Occluded Facial Expression Recognition using Self-supervised Learning

Jiahe Wang[1], Heyan Ding[1], and Shangfei Wang[1,2⋆]

[1] Key Lab of Computing and Communication Software of Anhui Province, University of Science and Technology of China
[2] Anhui Robot Technology Standard Innovation Base, University of Science and Technology of China
pia317@mail.ustc.edu.cn, dhy0513@mail.ustc.edu.cn, sfwang@ustc.edu.cn

Abstract. Recent studies on occluded facial expression recognition typically required fully expression-annotated facial images for training. However, it is time consuming and expensive to collect a large number of facial images with various occlusions and expression annotations. To address this problem, we propose an occluded facial expression recognition method through self-supervised learning, which leverages the profusion of available unlabeled facial images to explore robust facial representations. Specifically, we generate a variety of occluded facial images by randomly adding occlusions to unlabeled facial images. Then we define occlusion prediction as the pretext task for representation learning. We also adopt contrastive learning to make facial representation of a facial image and those of its variations with synthesized occlusions close. Finally, we train an expression classifier as the downstream task. The experimental results on several databases containing both synthesized and realistic occluded facial images demonstrate the superiority of the proposed method over state-of-the-art methods.

Keywords: Occluded facial expression recognition · Self-supervised Learning · Representation Learning.

1 Introduction

Facial expressions play an important role in our daily communication. In recent years, facial expression recognition has attracted increasing attention and achieved great progress [18, 4, 8, 21] because of its application in many fields, such as psychological treatment, security and service robots. However, it is still challenging to recognize facial expressions from occluded facial images.

Current methods of occluded facial expression recognition can be divided into four categories: robust facial representation, non-occluded facial image reconstruction, sub-region analysis, and non-occluded facial image help. Robust facial representation methods aim to locate the representation that is insensitive to occlusion but discriminative for expression recognition. It is very difficult to

⋆ Corresponding author

find a robust representation because the types of occlusion are diverse and positions of occlusion are infinite. Non-occluded facial image reconstruction methods aim to construct non-occluded facial images using a generative model and train the facial expression classifier from the reconstructed facial images. However, the generated facial images are typically not as realistic as the real images, and this effects the performance of facial expression recognition. Sub-region analysis methods divide the image into several regions and recognize expressions from these regions and the entire image. Dividing facial images typically requires facial landmarks, and the attention mechanism is used to select important regions. However, facial landmark detection from the occluded facial image remains challenging. Non-occluded image help methods adopt non-occluded facial images as privileged information to assist occluded facial expression recognition. During training, these methods typically construct two networks: one for non-occluded facial expression recognition and the other for occluded facial expression recognition. During testing, these methods assume that all facial images are occluded and only the network for occluded expressions is used, whereas in a realistic scenario, we do not know whether the facial image is occluded or not. Furthermore, all the above methods require fully expression-annotated images for training. Because the types and positions of occlusion are infinite, collecting a large-scale dataset with various facial expressions and occlusions is difficult.

To address this, we propose an occluded facial expression recognition method through self-supervised learning [6]. We use a large number of unlabeled facial images in the pretext task to learn a robust and occlusion-insensitive facial representation. First, we synthesize many occluded images by randomly adding different occlusions to a large number of images. We apply occlusion detection as the pretext task to learn the representation. We also use contrastive learning to make the representation of facial image and those of its variations with synthesized occlusions similar. Finally, we set occluded expression recognition as the downstream task.

Our contributions are as follows: We are the fisrt to introduce a large number of unlabeled facial images for occluded expression recognition through self-supervised learning. We design an occluded detection and similarity constraint between the occluded and non-occluded facial images as the pretext task.

2 Related Work

Because of the variability of occlusion, occluded facial expression recognition is still a big challenge. Current work can be classified into four categories: sub-region analysis[3, 18, 11, 10], robust facial representation [20, 2], non-occluded facial image reconstruction[17, 12] and non-occluded facial image help[16, 19].

Sub-region analysis methods typically divide the image into several regions and obtain results from the regions and the entire image. These methods often apply the attention mechanism. Wang *et al.*[18] proposed a region attention network that adjusted the importance of facial parts and designed a region biased loss function to obtain a high attention weight for the important region.

Li *et al.*[11] presented a patch-gated convolutional neural networks (PG-CNN) for facial expression recognition under occlusion. The PG-CNN chose 24 interest patches that were fed into an attention network to extract local features and learn an unobstructed score. The final classifier was constructed based on the weighted concatenated local features of all regions. Then, Li *et al.*[10] further proposed the global gated unit to add the global information of facial images for facial expression recognition. Dapogny *et al.*[3] used random forests to train partially defined local subspaces of the face and adapted local expression predictions as high-level representations. Then they weighted confidence scores provided by an autoencoder network. However, to divide the images or obtain regions, these methods typically require facial landmarks. It is difficult to detect facial landmarks from occluded facial images, which greatly affects the results of sub-region analysis methods. The detected error may be propagated to the classification task.

