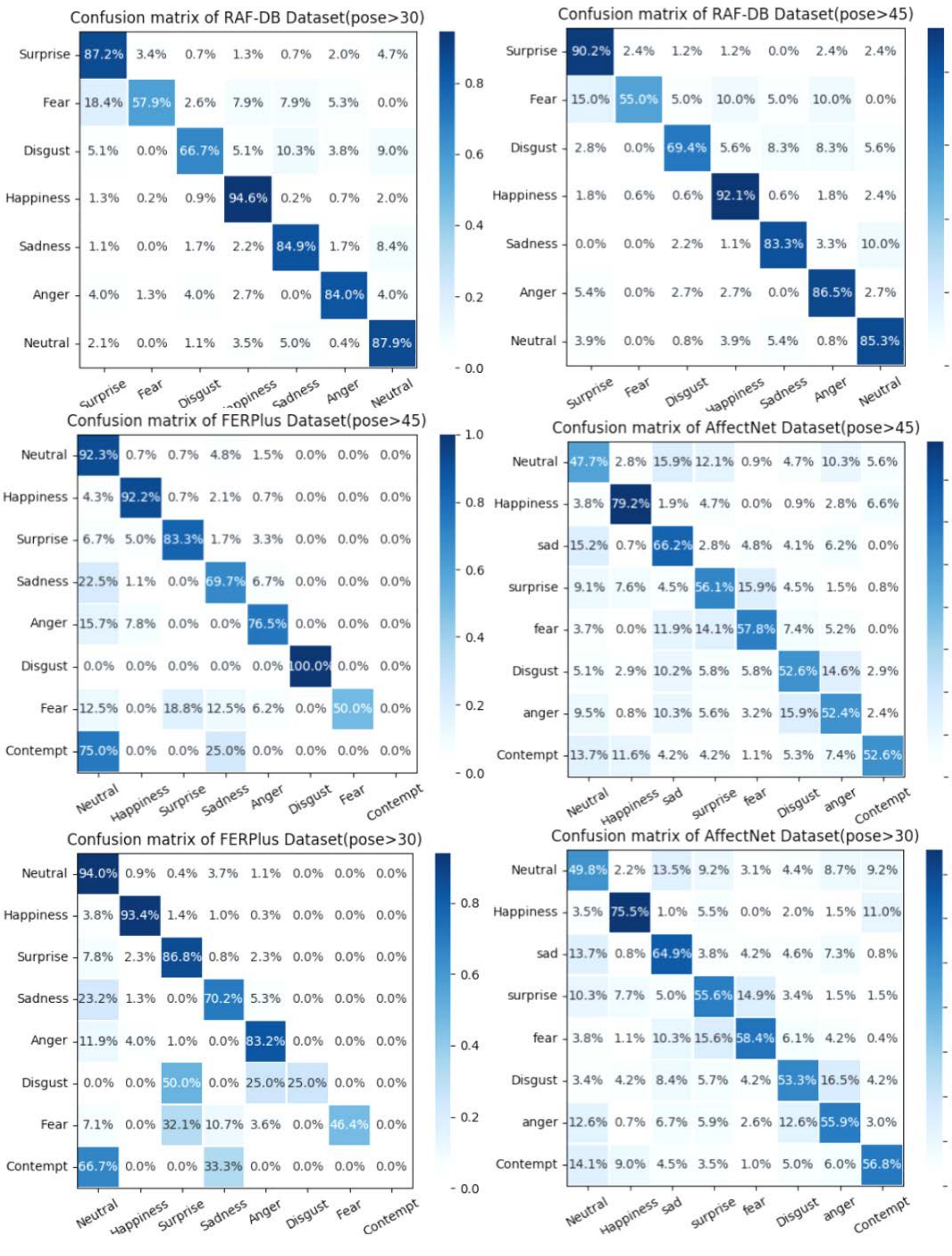

FIGURE 5. **The confusion matrices of our method on pose variation test datasets.**

In Figure 6, we provide examples correctly classified and misclassified by our method from RAF-DB, FERPlus, and AffectNet datasets. Generally, we find that micro-expression, extreme facial poses, and confusing facial expressions are misclassified.

