TransFER: Learning Relation-aware Facial Expression Representations with Transformers

Fanglei Xue[1,2 *], Qiangchang Wang[3 *], Guodong Guo[4,5,3 †]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China
[3]West Virginia University, Morgantown, USA
[4]Institute of Deep Learning, Baidu Research, Beijing, China
[5]National Engineering Laboratory for Deep Learning Technology and Application, Beijing, China
xuefanglei19@mails.ucas.ac.cn, qw0007@mix.wvu.edu, guoguodong01@baidu.com

Abstract

Facial expression recognition (FER) has received increasing interest in computer vision. We propose the TransFER model which can learn rich relation-aware local representations. It mainly consists of three components: Multi-Attention Dropping (MAD), ViT-FER, and Multi-head Self-Attention Dropping (MSAD). First, local patches play an important role in distinguishing various expressions, however, few existing works can locate discriminative and diverse local patches. This can cause serious problems when some patches are invisible due to pose variations or viewpoint changes. To address this issue, the MAD is proposed to randomly drop an attention map. Consequently, models are pushed to explore diverse local patches adaptively. Second, to build rich relations between different local patches, the Vision Transformers (ViT) are used in FER, called ViT-FER. Since the global scope is used to reinforce each local patch, a better representation is obtained to boost the FER performance. Thirdly, the multi-head self-attention allows ViT to jointly attend to features from different information subspaces at different positions. Given no explicit guidance, however, multiple self-attentions may extract similar relations. To address this, the MSAD is proposed to randomly drop one self-attention module. As a result, models are forced to learn rich relations among diverse local patches. Our proposed TransFER model outperforms the state-of-the-art methods on several FER benchmarks, showing its effectiveness and usefulness.

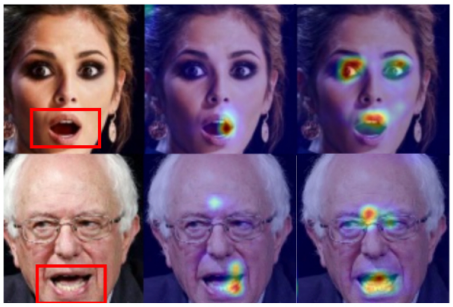

Figure 1. Attention visualizations [5] on two example images: Surprise (Row 1) and Anger (Row 2). Column 1: Original images. Column 2: Attention visualizations of our ViT-FER model. Column 3: Attention visualization of our TransFER model.

1. Introduction

In the past several decades, facial expression recognition (FER) has received increasing interest in the computer vision research community, as it is important to make computers understand human emotions and interact with humans.

Despite it had obtained excellent performance recently, FER is still a challenging task mainly due to two reasons: 1) Large inter-class similarities. Expressions from different classes may only exhibit some minor differences. As illustrated in Fig. 1, Surprise (Row 1) and Anger (Row 2) share a similar mouth. Critical clues to distinguish them lie in both eyes and areas between eyes; 2) Small intra-class similarities. Expressions belonging to the same class may have dramatically different appearances, varying with races, genders, ages to cultural backgrounds.

Existing works can be divided into two categories: global-based and local-based approaches. For the former,

*The first two authors contributed equally. This work was done when Fanglei Xue and Qiangchang Wang were interns at IDL, Baidu Research.
†Corresponding author

many loss functions are proposed to enhance the representational ability of features [18, 11]. However, since these methods take global facial images as the input, they may neglect some critical facial regions which would play an important role in distinguishing different expression classes. To overcome this issue, many local-based methods are proposed to learn discriminative features from different facial parts which can be divided into two sub-categories: landmark-based and attention-based approaches. [30, 15, 31] extracted features on facial parts which are cropped around landmarks. However, there are several issues: 1) Pre-defined facial crops may not be flexible to describe local details which may vary from different images. This is because important facial parts may appear at different locations, especially for faces with pose variations or viewpoint changes;

2) Facial landmark detection may be inaccurate or even fail for faces which are affected by various challenging factors, such as strong illumination changes, large pose variations, and heavy occlusions. Therefore, it is necessary to capture important facial parts and suppress useless ones.

To achieve the aforementioned goal, [19, 28] applied attention mechanisms. However, they may have redundant responses around similar facial parts, while neglecting other potentially discriminative parts which would play an important role in FER. This issue is especially serious for faces with occlusions or large pose variations where some facial parts are invisible. Therefore, diverse local representations should be extracted to classify different expressions. Consequently, more diverse local patches can contribute even when some patches are invisible. Meanwhile, different local patches can be complementary to each other. For example, as illustrated in Fig. 1, it is difficult to distinguish between surprise (Row 1) and anger (Row 2) based on the mouth areas only (Column 2). Our TransFER model explores diverse relation-aware facial parts, like eyes (Column 3, Row 1) and areas between the brows (Column 3, Row 2), which help distinguish these different expressions. Thus, the relations among different local patches should be explored in a global scope, highlighting important patches and suppressing the useless.