Robust facial representation methods aim to find a visual representation that is robust to occlusion. Zhang *et al.*[20] used a Monte Carlo algorithm to extract a set of Gabor templates from images and converted these templates into template match distance features. Cornejo and Pedrini[2] used robust principal component analysis to reconstruct occluded facial regions and then extracted census transform histogram features. However these methods do not have good generalization ability. It is difficult to find features that are insensitive to occlusion, because the positions of occlusion are unlimited.

Non-occluded facial image reconstruction methods exploit a deep generative model to construct non-occluded facial images. Lu *et al.* [12] exploited a generator to complement the non-occluded image and then the generated complementation image was used to predict the expression. They used reconstruction loss, triplet loss and adversarial loss to implement occluded facial image complementation. Ranzato *et al.* [17] used a deep belief network to construct a non-occluded face from the occluded face and then predicted the expression from the complete face. However, their visualization of occluded images was not good because of the unlimited positions and types of occlusion. Errors caused by an inaccurate reconstruction facial image may be propagated to the final task.

Non-occluded facial image help methods train facial expression classifiers from occluded facial images with the assistance of non-occluded facial images. Generally, non-occluded facial images have more useful information than occluded facial images. Pan *et al.*[16] used non-occluded images as privileged information to enhance the occluded classifier. Pan *et al.* trained two deep neural networks for occluded and non-occluded images separately and then used the non-occluded network to guide the occluded network. Xia *et al.*[19] proposed a stepwise learning strategy to obtain a robust network. They divided occluded and non-occluded images into three subsets from simple to difficult. Then they input the three subsets into the network to learn parameters in stages. They also used least squares generative adversarial networks [14] to reduce the feature gap between occluded facial images and non-occluded facial images. However, non-occluded facial image help methods trained two distinct networks for occluded

and non-occluded images. During the tests, all facial images were assumed to be occluded, which may have affected the results of non-occluded facial images.

To summarize, existing methods require a large number of expression-labeled occluded images. However, occluded facial images with expression labels are difficult to collect. Thus, we generate a large number of occluded facial images from unlabeled facial images to simulate real occluded images, where the positions and types of occlusion vary. Therefore, in this study, we propose an occluded facial expression recognition method that uses self-supervised learning. Specifically, we design a pretext task related to occluded facial images. We apply contrastive learning to maximize the similarity between the facial image and its variations with synthesized occlusions. We also design an occlusion detection task as the pretext task. Then we set expression recognition as the downstream task and add a classifier to fine-tune the model.

3 Method

The framework of the proposed occluded facial expression recognition through self-supervised learning is shown in Figure 1. The training process is mainly composed of two tasks. In the pretext task, we use the pre-training set to obtain an initial feature extractor F. In the downstream task, we add a classifier C after feature extractor F, and use the training set to fine-tune the parameters of extractor F and classifier C.

3.1 Problem Statement

Let $\mathcal{D}_{pre\_train} = \left\{x_{pc}^{(i)}\right\}_{i=1}^{N_p}$ denote the pre-training set of N_p training samples, where $x_{pc}^{(i)} \in \mathbb{R}^{H\times W\times 3}$ is a non-occluded facial image, which has no expression label. We generate occluded facial image $x_{po}^{(i)}$ from non-occluded facial image $x_{pc}^{(i)}$. We randomly select the type of occlusions from N_c types of occlusions, and the positions of occlusions are random. $x_{po}^{(i)}$ has a corresponding occlusion mask $\mathbf{M} \in \{0,1\}^{H\times W}$, where H and W denote the height and width of the input image respectively. $\mathcal{D}_{fine\_train} = \left\{x_{fo}^{(i)}, y^{(i)}\right\}_{i=1}^{N_{fo}} \cup \left\{x_{fc}^{(i)}, y^{(i)}\right\}_{i=1}^{N_{fc}}$ denotes the training set of $N_{fo} + N_{fc}$ training samples, which have expression labels. We obtain $\left\{x_{fo}^{(i)}\right\}_{i=1}^{N_{fo}}$ by adding occlusion to $\left\{x_{fc}^{(i)}\right\}_{i=1}^{N_{fc}}$. $y^{(i)} \in \{0,1,...,N_e-1\}$ represents the expression label of the i^{th} sample. $\mathcal{D}_{test} = \{x_o^{(i)}\}_{i=1}^{N_1} \cup \{x_c^{(i)}\}_{i=1}^{N_2}$ denotes the N_1 occluded and N_2 non-occluded testing samples. Given $\mathcal{D}_{pre\_train}$ and $\mathcal{D}_{fine\_train}$, we first use self-supervised learning to obtain the initial parameters of extractor F on the pre-training set, and then fine-tune extractor F and classifier C on the training set. Our goal is to learn a network $f : \mathbb{R}^{H\times W\times 3} \rightarrow \{0,1,\cdots,N_e-1\}$, which improves prediction for occluded expression images.