F. ABLATION STUDY

We conduct ablation studies on the validation set of AffectNet.

1) EVALUATION OF THE α

α is the ratio of considering L_{map} loss. In other words, this ratio controls the effect of the attention guidance map to produce the local attention map. We evaluate the parameter α values of our loss function in Figure 7. We study different ratios from 0 to 1.0 on AffectNet dataset. Our default ratio that achieves the best performance is 0.5. Small α value degrades the ability of our LAN since it reflects less attentive features. Large α value leads to reduce

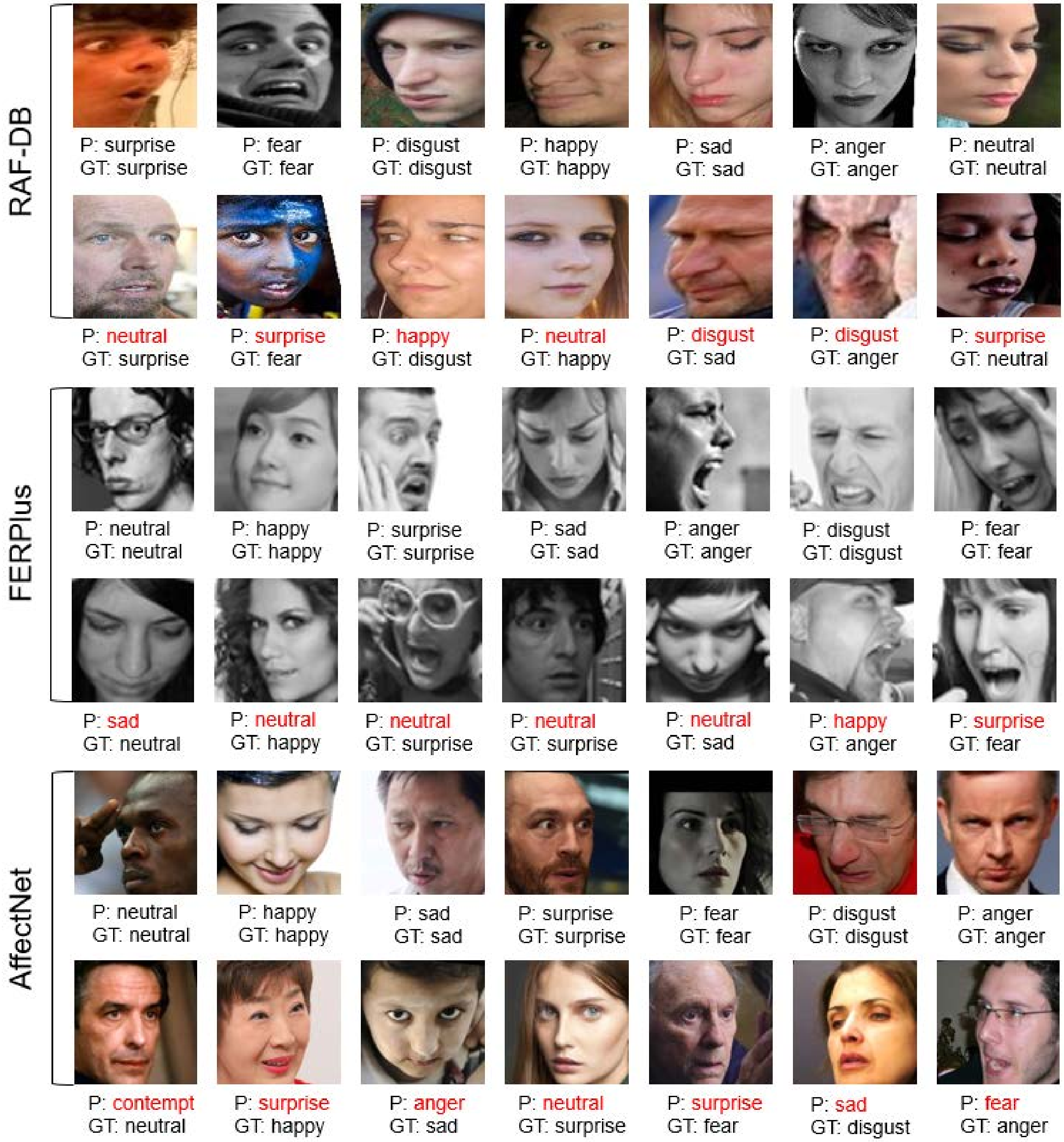

FIGURE 6. Selected examples of FER results on pose variation datasets. P: predictions by our method, GT: ground truth. Red text denotes misclassification.

ability of classification by over-consideration of the attention map.