To achieve the above two goals, we propose the TransFER model to learn diverse relation-aware local representations for FER. First, the Multi-Attention Dropping (MAD) is proposed to randomly drop an attention map. In such a way, models are pushed to explore comprehensive local patches except for the most discriminative ones, focusing on diverse local patches adaptively. This is especially useful if some parts are invisible due to pose variations or occlusions. Second, Vision Transformer (ViT) [10] is adapted to FER, called ViT-FER, to model connections among multiple local patches. Since the global scope is used to reinforce each local patch, the complementarity among multiple local patches are well explored, boosting the recognition performance. Third, multi-head self-attention allows ViT to jointly attend to features from different information subspaces at different positions. Redundant relations may be built, however, since there is no explicit guidance. To address this, Multi-head Self-Attention Dropping (MSAD) is proposed to randomly drop one self-attention. In such a manner, if a self-attention is dropped, models are forced to learn useful relations from the rest. Consequently, rich relations among different local patches are explored to benefit the FER.

Combining the novel MAD and MSAD modules, we propose the final architecture, termed as TransFER. As illustrated in Fig. 1, compared with the ViT-FER baseline (Column 2), the TransFER locates more diverse relation-aware local representations (Column 3), distinguishing these different expressions. It achieves the state-of-the-art performance on several FER benchmarks, showing its effectiveness. The contributions of this work can be summarized as follows:

1. We apply ViT to characterize the relations between different facial parts adaptively, called ViT-FER, showing their effectiveness for FER. To the best of our knowledge, this is the first effort to explore Transformers and investigate the importance of relation-aware local patches for FER.

2. A Multi-head Self-Attention Dropping (MSAD) is introduced to randomly remove self-attention modules, forcing models to learn rich relations between different local patches.

3. An Multi-Attention Dropping (MAD) is designed to erase attention maps, pushing models to extract comprehensive local information from every facial part beyond the most discriminative parts.

4. Experimental results on several challenging datasets show the effectiveness and usefulness of our proposed TransFER model.

2. Related Work

In this section, related work about facial expression recognition, Transformers, and regularization methods are reviewed briefly.

2.1. Facial Expression Recognition

Facial expression recognition (FER) has remained as an active research area in the past decades. Traditionally, the hand-crafted features were developed to describe different facial expressions, such as LBP [22], HOG [7], and SIFT [21]. However, these features lack of generalization ability under some challenging scenarios, such as poor illumination conditions.

Recently, deep learning has greatly improved the FER research. Loss functions are designed in [18, 11] to enhance the discriminative ability of expression features. Each ROI

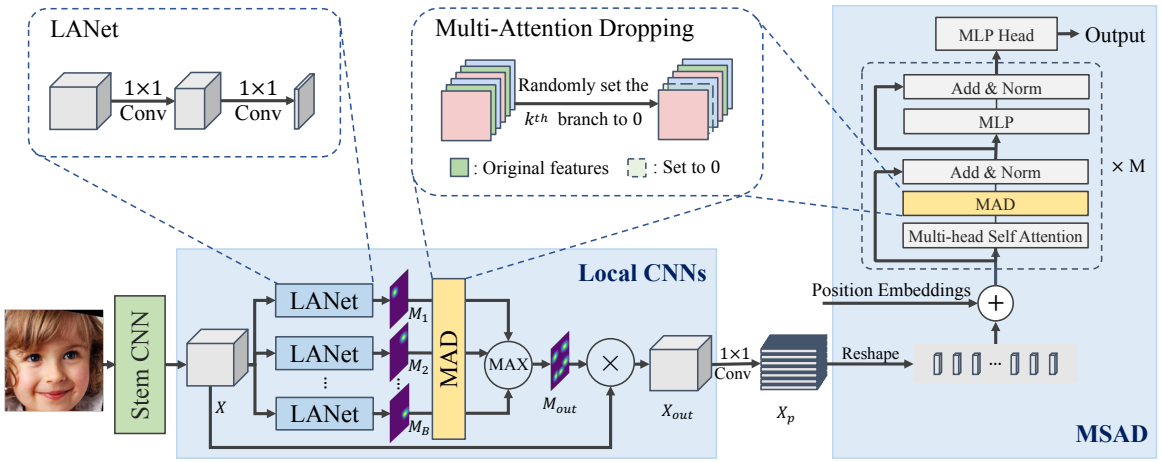

Figure 2. The overall architecture of our TransFER model. Firstly, facial images are first fed into a stem CNN to extract feature maps. Secondly, feature maps are then passed through the local CNNs to locate diverse useful feature areas. Thirdly, a 1×1 convolution and reshape operations are used to project feature maps to a sequence of feature vectors which can be directly input into MSAD (MAD in a Transformer encoder) where the relationships between these local patches are explored. An MLP Head is attached to generate the final classification result. MAD guides multiple local branches to locate diverse local patches. MSAD pushes multi-head self-attention to explore rich relations among different local patches.