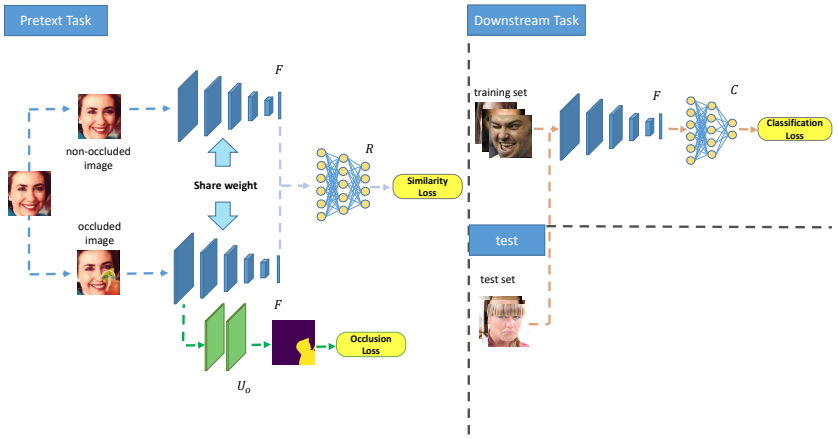


Fig. 1. Framework of the method. In the pretext task, extractor F extracts features; mask recognition network U_o predicts the location of the occlusion; and projection head R outputs the facial representation. In the downstream task, classifier C predicts the expression.

3.2 Pretext Task

Because of the infinite types and positions of occlusion, we always lack expression-annotated facial images with occlusions for learning robust facial representation. Fortunately, we may synthesize a large number of facial images with various occlusions. Therefore, we design pretext tasks to learn representations from facial images with synthesized occlusions.

To extract distinguishing and occlusion-insensitive features, we use many unlabeled facial images to obtain good initialization parameters of extractor F in the pretext task.

Specifically, we add occlusion to a non-occluded facial image to obtain an occluded facial image. Then we introduce an occlusion detector after a certain layer of the feature extractor to predict the occlusion position. Through occlusion detection, the learned features from a certain layer of the feature extractor may be aware of the importance of facial areas. Then, we adopt contrastive learning to to make the learned representation of facial image and those of its variations with synthesized occlusions similar. Therefore, the learned features are robust to occlusion.

Similarity Loss After we generate occluded facial images from non-occluded facial images, we expect that these representation will be similar. Contrastive learning maximizes the similarity between positive pairs and minimizes the similarity between negative pairs, which meets our needs. Given a mini-batch containing K samples $\{x_{pc}^{(j)}\}_{j=1}^{K}$, we randomly add occlusion to $\{x_{pc}^{(j)}\}_{j=1}^{K}$ to obtain $\{x_{po}^{(j)}\}_{j=1}^{K}$. Then both $x_{po}^{(j)}$ and $x_{pc}^{(j)}$ are fed into extractor F to extract features

$h_{po}^{(j)} = F(x_{po}^{(j)})$ and $h_{pc}^{(j)} = F(x_{pc}^{(j)})$. We use projection head R to obtain facial representations $z_{po}^{(j)} = R(h_{po}^{(j)})$ and $z_{pc}^{(j)} = R(h_{pc}^{(j)})$. Because $x_{po}^{(j)}$ is transformed from $x_{pc}^{(j)}$, we consider $(z_{po}^{(j)}, z_{pc}^{(j)})$ as the positive pair, $(z_{po}^{(j)}, z)_{z \in \{z_{po}^i, z_{pc}^i\}_{i=1}^K / \{z_{po}^j, z_{pc}^j\}}$ and $(z_{pc}^{(j)}, z)_{z \in \{z_{po}^i, z_{pc}^i\}_{i=1}^K / \{z_{po}^j, z_{pc}^j\}}$ as the negative pair. In our case, the positive pair has a higher similarity than the negative pair. We adopt cosine similarity to measure the similarity.

The similarity loss is

$$\mathcal{L}_{SS} = \frac{1}{2K} \sum_{j=1}^{K} (\mathcal{L}_{spo}^j + \mathcal{L}_{spc}^j) \tag{1}$$

where $\mathcal{L}_{spo}^j$ and $\mathcal{L}_{spc}^j$ are the similarity loss of $z_{po}^{(j)}$ and $z_{pc}^{(j)}$, respectively. The specific forms of $\mathcal{L}_{spo}^j$ and $\mathcal{L}_{spc}^j$ are

$$\begin{aligned} \mathcal{L}_{spo}^j &= -\log \frac{\exp\left(sim\left(z_{po}^{(j)}, z_{pc}^{(j)}\right)/\tau\right)}{\sum\limits_{z \neq z_{po}^{(j)}} \exp\left(sim\left(z_{po}^{(j)}, z\right)/\tau\right)} \\ \mathcal{L}_{spc}^j &= -\log \frac{\exp\left(sim\left(z_{po}^{(j)}, z_{pc}^{(j)}\right)/\tau\right)}{\sum\limits_{z \neq z_{pc}^{(j)}} \exp\left(sim\left(z_{pc}^{(j)}, z\right)/\tau\right)} \end{aligned} \tag{2}$$

where τ is the temperature parameter and $sim(u, v) = u^\top v / \|u\| \|v\|$ denotes the cosine similarity between two vectors u and v. By minimizing $\mathcal{L}_{spo}^j$ and $\mathcal{L}_{spc}^j$, the similarity of positive pairs in the numerator is increased and the similarity of negative pairs in the denominator is decreased.

