OCCLUSION-ROBUST FACIAL EXPRESSION RECOGNITION BY ADVERSARIAL LEARNING

Anonymous ICME submission

ABSTRACT

Although facial expression recognition has achieved huge success in laboratory environments, it is still challenging under unconstrained environments, especially under partial occlusion conditions. In this paper, we propose an occlusion-robust neural network, which can improve the model's generalization under partial occlusion conditions. Specifically, we design a feature extractor to alleviate the effect of occlusion regions, which exploits attention mechanism with pixel-level occlusion supervision to enhance the perception of non-occluded areas. In order to extract occlusion-independent facial features, we introduce feature discriminator to measure the correlation between the extracted facial features and the occlusion labels. Adversarial learning technique is applied between the feature extractor and feature discriminator so that feature extractor can extract occlusion-robust facial features discriminative for facial expression recognition. Experimental results on the benchmark databases show the superior performance of the proposed method.

Index Terms— facial expression recognition, occlusion supervision, adversarial learning

1. INTRODUCTION

Due to wide potential application in human-computer emotional interaction and video generation, facial expression recognition receives much attention and achieves much progress. However, facial expression recognition is still challenging due to the existence of partially occluded face.

Previous occluded facial expression recognition works can be categoried into two categories: reconstruction-based and weighting-based methods. Reconstruction-based methods reconstruct landmark information or non-occluded images, and then extract facial features from reconstructive counterpart to conduct facial expression recogniton [1, 2, 3]. However, these methods hardly reconstruct non-occluded images in detail because of complexity and variation of occlusion. To lessen the influence of the occlusion, weighting-based methods pay less attention to occluded areas[4, 5, 6, 7]. Most of them adopt attention mechanism without occlusion supervision to make the model learn weights for occluded areas automatically. However, due to the limited data and variation of occlusions, the attention mechanism without occlusion supervision hardly learn good attention weights for complex and realistic occluded images.

Different from previous works, we propose to learn occlusion-robust facial features by attention mechanism with pixel-level occlusion supervision and adversarial training, which can lessen the influence of the occlusion to facial expression recognition. Specifically, the proposed network consists of a feature extractor, an expression classifier and a feature discriminator. The feature extractor includes an attention module with pixel-level occlusion supervision, which can make the model pay less attention to occluded areas. In order to extract occlusion-robust but expression-discriminative features, adversarial learning mechanism is applied between the feature extractor and feature discriminator. During adversarial training, the feature extractor hopes to generate facial features independent from corresponding occlusion labels, while the feature discriminator wants to distinguish correlation between them as much as possible. Thus the proposed method can exploit extracted occlusion-robust features to improve facial expression performance under occlusion conditions.

2. RELATED WORK

Because of the variability of occlusion, occluded facial expression recognition is challenging. Previous approachs can be classified into two categories: reconstruction-based methods and weighting-based methods.

Reconstruction-based methods. Bourel *et al.* [1] recovered lost or drifting facial landmarks by an enhanced Kanade-Lucas tracker, and then extracted facial geometric features based on landmark information. However, low-level geometric features hardly solve complex occluded facial expression recognition problem. Mao *et al.* [2] used robust principal component analysis(RPCA) technology to reconstruct images and extract local features from the reconstructed images. However, generating fine-enough non-occluded facial images is still a thorny issue, and low-quality reconstructed images may bring about noise which badly influences the classifier. Pan *et al.* [3] introduced a decoder network to insure that the extracted facial features can reconstruct the corresponding non-occluded images. At the same time, they extracted facial features from occluded images and corresponding non-occluded images by distinct networks respectively, and compeled distributions of the two kinds of facial features to be

close by adversarial training. However, they train two distinct networks for occluded and non-occlude images, and thus need to know whether the images are occluded or not before prediction, which greatly limits its application in the realistic occluded environment.