is weighed in [19] via a gate unit that computes a weight from the region itself. A region attention network is proposed in [28] to adaptively capture the importance of facial regions for occlusion and pose variant FER.

Differently, a new adaptive loss is proposed in [17] to reweight category importance coefficients, alleviating the imbalanced class distribution. Besides, several works [27, 6] address label uncertainties in FER. In our approach, a mechanism is designed to locate diverse local patches. Besides, relations among different local patches are captured, which is the first attempt for FER, to the best of our knowledge.

2.2. Transformers in Computer Vision

Recently, Transformers [26] are applied to address computer vision problems [4, 10, 25].

An end-to-end object detection method reason about the locations of the objects, utilizing the Transformer decoder [4]. More recently, Vision Transformer (ViT) [10] treats images as a sequence of patches for image classification. Pre-trained on large-scale datasets, it obtained a competitive performance. Without the requirement of large-scale training data, DeiT [25] can accelerate training using a teacher-student strategy. In our work, it is the first attempt to explore Transformers for FER, to the best of our knowledge. Besides, this is also the first effort to show the importance of relations among local patches for FER.

2.3. Regularization Method

Overfitting is an important issue in deep neural networks. Dropout [23] randomly zeroes some of the elements in fully-connected layers, alleviating the overfitting problem. Despite its effectiveness, it is less effective in convolutional operations. This is because features are spatially correlated in CNNs. To address this challenge, Cutout [9] is proposed to randomly erase contiguous regions in the input image. DropBlock [12] further improves the Cutout by applying Cutout at every feature map. We propose the MSAD to effectively regularize the Transformers, exploring rich relations among different local patches.

3. TransFER

The overall architecture of our approach is shown in Fig 2, which mainly consists of the stem CNN, Local CNNs, and Multi-head Self-Attention Dropping (MSAD). The stem CNN is used to extract feature maps. The IR-50 [8] is adopted here since it has a good generalization.

As mentioned above, due to the small inter-class differences among different emotions, it is highly desired to extract diverse local patches. To achieve this goal, Multi-Attention Dropping (MAD) is devised to randomly drop facial parts. In such a way, multiple local branches in local CNNs are encouraged to locate diverse discriminative local patches. In MSAD, rich relations between different local patches are captured to boost the FER performance. This is achieved by randomly dropping self-attention modules. As

a result, multi-head self-attentions are complementary with each other, learning rich useful relations among different local patches. More details are illustrated as follows.

3.1. Local CNNs

As described earlier, given a facial image, our approach first uses a stem CNN to extract feature maps. Then, multiple spatial attentions are used to capture local patches automatically. However, if without a proper guidance, it is not guaranteed that comprehensive discriminative facial parts are located. If models focus on few discriminative facial parts, FER would suffer from performance degradation when these parts are hard to recognize or totally occluded, especially for faces with large pose variations, or strong occlusions. To address this, local CNNs are developed to extract diverse local features which are guided by the MAD.

The framework is shown in Fig. 2, mainly consists of three steps, which are detailed as follows.

Firstly, multiple attention maps are generated. Let $X \in R^{h\times w\times c}$ denote the input feature maps where h, w, and c refer to the height, width, and the number of feature maps, respectively. Since LANet [29] allows models to automatically locate important face parts, it is used in multiple local branches, as illustrated in Fig. 2. It consists of two 1×1 convolution layers. The first one outputs c/r feature maps where r is the reduction ratio to reduce the dimension of feature maps, followed by a ReLU layer to enhance the non-linearity. The second layer reduces the feature map number to one and generates an attention map by a Sigmoid function, which is denoted as M_i. Suppose there are B LANet branches, then attention maps $[M_1, M_2, ..., M_B]$ are generated where $M_i \in R^{h\times w\times 1}$.

Secondly, the MAD forces multiple local branches to explore diverse and useful facial parts, which would be presented in Section 3.2. Generally speaking, it takes several branches of data as input and randomly drops one branch by setting the values in this branch to zeroes (without changing the input shape). As a result, MAD takes B attention maps as input, randomly set one attention map to zeroes, and output B attention maps.

