Occluded Facial Expression Recognition Enhanced through Privileged Information

Bowen Pan, Shangfei Wang* and Bin Xia
bowenpan@mail.ustc.edu.cn;sfwang@ustc.edu.cn;xiabin@mail.ustc.edu.cn
Key Lab of Computing and Communication Software of Anhui Province
University of Science and Technology of China
HeFei, Anhui, P.R.China

ABSTRACT

In this paper, we propose a novel approach of occluded facial expression recognition under the help of non-occluded facial images. The non-occluded facial images are used as privileged information, which is only required during training, but not required during testing. Specifically, two deep neural networks are first trained from occluded and non-occluded facial images respectively. Then the non-occluded network is fixed and is used to guide the fine-tuning of the occluded network from both label space and feature space. Similarity constraint and loss inequality regularization are imposed to the label space to make the output of occluded network converge to that of the non-occluded network. Adversarial leaning is adopted to force the distribution of the learned features from occluded facial images to be close to that from non-occluded facial images. Furthermore, a decoder network is employed to reconstruct the non-occluded facial images from occluded features. Under the guidance of non-occluded facial images, the occluded network is expected to learn better features and classifier during training. Experiments on the benchmark databases with both synthesized and realistic occluded facial images demonstrate the superiority of the proposed method to state-of-the-art.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

facial expression recognition, facial occlusion, privileged information

ACM Reference Format:
Bowen Pan, Shangfei Wang* and Bin Xia. 2019. Occluded Facial Expression Recognition Enhanced through Privileged Information. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19), Oct. 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3343031.3351049

*Dr. Shangfei Wang is the corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351049

1 INTRODUCTION

In recent years, facial expression recognition (FER) has attracted more and more attention due to its wide application in many areas such as human-computer interaction, security and psychological treatment. Although much progress has been achieved at building robust facial expression classifier [14–16], it is still very challenging to recognize expressions from facial images with occlusions.

Currently, most benchmark databases used in facial expression recognition are collected in laboratory environment , and the recorded facial images are mainly frontal without any occlusion. Classifiers that are trained on such databases do not have good generalization ability to the real scenario in the wild. Although there exist large-scale facial databases in the wild like RAF-DB [6] and AffectNet [10], the amount of training instances is still limited due to the diversity of the type and position of occlusions. The limitation of the databases prevents us from learning a robust facial expression classifier directly from the available data.

To address this issue, one solution is to construct feature representations that are robust to occlusions [17]. However, these hand-crafted feature representations are still limited and unable to generalize well on new types of occlusion. Another solution is to split the face into multiple regions and extract feature representations locally [1, 7, 8]. And then, a set of weights are learned to indicate the importance (unobstructed-ness) of facial regions. The final classifier is learned based on the weighted feature representations of facial regions. The main shortcoming of the weighted strategy is that the learned weights may be biased since no corresponding non-occluded images are used as guidance.

Although the databases of occluded facial expression are limited, several benchmark databases consisting of facial images without occlusions are available. Such data can be used to construct a non-occluded facial expression classifier. Since facial images without occlusions usually provide more information for facial expression recognition compared to occluded facial images, the non-occluded facial expression classifier can be used as the guidance to facilitate the learning process of a occluded facial expression classifier. Therefore, we propose using non-occluded facial images as privileged information [12] to assist the learning process of the occluded view. Specifically, two deep neural networks are first trained from occluded and non-occluded images respectively. During training, in order to take advantage of the guidance of the non-occluded images, we introduce similarity constraint and loss inequality regularization in label space, and apply adversarial learning and reconstructed loss in feature space. Thus, the occluded network is fully fine-tuned with the guidance of the non-occluded network. During testing, the occluded network is used to recognize expressions from facial

images with occlusions. The proposed approach makes full use of the readily available facial images without occlusions as privileged information, only needed during training, to construct better facial representations and a better expression classifier from facial images with occlusions. The facial images without occlusions are not required during testing. Thus, our approach can be used in real-world applications.

2 RELATED WORK

Since the occlusion is extremely diverse and current databases can not provide sufficient facial images with occlusions, it is intractable to learn the facial expression classifier in a data-driven manner. Many solutions are proposed to tackle the occlusion such as feature representations that are robust to occlusion and network structure that perceives the occlusion automatically.

Zhang *et al.* [17] proposed a robust approach that employs Monte Carlo algorithm to extract a set of Gabor based part-face templates and converted these templates into template match distance features. The template match distance features rely on the selection of the part-face templates on a specific database and do not have good generalization ability under the cross-database condition.

Dapogny *et al.* [1] proposed to train Random Forests upon spatially defined local subspaces of the face and use Local Expression Predictions (LEPs) as high-level representations. LEPs can be further weighted by confidence scores provided by an autoencoder network for classification. The main shortcoming of Dapogny *et al.*'s method is that the whole framework is not trained end-to-end and there is a lack of guidance of non-occluded facial images.