Occlusion Loss In facial images, the occluded area typically contains less or even no information about the facial expression. In the pretext task, the occluded position of synthesized occluded facial images is easy to obtain. When generating a synthesized occluded facial image, we can obtain a binary mask $\mathbf{M}$ about the occlusion position, where 1 represents occlusion and 0 represents no occlusion. We expect that the network can contain the occluded position information; hence, we add a mask recognition network U_o to obtain $\hat{\mathbf{M}}$ to predicte binary mask $\mathbf{M}$. Because $\mathbf{M}$ is a binary mask, we use the cross-entropy loss to optimize the result. The occlusion loss is

$$\begin{aligned} \mathcal{L}_{mask} = -\frac{1}{H \times W} \sum_{j,k} (&\mathbf{M}[j,k] \log \hat{\mathbf{M}}[j,k] \\ &+ (1 - \mathbf{M}[j,k]) \log(1 - \hat{\mathbf{M}}[j,k])) \end{aligned} \tag{3}$$

where $\mathbf{M}[j,k]$ denotes whether the point (j,k) is occluded. $\hat{\mathbf{M}}[j,k]$ represents the probability of predicting whether the point (j,k) belongs to occlusion. The point (j,k) represents the point in the k^{th} row and j^{th} column.

The overall pretext task loss is defined as

$$\mathcal{L}_{pre} = \mathcal{L}_{SS} + \lambda \mathcal{L}_{mask} \tag{4}$$

where λ is a hyperparameter that balances the trade-off between the two losses.

The overall loss considers both the occluded position and facial features. We optimize the model by minimizing the overall pretext task loss.

3.3 Downstream Task

We set occluded expression recognition as the downstream task. In the pretext task, we obtain extractor F for occluded facial images. In the downstream task, we obtain feature extractor F obtained in the previous step and add a classifier C after it. Then we use the training set to fine-tune extractor F and classifier C. $\hat{y} = C(F(x))$ performs facial expression recognition for facial image x, where $x \in \mathcal{D}_{fine\_train}$. We use cross-entropy loss to measure the difference between it and the truth label. The classification loss is

$$\mathcal{L}_{cla} = \mathcal{L}_{CE}\left(\hat{y}, y\right) \tag{5}$$

3.4 Optimization

Among the losses, we only use $\mathcal{L}_{SS}$ and $\mathcal{L}_{mask}$ in the pretext task, and $\mathcal{L}_{cla}$ in the downstream task. We first use $\mathcal{L}_{SS}$ and $\mathcal{L}_{mask}$ to obtain extractor F, and then use $\mathcal{L}_{cla}$ to fine-tune extractor F and classifier C. We use Adam[7] to update the parameters.

4 Experiment

4.1 Experimental Conditions

For the pretext task, we chose a large-scale face recognition dataset VGGFace2[1] as the pre-training set. Following the experimental conditions in Pan *et al.*'s study [16] and Li *et al.*'s study [11], we conducted within-database experiments on synthesized occluded databases, that is, the Real-world Affective Faces Database (RAF-DB)[9], AffectNet database[15], and extended Cohn-Kanade database (CK+)[13]. When testing our method on the Facial Expression Dataset with Real Occlusions (FED-RO) [10], we merged AffectNet and RAF-DB as the training set. We also tested our methods on the original test databases, that is, RAF-DB and AffectNet. Following the experimental conditions in Wang *et al.*'s study[18], we tested our model on Occlusion-AffectNet[18] and Occlusion-RAF-DB[18]. The details are as follows:

VGGFace2 includes 3.31M images from $9,131$ subjects. Each subject has an average of 362.6 images. The database is downloaded from Google Image Search and has large variations in ethnicity, age, and pose. In our experiment, we used the VGGFace2 database as our pre-training set.

RAF-DB contains approximately 30K real-world images annotated with basic or compound expressions using 40 annotators. In our experiment, we only used images with seven basic expressions (i.e., neutral, happiness, sadness, surprise, fear, disgust and anger); hence, we used 12,271 images as the training set and 3,068 images as the test set.

AffectNet was collected from the internet by querying expression-related keywords. It contains more than 1M facial images, of which approximately 450K images were annotated by 12 human experts. In our experiment, we used facial images with seven basic expressions, which included approximately 280K images as the training set and $3,500$ as the test set.