Thirdly, multiple attention maps are aggregated together to generate one attention map. To be specific, an element-wise maximum operation is used to aggregate multiple attention maps. Given a list of feature maps $[M_1, M_2, \ldots, M_B]$, the output M_{out} can be formulated as follows:

$$M_{out}(x, y) = \max\{M_1(x, y), M_2(x, y) \ldots M_B(x, y)\} \tag{1}$$

where $1 \le x \le w$ and $1 \le y \le h$.

Finally, we multiply M_{out} with the original feature map X using an element-wise production. Thus, unimportant areas in the original feature map are suppressed and vise versa.

To summarize, local CNNs are able to locate diverse local patches. This is achieved by using multiple LANets to locate multiple discriminative areas and aggregate them by a maximum operation, followed by element-wise multiplication with the input feature maps.

3.2. Multi-Attention Dropping

Dropout [23] is proposed to prevent neural networks from overfitting. It adapts a feature vector or feature map as input. During the training process, some of the elements of the input are randomly set to zeroes with a probability p using samples from a Bernoulli distribution. If there is more than one channel, each channel would be zeroed out independently. Inspired by this works, a dropout-like operation is developed for the FER task, called Multi-Attention Dropping.

In contrast with the standard Dropout, our proposed MAD adopts a group of feature maps (or vectors) as input and treats every feature map as a whole. As shown in the middle-upper part of Fig. 2, during the training process, one feature map is selected from a uniform distribution which is entirely set to zeroes. The drop operation is performed with a probability p_1. Dropped feature maps would not be activated in the following layers. Thus, a dropout-like stop-gradient operation is proposed, which can guide local CNNs to explore diverse and discriminative facial parts. As a consequence, well-distributed facial parts can be located, leading to comprehensive local representations to benefit the FER.

3.3. Multi-head Self-Attention Dropping

In order to explore the rich relationships among different local features generated by local CNNs, the Multi-head Self-Attention Dropping (MSAD) module is proposed. It mainly consists of a Transformer encoder with MAD injected behind every Multi-head Self Attention module and an MLP classification head like Vision Transformer (ViT) [10] does. The following are the details:

Projection. After local CNNs, feature maps $X_{out} \in R^{h\times w\times c}$ are generated which contain information about diverse local patches. To capture rich relations among multiple local patches, the Transformer is used which contains multiple encoder blocks. However, since the Transformer is first proposed for NLP tasks and adopts a sequence of 1D feature vectors as input. To adapt the Transformer, a projection module is developed to transform 2D sequence input to 1D.

As illustrated in Fig. 2, a 1×1 convolution layer is first applied to the X_{out}, projecting to feature maps $X_p \in R^{h\times w\times c_2}$ where the number of channels is denoted as c_2. So far, we do not change the height and width relationship between the feature maps X_p and the original image. So, every $c_2 \times 1$ vector can be considered as a representation of

a corresponding patch of the input image. So, we slice the X_p feature maps along the channel dimension and realign them as a sequence of feature vectors $x \in R^{(h \cdot w) \times c_2}$ which can be fed into the Transformer encoder directly.

Following [10], the learnable $[class]$ token is also appended to the sequence of input vectors. And standard learnable 1D position embeddings are added to the expanded sequence of vectors to inject position information.

Transformer Encoder. The Transformer encoder [26] is composed of a stack of M encoder blocks. Every block is composed of multiple layers of Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) with skip connections, as shown in the right of Fig. 2. A classification head implemented by a single layer of MLP is attached to perform classification output.

Firstly, the input $x \in R^{\mathrm{N} \times \mathrm{d}}$ is linearly transformed to queries q, keys k, and values v as follows:

$$[q, k, v] = x[w_q, w_k, w_v], \tag{2}$$

where $w_q, w_k \in R^{\mathrm{d} \times \mathrm{d_k}}$, $w_v \in R^{\mathrm{d} \times \mathrm{d_v}}$.

Secondly, the attention weights are computed as follows:

$$A = Softmax(\frac{qk^T}{\sqrt{d_k}}). \tag{3}$$

Thirdly, a weighted sum over all values is computed as follows:

$$O = Av. \tag{4}$$

The MSA runs self-attention operations k times in parallel and linearly embeds their concatenated outputs to form the final output.

The MLP consists of two fully-connected layers for feature projection and GELU [16] for non-linearity.

The MSA is designed to embed the projections in their respective space. However, if without an explicit signal, multiple self-attention modules tend to have redundant projections, limiting the representational ability. To address this issue, we utilize Multi-Attention Dropping (MAD) which is presented in Section 3.2 to randomly drop one of the attention heads, pushing models to learn comprehensive relations among different local patches.

Suppose there are k SAs in the MSA. One SA module is randomly selected from a uniform distribution and it is set to zeroes with a probability p_2.