Recently, Li *et al.* [7] proposed Patch-Gated CNN (PG-CNN) to perceive the occlusion automatically and perform expression classification. PG-CNN extracts 24 regions of interest from the convolutional feature maps. A specific structure named Patch Gated Unit (PG-Unit) is designed to further extract local features from the region and to learn an unobstructed score by an attention net. The final classifier is built based on the weighted concatenated local features of all regions. Although PG-CNN perceives the occlusion automatically by learning unobstructed scores for multiple facial regions, unobstructed scores are learned without any ground truth of the occlusion information and may be biased. Li *et al.* [8] further improved their work by introducing Global Gated Unit (GG-Unit) to complement the global information of the facial images for expression recognition. Although their method perceives the occlusion at both patch-level and image-level and achieves the state-of-the-art performance, the shortcoming of lacking guidance from the non-occluded images still exists.

To summarize, current works only consider building classifiers from the occluded images and ignore the potential assist of non-occluded images. Therefore, in this paper, we propose a new framework for occluded facial expression recognition, which utilizes both occluded and non-occluded images effectively. Specifically, two deep neural networks are first trained from occluded and non-occluded images respectively. Non-occluded facial images are viewed as privileged information and guide the fine-tuning process of the occluded network. During training phase, similarity constraint and loss inequality regularization are imposed in label space in order to calibrate the output values of the occluded network. Adversarial learning is applied in feature space and forces the distribution of the occluded features to be close to that of the non-occluded features. A decoder network is further built upon feature representations and reconstructed loss is introduced to preserve the critical information for expression recognition. During testing phase, the learned occluded network is expected to be robust for unknown facial images with occlusions.

Compared with related work, our contribution can be summarized as follows: (1) we are the first to introduce non-occluded facial images as privileged information to guide the learning process of the occluded classifier. (2) we propose to guide the learning process of the occluded classifier in both label and feature spaces, and achieve state-of-the-art performance on occluded facial expression recognition.

3 PROBLEM STATEMENT

Let $\mathcal{D}_{\text{train}} = \left\{x_o^{(i)}, x_c^{(i)}, y^{(i)}\right\}_{i=1}^{N}$ denotes a training set of N training instances, where $x_o^{(i)}, x_c^{(i)} \in \mathbb{R}^{H\times W\times 3}$ are the facial image with occlusions and the facial image without occlusion. $y^{(i)} \in \{0, 1, \cdots, K-1\}$ represents the ground truth of the i^{th} instance. Denote test set as $\mathcal{D}_{\text{test}} = \left\{x_o^{(i)}\right\}_{i=1}^{M}$. Each test instance only contains the facial image with occlusions. Given $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, our goal is to learn a network $F_o : \mathbb{R}^{H\times W\times 3} \rightarrow \{0, 1, \cdots, K-1\}$, which gives good predictions for occluded images. It is worth to note that requiring paired occluded and non-occluded images is not mandatory. Since it is the first trial to explore the assist role of the non-occluded images, we only consider the paired situation in this paper.

4 METHODOLOGY

The framework of our proposed method is summarized in Figure 1. We adopt a strategy of learning with privileged information, in which non-occluded images are treated as privileged information and assist the learning process of the occluded classifier during training phase. During testing phase, the prediction of unknown occluded images is given by the occluded classifier.

In the following subsections, we first introduce the basic networks. Then we elaborate the guidance of privileged information in label and feature spaces respectively. Finally, we conclude the overall loss function and optimization.

4.1 Basic Networks

Two deep convolutional neural networks with the same architecture are built in our framework. The first one $\hat{y}_o = F_o(x_o; \theta_o)$ performs expression recognition from occluded images, and is the desired classifier in this paper. The second one $\hat{y}_c = F_c(x_c; \theta_c)$ performs expression recognition from non-occluded images, and is used as guidance. These two networks are first pretrained with the supervised multi-class cross entropy losses $\ell_{\text{sup}}(y, \hat{y}_o)$ and $\ell_{\text{sup}}(y, \hat{y}_c)$. After pretraining, parameters of non-occluded network are fixed and the occluded network is further fine-tuned under the guidance of non-occluded network.