2) EVALUATION OF DIFFERENT FUSION SCHEMES

We also conduct experiment to evaluate different feature fusion strategies to combine global and local features. We consider three popular feature fusion methods, such as feature concatenation, feature averaging, and feature addition. Table 6 shows the evaluation of different feature fusion schemes on AffectNet dataset. Our feature addition scheme achieves the best performance. For the rest of the methods, feature concatenation is superior to feature averaging.

TABLE 6. Evaluation of different feature fusion schemes on AffectNet dataset.

Fusion method	Accuracy (%)
Feature concatenation	60.20
Feature averaging	59.40
Feature addition	**60.88**

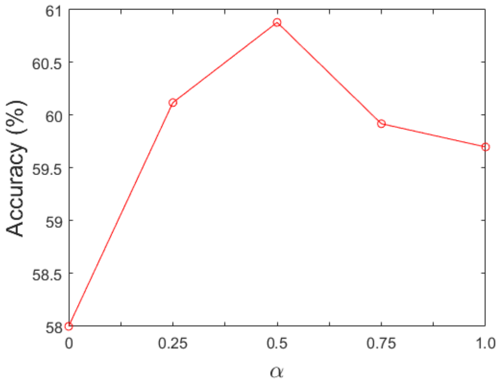

FIGURE 7. The evaluation of different α values on AffectNet dataset.

V. CONCLUSION

This paper presents a Local Attention Network (LAN) to adaptively capture the important facial regions according to the pose variations. LAN produces facial attention maps using coarse-level feature clustering information. The resulting facial local attention maps modulate the facial features by emphasizing on more attentive regions while suppressing the regions deformed by pose variations. To determine the region importance, we employ coarse-level pose information as a guidance by clustering the features and selecting the attentive cluster based on the sample distributions. Extensive experiments on four public datasets show that our LAN achieves previous results and can handle pose variations in the real-world.

Although our results show the robustness of our model under pose variations, our model still fails to images with extreme poses. In the future work, we will extend our model to be able to recognize for extreme facial poses. We will also conduct a more extensive evaluation with a variety of different facial angles and extreme facial poses.

REFERENCES

[1] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 292–301.

[2] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 279–283.

[3] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13981–13990.

[4] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5202–5211.

[5] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 189–204, Jan. 2015.

[6] A. H. Farzaneh and X. Qi, "Discriminant distribution-agnostic loss for facial expression recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1631–1639.

[7] B. Fasel, "Robust face analysis using convolutional neural networks," in *Proc. Int. Conf. Pattern Recognit.*, 2002, pp. 40–43.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[9] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 87–102.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[11] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Apr. 2020.

[12] C. Huang, "Combining convolutional neural networks for emotion recognition," in *Proc. IEEE MIT Undergraduate Res. Technol. Conf. (URTC)*, Nov. 2017, pp. 1–4.

[13] M. Jampour and M. Javidi, "Multiview facial expression recognition, a survey," *IEEE Trans. Affect. Comput.*, early access, Jun. 21, 2022, doi: 10.1109/TAFFC.2022.3184995.

[14] J. He, D. Li, B. Yang, S. Cao, B. Sun, and L. Yu, "Multi view facial action unit detection based on CNN and BLSTM-RNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 848–853.

[15] S. E. Kanou *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. 15th ACM Int. Conf. Multimodal Interact.*, 2013, pp. 543–550.

[16] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2584–2593.

[17] Y. Li, Y. Lu, J. Li, and G. Lu, "Separate loss for basic and compound facial expression recognition in the wild," in *Proc. Asian Conf. Mach. Learn.*, 2019, pp. 897–911.

[18] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[19] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2002, pp. 229–234.

[20] M. Liu, S. Li, S. Shan, and X. Chen, "AU-inspired deep networks for facial expression feature learning," *Neurocomput.*, vol. 159, pp. 126–136, Jul. 2015.