MAD is performed across different MSA, that is, every sample in the same mini-batch and every MSA in the different block randomly select one SA from their self k SAs in every training iterations. This brings sufficient randomness to the samples and different MSA blocks in Transformer. Similar to Dropout, MAD is only performed during the training time. But during the inference time, unlike Dropout, MAD did not rescale the weights due to the different mechanisms with fully connected layers.

In such a way, models are encouraged to learn useful information since multiple self-attentions are pushed to complement to each other.

In general, more than one SA can be selected and dropped, but by our observation, dropping two or more SAs at the same time did not increase the performance. So for simplicity, we only consider the drop rate as a hyper-parameter and conduct all of our experiments with dropping only one branch in MAD.

4. Experiments

4.1. Datasets

RAF-DB [18] is a real-world expression dataset. It contains 29,672 real-world facial images which are collected by Flickr's image search API and independently labeled by about 40 trained human workers. In the experiments, the single-label subset provided in RAF-DB is utilized. It contains 15,339 expression images with six basic expressions (happiness, surprise, sadness, anger, disgust, fear) and neutral expression where 12,271 images of them are used in training and the rest are used for testing. The overall accuracy on the test set is reported.

FERPlus [3] is extended from FER2013 [13] which is a large-scale dataset collected by APIs in the Google image search. It contains 28,709 training, 3,589 validation, and 3,589 test images. They relabeled the dataset with ten labelers to eight emotion categories (six basic expressions, plus neutral and contempt). The overall accuracy is reported on the test set.

AffectNet [20] is the largest publicly available FER dataset so far. It contains about 1M facial images collected by three major search engines where about 420K images are manually annotated. Follow the settings in [17], we used 280K training images and 3.500 validation images (500 images per category) with seven emotion categories. The mean class accuracy on the validation set is reported.

4.2. Implementation Details

Since RAF-DB and FERPlus datasets provide annotated landmarks, these landmarks are used for face detection and alignment. For FERPlus dataset, the MTCNN [33] is used to detect and align faces. All images are aligned and resized to 112×112 pixels. Pre-trained on Ms-Celeb-1M [14], the IR-50 [8] is used as the stem CNN where only the first three stages in IR-50 are used. Pre-trained on ImageNet[1], ViT [10] with eight self-attention heads and a stack of $M = 8$ identical encoder layers are adopted as the Transformer Encoder.

Due to the class imbalance problem that is widely existed in FER, upsampling the training data is used to balance the

[1]The pre-trained weight is downloaded from https://github.com/rwightman/pytorch-image-models/.

Table 1. Evaluation (%) of local CNNs, MAD, and MSAD on RAF-DB and AffectNet.

Local CNNs	MAD	MSAD	RAF-DB	AffectNet
			89.93	65.63
✓			90.03	65.74
✓	✓		90.35	65.94
✓	✓	✓	90.91	66.23

Table 2. Evaluation (%) of different output stages of IR-50 on RAF-DB.

Stage	2	3	4
Acc.(%)	84.94	90.91	90.32

class distribution. The drop rates of MAD in local CNNs (p_1) and MSAD (p_2) are set to 0.6 and 0.3 for RAF-DB and FERPlus, and 0.2 and 0.6 for AffectNet, respectively based on our grid search.

Our TransFER is trained with the SGD optimizer to minimize the cross-entropy loss. We use the momentum of 0.9 and no weight decay, a mini-batch size of 256 in our experiments. During training, data augmentation is utilized on-the-fly including random rotate and crop, random horizontal flip, and random erasing. At test time, we only resize the original image to 112×112 pixels and feed it to the model directly. For RAF-DB and FERPlus, we train 40 epochs with an initial learning rate of 1e-3 decayed by a factor of 10 at the 15 and 30 epochs. For AffectNet, due to its large number of samples, we train 20K iterations with an initial learning rate of 3e-4 decayed by a factor of 10 at 9.6K and 19.2K iterations. We train our model on two NVIDIA V100 GPU with 32GB RAM.

4.3. Ablation Studies

Effectiveness of the proposed modules. To validate the proposed modules in our TransFER model, an ablation study is designed to investigate the effects of local CNNs, MAD, and MSAD on RAF-DB and AffectNet, as shown in Table 1. To efficiently show results, a tuple (a, b) is used where a and b denote the performance on RAF-DB and AffectNet, respectively.