Weighting-based methods. Dapogny *et al.* [4]proposed to train random forests on spatially defined local subspaces of the face, and used local expression predictions (LEPs) as high-level representations. LEPs can be further weighted by confidence scores provided by an autoencoder network. However, the definition of confidence coefficients depends on human experience and is not necessarily optimal. Recently, many researchers used attention mechanism to make the model learn weights for occluded areas automatically. Li *et al.* [5] extracted 24 facial patches from the feature maps based on landmarks information, and then extracted local features from the facial patches by attention-based Patch Gated Units (PG-Unit). Considering the lack of global information, they further introduced attention-based Global Gated Unit (GG-Unit) to extract global features[6]. Wang *et al.* [7] introduced relation-attention module on the basis of self-attention module to learn better attention weights. However, these attention-based works don't introduce any occlusion supervision to optimize attention module, which hardly learn good attention weights for complex and realistic occluded images.

Compared with previous works, our contribution can be summarized as follows: (1) We introduce attention module with pixel-level occlusion supervision, which can enhance the perception ability of the model to the non-occluded areas. (2) We further extract facial features independent from occlusion labels by adversarial training.

3. METHODOLOGY

The overall network architecture of the proposed method is shown in Figure 1. It is composed of three components, including the feature extractor F, the expression classifier R and the feature discriminator D^F.

3.1. Problem Statement

Let $T = \{(x, y)\}^N$ denote N training samples, where x represents the input image with or without occlusions and $y = \{y^{occ}, y^{occ}_{mask}, y^e\}$ represents the ground truth label. $y^{occ} \in \{0, 1\}$ denotes whether there are artificial occlusions in the input image. $y^{occ}_{mask} \in \{0, 1\}^{W \times H}$ denotes the artificial occlusion mask of the input image. W and H denote the width and height of the input image respectively. $y^e \in \{1, 2, ..., C\}$ denotes the expression category where C denotes the number of expression categories. The purpose of our work is to train a feature extractor F and an expression classifier R for occlusion-robust facial expression recognition.

3.2. Attention Mechanism with Occlusion Supervision

We use the four bottlenecks of Resnet[8] as the backbone of the feature extractor and introduce attention module to make the model pay less attention to occluded areas automatically. Let ϕ, Res^{low} and Res^{up} denote the attention module, the first bottleneck and the last three bottlenecks respectively. The feature extractor can be formalized as followings,

$$F(x) = Res^{up}(Res^{low}(x) \odot (1 - \phi(Res^{low}(x)))) \tag{1}$$

where $\phi(Res^{low}(x))$ denotes the predicted occluded mask of image x and $\odot$ denotes element-wise multiplication.

To further enhance the perception of attention module to non-occluded areas, we use artificial occlusion masks y^{occ}_{mask} to supervise the learning of the attention module. We resize the synthesized occluded mask to the same width and height as the output feature map of attention module. The loss is shown in Equation 2,

$$\begin{aligned}\mathcal{L}_{noisy} = -\frac{1}{w \times h} \sum_{i,j} (y^{occ}_{mask}[i, j] \log p^{occ}_{mask}[i, j] \\ + (1 - y^{occ}_{mask}[i, j]) \log(1 - p^{occ}_{mask}[i, j]))\end{aligned} \tag{2}$$

where w and h denote the width and height of the output feature map of attention module respectively. $y^{occ}_{mask}[i, j]$ denotes whether i-row, j-col pixel belongs to synthesized occluded areas or not. $p^{occ}_{mask}[i, j]$ denotes the predicted probability that i-row, j-col pixel belongs to occluded areas.

3.3. Adversarial Learning Mechanism

To further remove occlusion-related information from the extracted features, we compel the extracted facial features to be independent on occlusion labels. We introduce feature discriminator to predict whether there are occlusions or not in the input images. When the extracted facial features are independent on occlusion labels, the formula $P(y^{occ}|F(x)) = P(y^{occ})$ holds. Therefore, we can use the Kullback-Leibler divergence between $P(y^{occ}|F(x))$ and $P(y^{occ})$ to measure the independence of the extracted facial features and occlusion labels. Further, Equation 3 holds when $H(P(y^{occ}))$ represents the information entropy of $P(y^{occ})$, and $H(P(y^{occ}), P(y^{occ}|F(x)))$ represents the cross entropy between $P(y^{occ})$ and $P(y^{occ}|F(x))$.

$$\begin{aligned}&KL(P(y^{occ})||P(y^{occ}|F(x))) = \\ &H(P(y^{occ}), P(y^{occ}|F(x))) - H(P(y^{occ}))\end{aligned} \tag{3}$$

Because $H(P(y^{occ}))$ is a constant, we could use the cross entropy to measure the independence.