CK+ consists of 593 sequences from 123 subjects. The image sequence begins with the onset frame and ends with the apex frame. We collected onset and apex frames as neutral and target expressions. In our experiment on CK+, we collected 636 facial images and adopted 10-fold subject-independent cross-validation.

Occlusion-AffectNet and Occlusion-RAF-DB were selected from the validation set of AffectNet and test set of the RAF-DB by Wang *et al.*[18]. In these two test sets, we used the same experimental conditions as those in the Wang *et al.*' study. Occlusion-AffectNet includes 683 realistic occluded facial images with eight basic expressions (i.e., neutral, happiness, sadness, surprise, fear, disgust, anger and contempt). Because Occlusion-AffectNet contains eight basic expressions, we used images with eight basic expressions from AffectNet, which included approximately 287K images as the training set and Occlusion-AffectNet as the test set. Occlusion-RAF-DB contains 735 realistic occluded facial images with seven basic expressions. In our experiment, we used images from RAF-DB with seven basic expressions, which included $12,271$ images as the training set and Occlusion-RAF-DB as the test set.

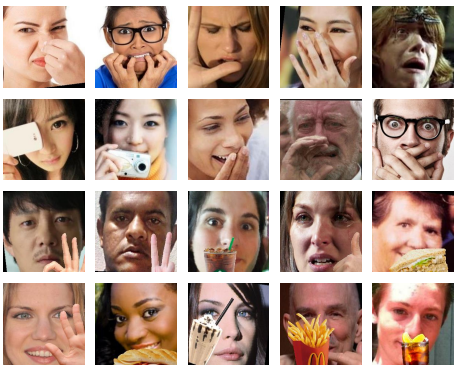


Fig. 2. Examples of realistic occluded facial images in FED-RO and the synthesized occluded facial images in AffectNet. The first and second rows are realistic occluded facial images, and the third and fourth rows are synthesized occluded facial images.

FED-RO is the first facial expression dataset to present real occlusions in the wild, and was collected by Li *et al.*[11]. FED-RO contains 400 images, which are labeled with seven basic expressions. Because FED-RO is small, we only used it for cross-database evaluation.

To mimic real-world scenarios, we artificially synthesized occluded facial images by adding occluding objects at random locations in all databases except the three real occlusion facial expression databases, that is, FED-RO, Occlusion-AffectNet, and Occlusion-RAF-DB. We used a variety of occlusion types to synthesize occluded facial images. The position of occlusion in each facial image was random. To compare our method with Pan *et al.*'s method [16], we used the same type of occlusion as that in Pan *et al.*'s study: food, hands and drinks. In Figure 2, we show some examples of realistic occluded facial images and synthesized occluded facial images in FED-RO and AffectNet. Because the Occlusion-AffectNet database on Wang *et al.*'s work and the AffectNet databases on Pan *et al.*'s work use different facial expression numbers, we use C7 to denote the seven classification task and C8 to denote the eight classification task. AffectNet(C7) and Occlusion-AffectNet(C8) represent seven and eight facial expression recognition tasks separately.

We conducted ablation experiments to verify the effect of the pretext task and the two different loss functions in the pretext task, that is, occlusion loss and similarity loss on AffectNet, Occlusion-AffectNet, RAF-DB, Occlusion-RAF-DB and FED-RO. First, we directly used ResNet-34 without pre-training as the baseline, which we refer to as the non-pretext task. Second, we used $\mathcal{L}_{SS}$ or $\mathcal{L}_{mask}$ as the pretext task, denoted by $\mathcal{L}_{SS}$ or $\mathcal{L}_{mask}$. Finally, we used $\mathcal{L}_{SS}$ and $\mathcal{L}_{mask}$ as the pretext task, denoted by $\mathcal{L}_{SS} + \mathcal{L}_{mask}$.

The implementation of the proposed method is based on the PyTorch framework. Because the AffectNet database is imbalanced, we resampled the data during training. We exploited ResNet-34 [5] to extract features. We built two fully connected layers as the small neural network projection head R to obtain the dimensional representation; two convolutional layers and an upsampling layer as the mask recognition network U_o and two fully connected layers as the classifier C. We add the mask recognition network U_o after the Conv1 layer of the ResNet-34. In our experiments, we resized the facial images in CK+ to 48×48 pixels and other images to 224×224 pixels. When conducting experiments on CK+, we used five types of occlusion, that is, 8×8 occlusion, 16×16 occlusion, 24×24 occlusion, eye occlusion and mouth occlusion to illustrate the robustness of the model to occlusions of different sizes. The batch size was 64. The hyperparameter λ was 0.2. The temperature parameter τ was 2. The learning rate of the network was $1e^{-4}$. We determined the hyperparameter in the loss function using grid search.

4.2 Experimental Results and Analysis

Analysis of facial expression recognition without occlusions The experimental results of facial expression recognition without occlusions are shown in Table 1. We trained the method on AffectNet(C7) and RAF-DB and tested it on

the original validation set of AffectNet(C7) and the original test set of RAF-DB, separately. Table 1 yields the following observations.