The baseline strategy (the first row) means there is no local CNN, no MAD, and no MSAD. The feature maps extracted from the stem CNN are directly fed into the standard Transformer encoder without any guide from MAD or MSAD. Compared with the baseline, local CNNs slightly improve the performance by $(0.1\%, 0.11\%)$, but gains significant improvement with the addition of MAD $(0.42\%, 0.30\%)$. It is hypothesized that multiple LANets cannot generate diverse attention maps without extra supervision. MAD achieves this by randomly dropping one LANet branch during the training process, guiding the local CNNs to explore more recognizable feature areas. The MSAD further improves the performance, which achieves the state-of-the-art performance of $(90.91\%, 66.23\%)$, improved by $(0.56\%, 0.29\%)$. We believe that this is due to the multiple self-attentions that are pushed to complement each other and learn comprehensive and useful representations.

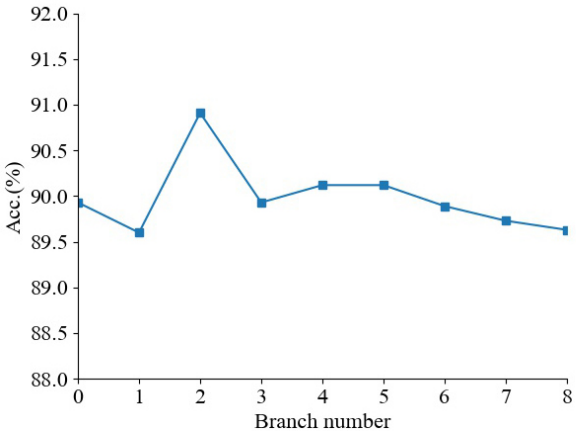

Figure 3. The evaluation of branch number (B) in Local CNNs on RAF-DB.

Determination of Stem CNN depth. As we know, given a CNN network, deeper layers generate more high-level and semantic information while shallow layers contain more texture and detailed information. For our proposed framework, we need both semantic information for local CNNs to locate more precise positions and detailed information to fed into MSAD for further extraction. So we designed this ablation study to determine which stage of IR-50 is the best for the FER task.

Like ResNet-50, IR-50 has four stages, every stage is made up of two convolutional layers and a max pooling layer to quarter the feature map size. Since the input image is in size 112×112, the output feature size in four stages are 56×56, 28×28, 14×14, 7×7, respectively. The feature map size of the first stage out is very large, making too many parameters for the following modules, so we only examine the stages 2 to 4.

From the results in Table 2, stage 3 achieves the best performance of 90.91% while stage 4 gives a comparative performance (90.32%), but with more parameters. Stage 2 performs much worse, only achieves 84.94%, proving that the LANet is not able to locate well semantic features.

Evaluation of B in local CNNs. As we have described in Section 3.1, B denotes the number of local branches in local CNNs. To explore the impact of branch number B, we evaluate the B from 0 to 8 on RAF-DB with other parameters as default. The evaluation results are shown in Figure 3. As B increases, the performance first increases and starts to decrease after $B = 5$.

Table 3. Evaluation of different drop rate in MAD (p_1) and MSAD (p_2) on RAF-DB.

p_1	p_2	Acc.(%)
0.4	0.3	89.28
0.5	0.3	89.24
0.6	0.3	**90.91**
0.7	0.3	89.89
0.8	0.3	89.89
0.6	0.1	89.80
0.6	0.2	89.83
0.6	0.3	**90.91**
0.6	0.4	89.93
0.6	0.5	89.24

The best performance (90.91%) is achieved when B is set to 2. Small B makes TransFER hard to locate robust and important feature parts and only achieved 89.60%. Large B degrades the ability of TransFER, since more branches may fall into a "collapsing solution" with almost the same outputs. AffectNet is a more difficult dataset, so the best performance is achieved with $B = 4$.

Evaluation of drop rate in MAD and MSAD. To evaluate the impact of drop rates in MAD and MSAD, Experiments of different drop rates are designed on RAF-DB. Denote p_1, p_2 as the drop rate in MAD and MSAD respectively, they are set to 0.6 and 0.3 by default. As shown in Table 3, both small and large p_1, p_2 values reduce the model performance. When p_2 is set to 0.3, p_1 changes from 0.4 to 0.8, the performance first increases from 89.28$ to 90.91%, after that, decreases back to 89.89%.

The same phenomenon is observed on p_2 when p_1 fixed to 0.6. The performance first increases from 89.80% to 90.91% when p_2 is set to 0.3, and decreases back to 89.24% as p_2 continues to increase.

There are eight self-attention heads in MSAD while MAD only has two LANet branches. The best p_2 value is smaller than the p_1 value indicates that self-attention in MSAD can grab important areas more effectively on RAF-DB.

Comparison among MAD, Dropout, Drop Block, and Spatial Dropout.