When fixing D^F and optimizing F, we hope to maximize the independence between the extracted facial features and occlusion labels, i.e., minimizing the difference between $P(y^{occ}|F(x))$ and $P(y^{occ})$. Thus we get the cross entropy loss of the feature extractor $\mathcal{L}_F$ as shown in Equation 4,

$$\mathcal{L}_F = -\sum_{i=0}^{1} P(y^{occ} = i) \log P(y^{occ} = i|F(x)) \tag{4}$$

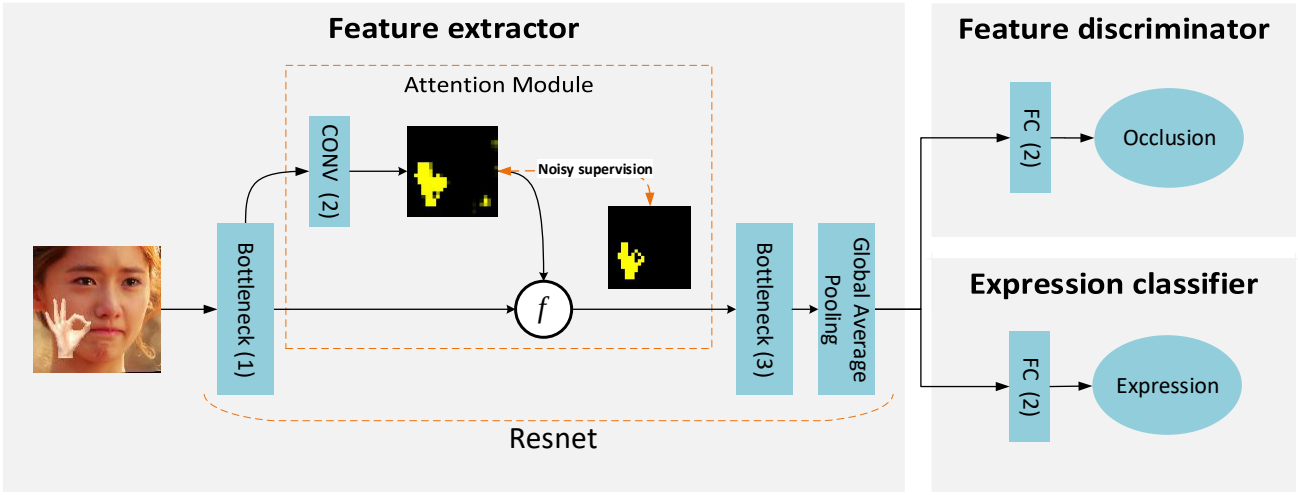

Fig. 1. Overall architecture of the proposed method. "FC" and " CONV" represent the fully connected layer and convolution layer respectively. The number in parentheses denotes the repeated times of the corresponding layer.

Because $P(y^{occ})$ is unknown, we replace $P(y^{occ})$ with occlusion labels' empirical distribution $P_E(y^{occ})$ in the training set T. And we further rewrite the loss function as follows,

$$\mathcal{L}_F = -\sum_{i=0}^{1} P_E(y^{occ} = i) \log P(y^{occ} = i|F(x)) \tag{5}$$

When fixing F and optimizing D^F, we hope to maximize the correlation between the extracted facial features and occlusion labels, i.e., enchancing the classification ability of D^F. Thus we directly minimize the binary cross entropy loss of the feature discriminator as follows,

$$\mathcal{L}_D^F = -\sum_{i=0}^{1} I(i = y^{occ}) \log P(y^{occ} = i|F(x)) \tag{6}$$

By adversarial training between the feature extractor F and the feature discriminator D^F, the proposed method can remove occlusion-related information as soon as possible and thus extract occlusion-robust facial features.