Table 1. Experimental results of facial expression recognition without occlusions on the AffectNet(C7) and the RAF-DB databases. (C7 represents seven classifications.)

Methods	AffectNet(C7)	RAF-DB
PG-CNN[11]	55.33	83.27
gACNN[10]	58.78	85.07
non-pretext	57.09	82.53
$\mathcal{L}_{SS}$	59.66	84.09
$\mathcal{L}_{mask}$	58.40	83.18
$\mathcal{L}_{SS} + \mathcal{L}_{mask}$	**60.20**	**85.95**

First, using $\mathcal{L}_{SS}$ or $\mathcal{L}_{mask}$ led to an improvement compared with the baseline non-pretext task. Specifically, the accuracies of $\mathcal{L}_{SS}$ and $\mathcal{L}_{mask}$ were 2.57% and 1.31% higher than that of the non-pretext task on AffectNet(C7). The experimental results on RAF-DB databases demonstrated a similar trend. Guidance regarding both the feature and occluded position helped the extractor F to learn more robust feature representations.

Table 2. Experimental results of facial expression recognition under synthesized occlusions on RAF-DB, AffectNet(C7) and CK+. (R8, R16 and R24 denote the sizes of the occlusions: 8 × 8, 16 × 16 and 24 × 24 respectively. The AffectNet(C7) represents seven classifications.)

Methods	**RAF-DB**	**AffectNet(C7)**	**CK+**				
			R8	**R16**	**R24**	**eye occluded**	**mouth occluded**
RGBT[20]	72.56	49.21	92.00	82.00	62.50	88.00	30.30
WLS-RF[3]	74.66	51.74	92.20	86.40	74.80	87.90	72.70
PG-CNN[11]	78.05	52.47	96.58	95.70	92.86	96.50	93.92
gACNN [10]	80.54	54.84	96.58	95.97	94.82	96.57	93.88
Pan *et al.*'s work[16]	81.97	56.42	97.80	**96.86**	94.03	96.86	93.55
Xia *et al.*'s work[19]	82.74	57.46	98.01	96.22	**95.91**	97.17	95.44
non-pretext	81.45	54.06	96.91	94.65	94.18	95.12	94.33
$\mathcal{L}_{SS}$	83.18	57.09	97.64	96.07	94.97	96.23	95.60
$\mathcal{L}_{mask}$	82.53	55.03	97.17	95.28	94.65	95.75	95.12
$\mathcal{L}_{SS} + \mathcal{L}_{mask}$	**84.06**	**58.40**	**98.27**	96.70	95.59	**97.33**	**96.07**

Second, similarity loss was more effective than occlusion loss, which were 1.26% and 0.91% higher than using only occlusion loss on AffectNet(C7) and RAF-DB. This may be because the test image was non-occluded; hence occlusion loss had little effect.

Third, our method achieved the best performance using similarity loss and occlusion loss together. Specifically, the accuracies of our method were 3.11%,

Table 3. Experimental results of facial expression recognition under realistic occlusions on FED-RO, Occlusion-AffectNet(C8) and Occlusion-RAF-DB. (C8 represents eight classifications.)

Methods	**FED-RO**	**Occlusion-AffectNet(C8)**	**Occlusion-RAF-DB**
PG-CNN[11]	64.25	-	-
gACNN[10]	66.50	-	-
Pan *et al.*'s work[16]	69.75	-	-
Xia *et al.*'s work[19]	**70.50**	-	-
RAN[18]	67.98	58.50	**82.72**
non-pretext	68.25	55.93	79.73
$\mathcal{L}_{SS}$	69.25	58.86	81.09
$\mathcal{L}_{mask}$	69.00	57.25	80.41
$\mathcal{L}_{SS} + \mathcal{L}_{mask}$	70.00	**59.30**	82.45

Table 4. Experimental results of λ sensitivity analysis on Occlusion-AffectNet(C8)

λ	Acc(%)
0.02	58.51
0.2	59.30
2	58.42
20	57.54

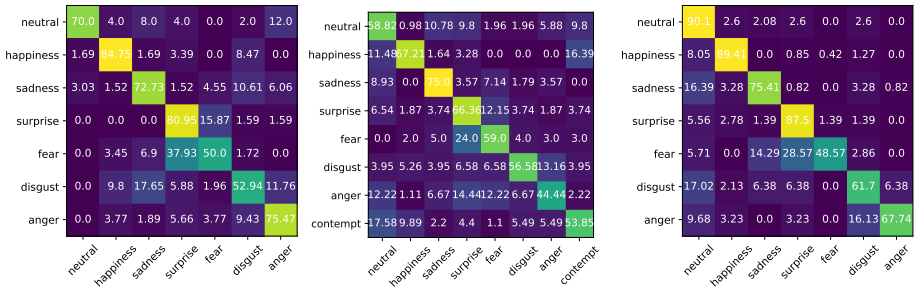
0.54% and 1.80% higher than those of the non-pretext task, $\mathcal{L}_{SS}$ and $\mathcal{L}_{mask}$ on AffectNet(C7), and 3.42%, 1.86% and 2.77% higher on RAF-DB, respectively. The pretext task of using both similarity loss and occlusion loss helped the downstream task to learn a more robust facial representation and make better predictions.