First, formally speaking, MAD accepts a set of attention maps as input, randomly selects one and drops the whole selected map. This is the reason why we call it Self-Attention Dropping. In contrast, Dropout [23], Drop Block [12] and Spatial Dropout [24] are directly applied to feature maps. Dropout treats all inputs equally and drops them independently, Drop Block drops units in a contiguous region, and Spatial Dropout drops the entire channel. They all perform independently element-wise or channel-wise, which is not suitable for input cases with multiple attention maps. To verify our hypotheses, we perform experiments on RAF-DB and AffectNet with these methods and our proposed MAD.

Table 4. Comparison among our MAD, Dropout, Drop Block and Spatical Dropout.

Dataset	**MAD**	Dropout	Drop Block	Spatial Dropout
RAF-DB	**90.91%**	90.35	90.25%	89.99%
AffectNet	**66.23%**	66.06	66.03%	65.54%

Other hyper-parameters are the default as described in 4.1.

We replace our MAD with these methods and perform a grid search to find the best hyper-parameters. The best result of each method is shown in Tab. 4. Dropout achieves the best performances with dropping rate 0 and 0.1 for RAF-DB and AffectNet, respectively. We also find that, with a dropping rate of 0.6, the model seems did not work on RAF-DB (39.05%) but achieves comparative performance on AffectNet (65.51%). This may because AffectNet contains more training data thus the model can learn from more diverse situations.

The best result of Drop Block is achieved with a dropping rate of 0.3, and the block size is 7 and 9, respectively on RAF-DB and AffectNet. The best dropping rate for Spatial Dropout is 0.2 for both two datasets. Our MAD achieves the best performances, we believe this is because other methods perform dropping independently channel-wise. This works for feature maps because the channel number is big, but not suitable for attention maps with few branches in our case.

Table 5. Performance comparison (%) with the state-of-the-art methods on RAF-DB and AffectNet.

Method	RAF-DB	AffectNet
DLP-CNN [18]	80.89	54.47
gACNN [19]	85.07	58.78
IPA2LT [32]	86.77	55.71
RAN [28]	86.90	52.97
CovPool [1]	87.00	-
SCN [27]	87.03	60.23
DACL [11]	87.78	65.20
KTN [17]	88.07	63.97
TransFER (Ours)	**90.91**	**66.23**

4.4. Comparison with the State of the Art

Table 5 compares our best results to the state-of-the-art methods on RAF-DB and AffectNet. RAF-DB is the latest facial expression dataset, and to our best knowledge, our proposed TransFER is the first model to achieve accuracy over 90% on this dataset, which is 2.84% better than KTN [17], the best result reported before. AffectNet is the largest dataset of facial expressions, a very challenging dataset. KTN [17] achieved the second-best performance in RAF-DB which is 1.23% lower than the best result reported previously on AffectNet. Our proposed approach outperforms the previous best result (DACL) by 1.03%.

Table 6. Performance comparison (%) with the state-of-the-art methods on FERPlus.

Method	FERPlus
PLD [3]	85.10
RAN [28]	88.55
SeNet50 [2]	88.80
RAN-VGG16 [28]	89.16
SCN [27]	89.35
KTN [17]	90.49
TransFER (Ours)	**90.83**

Tabel 6 compares the performance of our TransFER with the state-of-the-art methods on FERPlus. It can be seen that our method achieves the best accuracy of 90.83%. It is noting that both SCN [27] and KTN [17] achieve that reported performance by applying trivial loss functions, while we achieve better performance with the standard CE loss only.

4.5. Attention Visualization

To further investigate the effectiveness of our approach, we employ the method [5] to visualize the attention maps generated by our TransFER. To be specific, we first resize the visualization attention maps to the same size as the input images and visualize attention maps through COLORMAP_JET color mapping to the original image.

Fig. 4 shows the attention maps of different emotions in AffectNet. The figure has seven rows, each row shows one of the seven categories of expression. From top to bottom, the categories are anger, disgust, fear, happiness, neutral, sadness, and surprise. The first column shows the original aligned facial images, and the second to fifth columns show the results of four training strategies which have been listed in Table 1: (I) the baseline strategy; (II) adds multiple LANets to generate multiple attention maps but without MAD to guide; (III) has both multiple LANets and MAD; (IV) the whole architecure, including multiple LANets, MAD, and MSAD.

Firstly, comparing different columns (training strategies): local CNNs (II) can locate more potential interesting areas compared with the baseline (I), and MAD in local CNNs (III) and MSAD (IV) enhanced these candidate areas by exploring more interesting areas (*e.g.* (a), (b) in (III)) and constraining less dividing areas (*e.g.* (c) in (IV)).