3.4. Training Process

For expression recognition task, we use multi-category cross entropy loss $\mathcal{L}_{emp}$ as shown in Equation 7, where p_i^e represents the predicted probability that the sample belongs to i^{th} expression category, $I(i = y^e) = 0$ if $i \neq y^e$, $I(i = y^e) = 1$ if $i = y^e$.

$$\mathcal{L}_{emp} = -\sum_{i=1}^{C} I(i = y^e) \log p_i^e \tag{7}$$

We combine the classification loss $\mathcal{L}_{emp}$, attention loss $\mathcal{L}_{noisy}$ and the adversarial loss $\mathcal{L}_F$ to update feature extractor and expression classifier together.

$$\mathcal{L} = \mathcal{L}_{emp} + \lambda_1 \mathcal{L}_{noisy} + \lambda_2 \mathcal{L}_F \tag{8}$$

where λ_1 and λ_2 are the hyperparameters controlling loss coefficients.

We train the whole network in an adversarial style, i.e., updating the feature discriminator when fixing the feature extractor and expression classifier according to Equation 6, updating feature extractor and expression classifier when fixing the feature discriminator according to Equation 8.

4. EXPERIMENT

4.1. Experiment Condition

We evaluate the proposed method on the AffectNet [9], RAF-DB [10], CK+ [11] and FED-RO [6] like previous works [5, 6, 3, 7]. The AffectNet is a large-scale facial expression database with more than 1M facial images collected from the Internet. About 440K images are manually annotated for the presence of basic facial expressions and the intensity of valence and arousal. We use images with 8 basic emotions (i.e., neural, happy, sad, surprise, fear, disgust, anger and contempt), including 287K images as training set and 4K images as validation set. The Real-world Affective Faces Database(RAF-DB) contains around 30K images. We use 7 basic emotions (i.e., neural, happy, sad, surprise, fear, disgust and anger), containing 12,271 images as training set and 3,068 images as test set. The Extended Cohn-Kanade dataset (CK+) contains 593 video sequences recorded from 123 subjects in the laboratory environment. Each video starts with an onset frame and ends with an apex frame. We collect onset frames labelled as neutral category and apex frames with

six basic emotions (i.e., happy, sad, surprise, fear, disgust and anger). As a result, 636 facial images are collected. The Facial Expression Dataset with Real Occlusions (FED-RO) is the first facial expression database with realistic occlusions in the wild, which is collected and annotated by Li *et al.* [6]. In total, 400 facial images with various occlusion are downloaded from the Internet and labeled with 7 classes of basic emotions. Because of the lack of datasets with realisitc occlusions, Wang *et al.* [7] select 683 realisitc occluded images and 735 realistic occluded images from the validation set of the AffectNet and the test set of the RAF-DB respectively.

Following the way of previous works [5, 6, 3], occluded images are artificially synthesized by adding occluding objects at random locations of the faces on all databases as shown in the first row of Figure 2. The occluding objects include objects that appear frequently in occluded faces like food, hands and drinks. Each kind of occluding object has different templates. We restrain these occluders to suitable size which ranges from 1/3.75 to 1/2 the size of the original image in both width and height like [5, 6, 3].

To demonstrate the effectiveness of different parts of the proposed method, we apply four models on each database. The first model consists of the Resnet and the expression classifier, and minimizes the classification loss $\mathcal{L}_{emp}$. The second model introduces noisy-robust attention loss $\mathcal{L}_{noisy}$ and classification loss $\mathcal{L}_{emp}$. The third model adopts adversarial loss $\mathcal{L}_F$ and $\mathcal{L}_{emp}$. The last model optimizes $\mathcal{L}_{noisy}$, $\mathcal{L}_F$ and $\mathcal{L}_{emp}$ simultaneously. According to the differences of models' components, the four models are named as M_{Res}, M_{ResAtt}, M_{ResAdv} and $M_{ResAttAdv}$ respectively.