Analysis of facial expression recognition with synthesized occlusions The experimental results of facial expression recognition with synthesized occlusions are shown in Table 2. We trained our method on AffectNet(C7) and RAF-DB, and tested it on the synthesized occluded AffectNet(C7) and RAF-DB test set. The table yields the following observations:

First, our method achieved the best performance using both occlusion loss and similarity loss. Specifically, the accuracies of our method were 2.61%, 0.88% and 1.53% higher than that of non-pretext, $\mathcal{L}_{SS}$ and $\mathcal{L}_{mask}$ on RAF-DB, and 4.34%, 1.31% and 3.37% higher on AffectNet(C7), respectively . Such observations are consistent with experiments without occlusion.

Second, the experiment on CK+ obtained good results. The results demonstrated that our method was robust to occlusions of different sizes. As the size of the occlusion increased, the classification accuracy decreased. This demonstrated that the larger the size of the occlusion, the more useful information about the facial expression was occluded, and the more difficult it was to recognize facial expressions. The result of mouth occluded is lower than eye occluded.

This demonstrated that the mouth contains more expression information than the eyes. The experimental results for synthesized occluded facial expression recognition were similar to those of non-occluded facial expression recognition. This demonstrated that our method was helpful in both non-occluded facial expression recognition and occluded facial expression recognition.



(a) FED-RO (b) Occlusion-AffectNet(C8) (c) Occlusion-RAF-DB

Fig. 3. Confusion matrix on three realistic occlusion databases. (Row indices represent the ground truth labels, whereas column indices represent the predictions.)

Analysis of facial expression recognition with realistic occlusions The experimental results of facial expression recognition with realistic occlusions are shown in Table 3. We trained the method on AffectNet(C8) and RAF-DB and tested it on Occlusion-AffectNet(C8) and Occlusion-RAF-DB separately. We merged AffectNet(C7) and RAF-DB to obtain the training set and used FED-RO as the test set for the cross-database experiment. Table 3 yields the following observations.

First, the performance using both occlusion loss and similarity loss as the pretext task obtained better classification accuracy. For example, the accuracies of our method were 3.37%, 0.44% and 2.05% higher than those of the non-pretext task, $\mathcal{L}_{SS}$, and $\mathcal{L}_{mask}$ on Occlusion-AffectNet(C8); 1.75%, 0.75% and 1.00% higher on FED-RO; and 2.72%, 1.36% and 2.04% higher on Occlusion-RAF-DB , respectively. In addition, we also make a sensitivity analysis on λ on Occlusion-AffectNet(C8), and the results are shown in Table 4 .The results show that the lambda we set can well balance the trade-off between the two losses.

Second, we investigated the per expression category classification performance on FED-RO, Occlusion-AffectNet(C8) and Occlusion-RAF-DB. The confusion matrices based on our method are shown in Figure 3. These matrices show that our method achieved high accuracy in the happiness category. Many images of fear were mistakenly classified as surprise. The results demonstrated that it was difficult to distinguish between fear and surprise. The correct rate of disgust was relatively low on the three test sets. This may be because the

numbers of facial images showing disgust in AffectNet and the RAF-DB were relatively small.

Third, we obtained good results for realistic occlusions, synthesized occlusions, and the original test sets, which indecates the effectiveness of our method. The pretext task helped facial expression recognition. To visualize the perception ability of the mask recognition network U_o, we also created a result map of the generated mask, which is shown in Figure 4. Our network found the location of the synthesized occlusions. Our network also detected part of the real occluded positions.

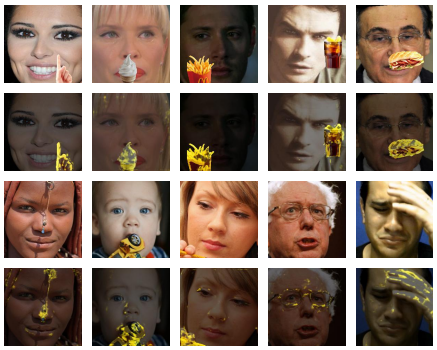


Fig. 4. Examples of the learned occluded masks on AffectNet and FED-RO. The first two rows are the synthesized occluded images and corresponding learned occluded areas. The last two rows are realistic occluded images and corresponding learned occluded areas.

4.3 Comparison with Related Work

To illustrate the superiority of our method, we compared it with state-of-the-art methods, that is, RGBT[20], WLS-RF[3], PG-CNN[11] and gACNN[10] on realistic occluded databases, synthesized occluded databases, and the original (non-occluded) test set. Because the Occlusion-AffectNet(C8) and the Occlusion-RAF-DB databases are new collected by Wang *et al.*[18], on these databases, we only compare our method to RAN [18].