Secondly, comparing different rows (emotions): It is generally assumed that mouth, nose, and eyes are the most useful regions to distinguish different emotions. But as discussed in [17], due to the high similarity of different emotions, these areas may be very similar even for different emotions. For example, fear (c), happiness (d), and surprise (g) often have an open mouth, so it's more important to explore other facial areas to discriminate against different emotions. Our proposed MSAD solves this problem by constraining the activation of the mouth area in (c) (IV) and exploring other useful areas in (g) (IV), compared to (III).

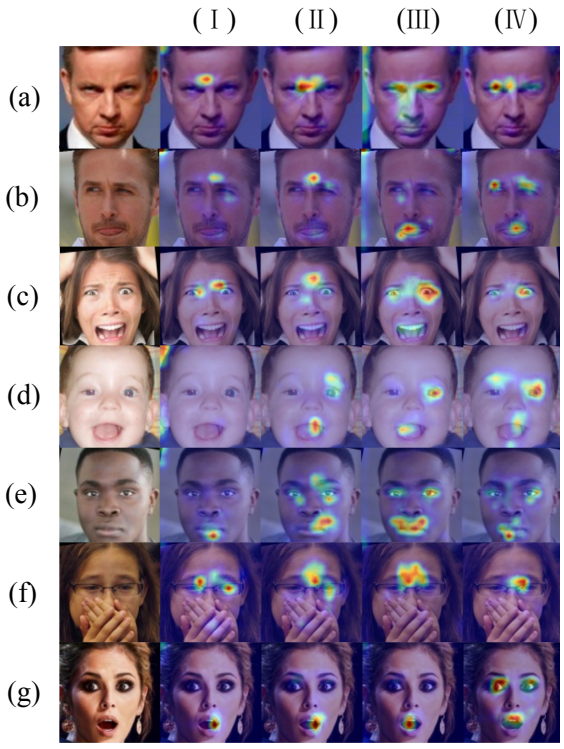

Figure 4. Attention visualization [5] of different expressions on some example face images from AffectNet dataset. (a) - (g) denote anger, disgust, fear, happiness, neutral, sadness, and surprise separately. (I) - (IV) denote four training strategies in Tab. 1, (I) denotes the baseline strategy, (II) denote training with local CNNs but without MAD, (III) denotes training with local CNNs and MAD, and (IV) denotes our proposed TransFER, training with local CNNs and MSAD. After applying MAD and MSAD, the whole framework can focus on more discriminative facial areas.

5. Conclusion

We have proposed a new architecture based on the Transformer for the FER task, called TransFER, which can learn rich, diverse relation-aware local representations. Firstly, a Multi-Attention Dropping (MAD) has been proposed to guide local CNNs to generate diverse local patches, making models robust to pose variations or occlusions. Secondly, the ViT-FER is applied to build rich connections upon multiple local patches where important facial parts are assigned with higher weights and useless ones are assigned smaller weights. Thirdly, the MSAD has been proposed to explore more rich relations among diverse facial parts. To the best of our knowledge, this is the first work to utilize the Transformers for the FER task. Extensive experiments on three public FER datasets demonstrated that our approach outperforms the state-of-the-art methods.

References

[1] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Covariance pooling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 367–374, 2018. 7

[2] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 292–301, 2018. 8

[3] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016. 5, 8

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[5] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. *arXiv preprint arXiv:2012.09838*, 2020. 1, 8

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13984–13993, 2020. 3

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 2

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 3, 5

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 4, 5

[11] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2402–2411, 2021. 2, 7

[12] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *arXiv preprint arXiv:1810.12890*, 2018. 3, 7

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

[14] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 5

[15] SL Happy and Aurobinda Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, 6(1):1–12, 2014. 2

[16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[17] Hangyu Li, Nannan Wang, Xinpeng Ding, Xi Yang, and Xinbo Gao. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Transactions on Image Processing*, 30:2016–2028, 2021. 3, 5, 7, 8

[18] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 2, 5, 7

[19] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018. 2, 3, 7

[20] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 5

[21] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003. 2

[22] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009. 2

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 3, 4, 7

[24] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using Convolutional Networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 648–656. IEEE Computer Society, Oct. 2015. 7

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 3

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3, 5

[27] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020. 3, 7, 8

[28] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 2, 3, 7, 8

[29] Qiangchang Wang and Guodong Guo. Ls-cnn: Characterizing local patches at multiple scales for face recognition. *IEEE Transactions on Information Forensics and Security*, 15:1640–1653, 2019. 4

[30] Siyue Xie and Haifeng Hu. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *IEEE Transactions on Multimedia*, 21(1):211–220, 2018. 2

[31] Siyue Xie, Haifeng Hu, and Yongbo Wu. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recognition*, 92:177–191, 2019. 2

[32] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018. 7

[33] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 5