Following Li *et al.*'s work [5, 6], training sets are mixed with synthesized images with a ratio of 1:1 during training process, we report accuracies on the validation set of the AffectNet and test set of the RAF-DB dataset respectively and do cross-database experiments on the CK+ database to verify the effectiveness of the proposed method on synthesized occluded images. Besides, to show the effectiveness of the proposed methods on realistic occluded images, we show accuracies on the selected realistic occluded part of the validation set of the AffectNet and the test set of the RAF-DB dataset respectively and do cross-database experiments on the FED-RO database. We implement the proposed method using Pytorch deep learning framework and use Resnet50 pretrained on ImageNet database to initialize the feature extractor. We use Adam optimizer, a weight decay of 1e-4, a batch size of 128, cosine periodic learning rate with $lr_{epochs} = 30$, $lr_{min} = 1e^{-5}$ and $lr_{max} = 1e^{-3}$. We set hyper-parameter $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$.

4.2. Experiment Results and Analysis

The experimental results of facial expression recognition on synthesized occlusions and realistic occlusions are shown in Table 1 and 2 respectively. From the Table 1 and Table 2, we have the following findings:

Table 1. Experimental results of facial expression recognition under synthesized occlusions. Values not in square parentheses and values in square parentheses denote the accuracy on the original images and the synthesized occluded images respectively. "AffectNet(C7)" denotes that the experiments ignore the contempt expression and are 7-category facial expression recognition. "CK+(AffectNet)" and "CK+(RAF-DB)" denote cross-database experimental results on the CK+ database of models trained on the AffectNet database and RAF-DB database respectively.

Methods	AffectNet (C7)	RAF-DB	CK+ (AffectNet)	CK+ (RAF-DB)
PGCNN[5]	0.5533 [0.5247]	0.8327 [0.7805]	0.9038 [0.8627]	0.8028 [0.7949]
gACNN[6]	0.5878 [0.5484]	0.8507 [0.8194]	0.9164 [0.8817]	0.8107 [0.7949]
Pan *et al.*[3]	[0.5642]	[0.8197]	[0.8990]	[0.7921]
M_{Res}	0.5766 [0.5597]	0.8520 [0.8211]	0.9009 [0.8947]	0.8019 [0.7704]
M_{ResAtt}	0.5854 [0.5704]	0.8615 [0.8357]	0.9135 [0.8994]	0.8208 [0.8097]
M_{ResAdv}	0.5864 [0.5706]	0.8605 [0.8366]	0.9150 [0.8010]	0.8176 [0.8082]
$M_{ResAttAdv}$	**0.5984 [0.5809]**	**0.8719 [0.8484]**	**0.9261 [0.9135]**	**0.8270 [0.8223]**

Firstly, the losses from the attention mechanism and adversarial mechanism both lead to a great improvement of facial expression recognition accuracy comparing with the baseline model M_{Res}. For example, the accuracies of M_{ResAtt} and M_{ResAdv} are 1.46%, 1.55% higher than that of M_{Res} on the RAF-DB database of synthesized occluded images. The experimental results on the AffectNet and CK+ databases also show similar trend. The model M_{ResAtt} introduces attention module with pixel-level occlusion supervision, which makes the model pay less attention to occluded areas and thus achieves better performance. M_{ResAdv} removes occlusion-related information from extracted facial features by adversarial training and thus achieves the improvement.

In order to further show the effectiveness of attention mechanism with pixel-level occlusion supervision, we visualize the learned occlusion attention masks of synthesized and realistic occluded images respectively. As show in Figure 2, we can find that the proposed method can percept non-occluded areas well on the synthesized and realistic occluded images.

Secondly, our method combines the strengths of the two introduced loss functions and achieves the best performance in both synthesized occluded and realistic occluded database. Specifically, the FER accuracy of our method $M_{ResAttAdv}$ is 2.11% and 3.23% higher than baseline model on the RAF-DB and AffectNet database of realistic occluded images, which is much better than using a single loss functions. This indicates that the different module will not cause discrepancy,

and losses in attention module and adversarial module can help network learn occlusion-robust feature representations and make better predictions.