Our method obtained better results on most datasets. Table 1 shows that our method obtained better performance on RAF-DB and AffectNet(C7) for non-occluded facial images. Specifically, the accuracies of our method were 4.87% and 1.42% higher than those of PG-CNN and gACNN on AffectNet(C7), and 2.68% and 0.88% higher on RAF-DB, respectively. We also compared our method on synthesized occluded databases. As shown in Table 2, our method achieved better accuracy than PG-CNN, gACNN, Pan *et al.*'s method, and Xia *et al.*'s method by 6.01%, 3.52%, 2.09% and 1.32% on RAF-DB, and 5.93%, 3.56%, 1.98% and

0.94% on AffectNet(C7), respectively. Our method also achieved superior performance under four types of occlusion: eye occluded, mouth occluded, 8×8 size of occlusion and 24×24 size of occlusion on CK+.

RGBT uses hand crafted features to help facial expression recognition, which lacks generalization. WLS-RF does not use the guidance of non-occluded facial images and the entire framework is not trained end to end. PG-CNN uses an attention network to extract local features from facial regions of the convolutional feature maps. gACNN introduces the global gated unit to complement the global information of facial images, which extends PGCNN. Although these two methods use the attention mechanism to pay attention to non-occluded regions, these methods typically require facial landmarks to locate sub-regions. Pan *et al.*'s method uses non-occluded images to guide the occluded expression classifier. However, it trains two distinct networks for non-occluded and occluded images, so it cannot predict occluded and non-occluded facial emotion in a single network. Xia *et al.*'s method divides the dataset into three parts based on the difficulty of the database, and then the network is learned in three stages. The end of each stage may not be optimal and requires human control. Our method obtained competitive results and recognized occluded and non-occluded facial expressions end to end, and effectively used the occluded and non-occluded facial image information.

Finally, we compared the generalization ability of our method with that of related methods for realistic facial images. The experimental results with realistic occlusions in Table 3 demonstrate that our method outperformed most related methods. Specifically, our method achieved better accuracy than PG-CNN, gACNN and Pan *et al.*'s method by 5.75%, 3.50% and 0.25% on FED-RO, respectively. Our method also achieved 2.02% and 0.80% higher accuracy than RAN on FED-RO and Occlusion-AffectNet(C8), respectively. These results demonstrate that our method also obtained good results on realistic facial images.

5 Conclusion

In this study, we proposed an occluded expression recognition method through self-supervised learning. We designed pretext tasks related to occluded facial images. We adopted similarity loss to make the representation of facial image and those of its variations with synthesized occlusions similar, and used occlusion loss to optimize the occlusion detection. Our method achieved better results than most state-of-the-art methods on both occluded and non-occluded facial images, which demonstrates its superiority.

Acknowledgements This work was supported by National Natural Science Foundation of China 92048203 and project from Anhui Science and Technology Agency 202104h04020011.

References

1. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 67–74. IEEE (2018)
2. Cornejo, J.Y.R., Pedrini, H.: Recognition of occluded facial expressions based on centrist features. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1298–1302. IEEE (2016)
3. Dapogny, A., Bailly, K., Dubuisson, S.: Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. International Journal of Computer Vision **126**(2), 255–271 (2018)
4. Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local learning with deep and handcrafted features for facial expression recognition. IEEE Access **7**, 64827–64836 (2019)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Jing, L., Tian, Y.: Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
8. Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing (2020)
9. Li, S., Deng, W., Du, J.: 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. pp. 2852–2861 (2017)
10. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Transactions on Image Processing **28**(5), 2439–2450 (2018)
11. Li, Y., Zeng, J., Shan, S., Chen, X.: Patch-gated cnn for occlusion-aware facial expression recognition. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 2209–2214. IEEE (2018)
12. Lu, Y., Wang, S., Zhao, W., Zhao, Y.: Wgan-based robust occluded facial expression recognition. IEEE Access **7**, 93594–93610 (2019)
13. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: 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. pp. 94–101. IEEE (2010)
14. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2017)
15. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
16. Pan, B., Wang, S., Xia, B.: Occluded facial expression recognition enhanced through privileged information. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 566–573 (2019)

17. Ranzato, M., Susskind, J., Mnih, V., Hinton, G.: On deep generative models with applications to recognition. In: CVPR 2011. pp. 2857–2864. IEEE (2011)
18. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing **29**, 4057–4069 (2020)
19. Xia, B., Wang, S.: Occluded facial expression recognition with step-wise assistance from unpaired non-occluded images. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2927–2935 (2020)
20. Zhang, L., Tjondronegoro, D., Chandran, V.: Random gabor based templates for facial expression recognition in images with facial occlusion. Neurocomputing **145**(dec.5), 451–464 (2014)
21. Zhao, R., Liu, T., Xiao, J., Lun, D.P.K., Lam, K.M.: Deep multi-task learning for facial expression recognition and synthesis based on selective feature sharing. International Conference on Pattern Recognition (2020)