[21] Y. Liu, W. Dai, F. Fang, Y. Chen, R. Huang, R. Wang, and B. Wan, "Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition," *Inf. Sci.*, vol. 578, pp. 195–213, Nov. 2021.

[22] Y. Liu, J. Peng, J. Zeng, and S. Shan, "Pose-adaptive hierarchical attention network for facial expression recognition," 2019, *arXiv:1905.10059*.

[23] Y. Liu, J. Zeng, S. Shan, and Z. Zheng, "Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 458–465.

[24] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska directed emotional faces (KDEF)," Dept. Clin. Neurosci., CD ROM, Psychol. Sect., Karolinska Institutet, Solna, Sweden, Tech. Rep., 1998.

[25] V. G. V. Mahesh, C. Chen, V. Rajangam, A. N. J. Raj, and P. T. Krishnan, "Shape and texture aware facial expression recognition using spatial pyramid Zernike moments and law's textures feature set," *IEEE Access*, vol. 9, pp. 52509–52522, 2021.

[26] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[27] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan./Mar. 2019.

[28] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 808–822.

[29] O. Rudovic, M. Pantic, and I. Patras, "Coupled Gaussian processes for pose-invariant facial expression recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1357–1369, Jun. 2013.

[30] O. Rudovic, I. Patras, and M. Pantic, "Regression-based multi-view facial expression recognition," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 4121–4124.

[31] H. Siqueira, S. Magg, and S. Wermter, ''Efficient facial feature learning with wide ensemble-based convolutional neural networks,'' in *Proc. AAAI*, 2020, pp. 5800–5809.

[32] Y. Tang, ''Deep learning using linear support vector machines,'' in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1–6.

[33] Z. Tösér, A. László Jeni, A. Lörincz, and F. J. Cohn, ''Deep learning for facial action unit detection under large head poses,'' in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 359–371.

[34] J. Wang, Y. Yuan, and G. Yu, ''Face attention network: An effective face detector for the occluded faces,'' 2017, *arXiv:1711.07246*.

[35] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, ''Suppressing uncertainties for large-scale facial expression recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6896–6905.

[36] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, ''Region attention networks for pose and occlusion robust facial expression recognition,'' *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.

[37] F. Xue, Q. Wang, and G. Guo, ''TransFER: Learning relation-aware facial expression representations with transformers,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3601–3610.

[38] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, ''Neural aggregation network for video face recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4362–4371.

[39] J. Zeng, S. Shan, and X. Chen, ''Facial expression recognition with inconsistently annotated datasets,'' in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 222–237.

[40] F. Zhang, Q. Mao, X. Shen, Y. Zhan, and M. Dong, ''Spatially coherent feature learning for pose-invariant facial expression recognition,'' *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 1–19, Mar. 2018.

[41] F. Zhang, T. Zhang, Q. Mao, and C. Xu, ''Joint pose and expression modeling for facial expression recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3359–3368.

[42] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, ''Joint face detection and alignment using multitask cascaded convolutional networks,'' *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[43] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, ''A deep neural network-driven feature learning method for multi-view facial expression recognition,'' *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.

[44] W. Zheng, ''Multi-view facial expression recognition based on group sparse reduced-rank regression,'' *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 71–85, Jan. 2014.

[45] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, ''Learning active facial patches for expression analysis,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2562–2569.

[46] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, ''Learning deep features for discriminative localization,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[47] Y. Zhou and B. E. Shi, ''Action unit selective feature maps in deep networks for facial expression recognition,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2031–2038.

SUNYOUNG CHO received the Ph.D. degree in computer science from Yonsei University, South Korea, in 2014. From 2015 to 2016, she was a Postdoctoral Fellow with the Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA. She is currently a Senior Researcher with the Agency for Defense Development, South Korea. Her research interests include facial expression recognition, emotion recognition, and object detection.

JWAJIN LEE received the B.S. degree in electrical and computer engineering (ECE) from the Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea, in 2020. Since 2020, he has been a Research Officer with the Defense Artificial Intelligence Center, Agency for Defense Development, Daejeon. His research interests include face recognition, facial expression recognition, and artificial intelligence.