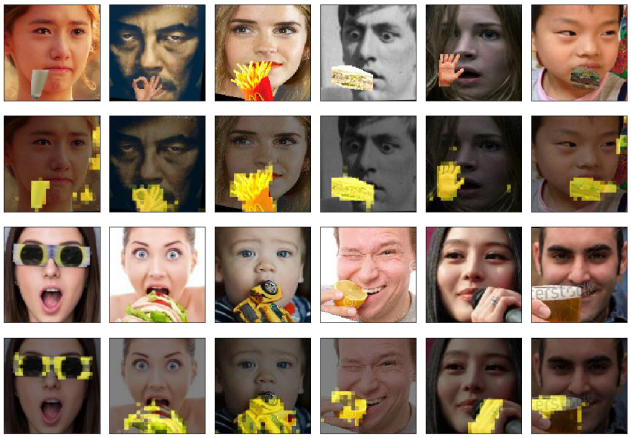

Fig. 2. Examples of the learned occluded area masks on the AffectNet and FED-RO databases. The first two rows are the synthesized occluded images and corresponding learned occlued areas. The last two rows are realistic occlued images and corresponding learned occlued areas.

4.3. Comparison with Related Work

To illustrate the superiority of the proposed method, we compare it with weighting-based methods (i.e., PGCNN [5], gACNN[6] and RAN [7]) and reconstruction-based methods (i.e., Pan *et al.*'work [3]). From Table 1, we find that the model $M_{ResAttAdv}$ shows the best performance compared with PGCNN, gACNN and Pan *et al.*'method. Different from PGCNN and pACNN which only apply attention mechanism to weight occluded areas automatically, our method $M_{ResAttAdv}$ introduces pixel-level occlusion supervision for the attention module, which enhances the perception of the model to non-occluded areas and thus $M_{ResAttAdv}$ shows better performance than PGCNN and pACNN. Pan *et al.* compels the distributions of the facial features extracted from original faces and occluded faces to be close. However, they use distinct networks for occluded images and non-occluded images, which is harmful for the learning of common information between them. Our method extracts occlusion-robust facial features by attention module with pixel-level occlusion supervision and adversarial training, which contributes to its performance.

From Table 2, we find that the model $M_{ResAttAdv}$ shows the best or comparable performance on the realistic occlusions. Specifically, our method achieves better accuracy than RAN by 2.11% and 3.23% on the AffectNet and RAF-DB database respectively. Similar to PGCNN and gACNN, RAN weights occluded areas and occluded samples by attention mechanism. However, these attention-based works don't introduce any occlusion supervision information to optimize attention module. We also compare the generalization ability of our method with related works by cross-database experiment. Experimental results with realistic occlusion on the FED-RO database show that our method outperforms related works. It demonstrates that the occluded feature learned by attention mechanical with pixel-level occlusion supervision can contain more important information for facial expression recognition with realistic occlusion.

Table 2. Experimental results of facial expression recognition under realistic occlusions. Values not in square parentheses and values in square parentheses denote the accuracy on the original datasets and the corresponding selected realisitc occluded datasets respectively. "AffectNet(C8)" denotes that the experiments use all the samples and are 8-category facial expression recognition. "FED-RO" denotes cross-database experimental results on the FED-RO, but the model is trained on the AffectNet and RAF-DB databases simultaneously.

Methods	**AffectNet (C8)**	**RAF-DB**	**FED-RO (AffectNet RAF-DB)**
PGCNN[5]	-	-	[0.6425]
gACNN[6]	-	-	[0.6650]
Pan *et al.*[3]	-	-	[0.6975]
RAN [7]	0.5950 [0.5850]	0.8690 [0.8272]	[0.6798]
M_{Res}	0.5875 [0.5850]	0.8520 [0.8059]	[0.6824]
M_{ResAtt}	0.5987 [0.5916]	0.8615 [0.8209]	[0.6925]
M_{ResAdv}	0.6003 [0.5922]	0.8650 [0.8224]	[0.6975]
$M_{ResAttAdv}$	**0.6115** [**0.6061**]	**0.8719** [**0.8382**]	[**0.7000**]

4.4. Analysis of Occlusions Variation

To analyze the robustness of the proposed method on occlusions with diverse sizes and locations, we synthesize occluded images with five occlusion modes on the CK+ database, i.e., eye occlusion, mouth occlusion, 8 × 8 occlusion, 16 × 16 occlusion and 24 × 24 occlusion. As shown in Table 3, we have the following findings:

Firstly, the proposed method $M_{ResAttAdv}$ achieves the best performance compared with other methods, which demonstrates that the proposed method can adapt to variation of occlusions on the size and shape. Results for larger or more difficult occlusion types, i.e., mouth occlusion and 24 × 24 occlusion, our framework shows particular improvement. Specificly, our framework achieves 2.51% and 2.36% increases than baseline M_{Res}under mouth occlusion and 24 × 24 occlusion. Our method $M_{ResAttAdv}$ can percept non-occluded areas well by attention module with pixel-level oc-

Table 3. Experimental results of facial expression recognition with five types of synthesized occlusions on the CK+ database. R8, R16 and R24 denote the size of the occlusion as 8×8, 16×16 and 24×24 respectively. Both the occlusion size of "eye occluded" and "mouth occluded" are 24×24.

Methods	R8	R16	R24	R24(eye occluded)	R24(mouth occluded)
WLS-RF[4]	0.9220	0.8640	0.7480	0.8790	0.7270
PGCNN[5]	0.9658	0.9570	0.9286	0.9650	0.9392
gACNN[6]	0.9658	0.9597	0.9286	0.9657	0.9388
Pan *et al.*[3]	0.9780	0.9686	0.9403	0.9686	0.9355
M_{Res}	0.9638	0.9544	0.9513	0.9513	0.9308
M_{ResAtt}	0.9733	0.9638	0.9544	0.9638	0.9450
M_{ResAdv}	0.9748	0.9701	0.9544	0.9670	0.9481
$M_{ResAttAdv}$	**0.9827**	**0.9764**	**0.9654**	**0.9748**	**0.9544**

clusion supervision and will not suffer notable performance degradation under different occlusion conditions.

Secondly, comparing the first three columns of the Table 3, we find that the recognition accuracies of all methods decrease significantly as the occlusion size increases, which is in line with our common sense. In the last three columns, the occlusion sizes are the same but there are significant differences. As is known to all, eyes and mouth are the areas with the most abundant facial movements, therefore, the accuracies under eyes or mouth occluded should be lower than the accuracies under random areas occluded. However, the experimental results support condition of mouth occluded but don't support condition of eyes occluded. In deed, eyes and mouth are important areas when people making expressions, but the movement of the mouth is much greater than that of the eyes. Obviously, the model tends to concern facial areas with a large range of movements. Therefore, mouth areas play a more important role compared with other facial patches in facial expression recognition.

5. CONCLUSION

In this paper, we propose an occlusion-robust neural network for facial expression recognition under partial occlusion conditions. On the one hand, we introduce attention mechanism with pixel-level occlusion supervision to enhance the perception of the model to non-occluded areas. On the other hand, we extract occlusion-independent facial features by adversarial training between the feature extractor and feature discriminator. Experimental results show the proposed method has good generalization on synthesized and realistic occluded facial expression recognition.

6. REFERENCES

[1] Fabrice Bourel, Claude C Chibelushi, and Adrian A Low, "Recognition of facial expressions in the presence of occlusion.," in *BMVC*, 2001, pp. 1–10.

[2] Xia Mao, YuLi Xue, Zheng Li, Kang Huang, and Shan-Wei Lv, "Robust facial expression recognition based on rpca and adaboost," in *2009 10th Workshop on Image Analysis for Multimedia Interactive Services*. IEEE, 2009, pp. 113–116.

[3] Bowen Pan, Shangfei Wang, and Bin Xia, "Occluded facial expression recognition enhanced through privileged information," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019, pp. 566–573.

[4] Arnaud Dapogny, Kevin Bailly, and Séverine Dubuisson, "Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 255–271, 2018.

[5] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen, "Patch-gated cnn for occlusion-aware facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2209–2214.

[6] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.

[7] 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*, vol. 29, pp. 4057–4069, 2020.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] 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*, vol. 10, no. 1, pp. 18–31, 2017.

[10] 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*, 2017, pp. 2852–2861.

[11] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 94–101.