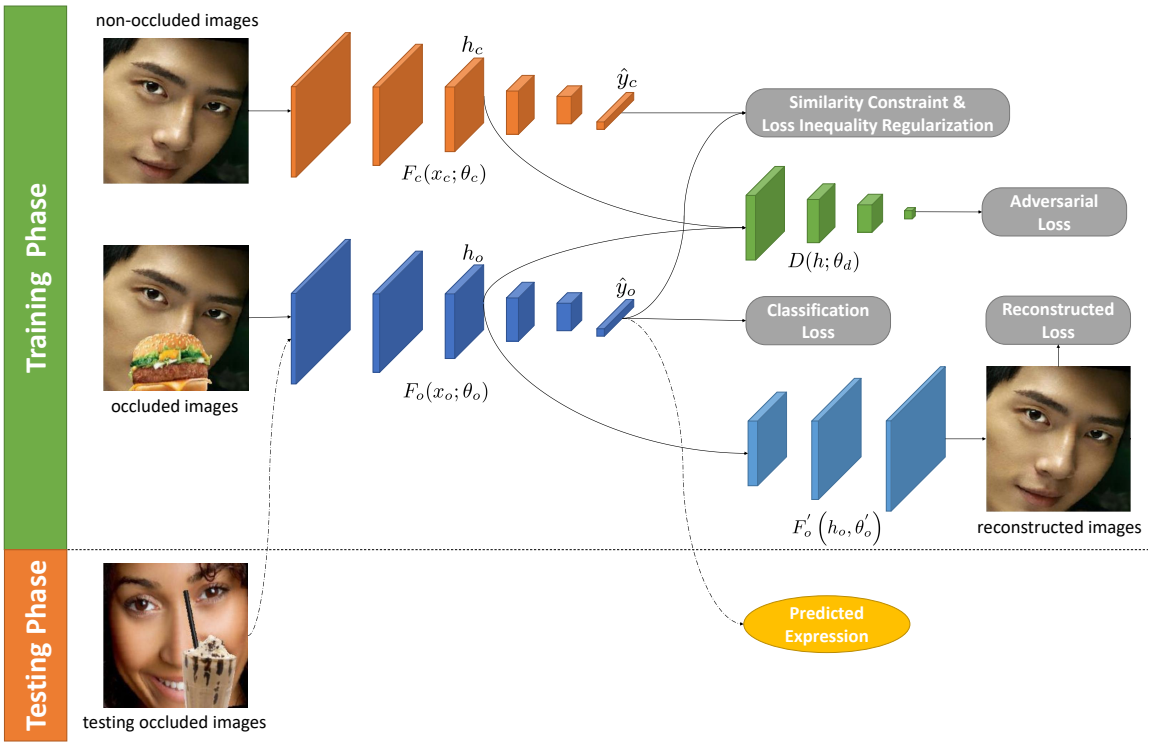

Figure 1: The framework of proposed facial expression recognition with occlusions.

4.2 Guidance in Label Space

4.2.1 Similarity Constraint. The similarity constraint was first introduced in a multi modal approach named SVM2K [3]. It forces the outputs of two modalities to be similar and large difference between them will be penalized. Here we introduce the similarity constraint as guidance and generalize it to the multi-dimensional case, leading to a similarity constraint loss as shown in Eq. 1.

$$\ell_{\text{sim}}(\hat{y}_o, \hat{y}_c) = \|\hat{y}_o - \hat{y}_c\|^2 \tag{1}$$

4.2.2 Loss Inequality Regularization. In Wang and Ji's work [13], a new regularization approach named Loss Inequality Regularization (LIR) was proposed to explore the privileged information effectively and further boost the performance of the primary classifier. The precondition of LIR is that privileged information is more discriminative than the primary features. In our case, non-occluded images are obviously more discriminative than occluded images for facial expression recognition. Then the loss of occluded image $\ell_{\text{sup}}(y, \hat{y}_o)$ should be higher than the loss of non-occluded image $\ell_{\text{sup}}(y, \hat{y}_c)$ as shown in Eq. 2.

$$\ell_{\text{sup}}(y, \hat{y}_o) \geqslant \ell_{\text{sup}}(y, \hat{y}_c) + \epsilon, \quad \epsilon \geqslant 0 \tag{2}$$

The above constraint is further formulated, leading to the loss function in Eq. 3.

$$\ell_{\text{LIR}}(y, \hat{y}_o, \hat{y}_c) = \left[\ell_{\text{sup}}(y, \hat{y}_c) - \ell_{\text{sup}}(y, \hat{y}_o)\right]_+ \tag{3}$$

where $[\cdot]_+ = \max(0, \cdot)$ is the hinge function.

4.3 Guidance in Feature Space

4.3.1 Adversarial Loss. Up to now, the guidance of privileged information merely happens in label space. There is a lack of the guidance in feature space. Therefore, the distribution of the occluded features should be close to that of the non-occluded features. Fortunately, generative adversarial networks proposed by Goodfellow *et al.* [4] provide us a mean to achieve this goal. Concretely, we consider the feature maps $h_o = \tilde{F}_o(x_o; \tilde{\theta}_o)$ and $h_c = \tilde{F}_c(x_c; \tilde{\theta}_c)$, which are taken from the same layer of the occluded and non-occluded networks. We view h_o as fake features and h_c as real features. Next, we introduce a discriminator $p = D(h; \theta_d)$, where h is a feature vector from either occluded view or non-occluded view and p is the probability that h comes from the non-occluded view. The learning objective of the discriminator is to classify the source of feature vectors accurately while the occluded network tries to fool the discriminator. Therefore, we get the following objective:

$$\begin{aligned}\min_{\tilde{\theta}_o} \max_{\theta_d} \Big[& \mathbb{E}_{h_c \sim \mathrm{P}(h_c)} \log D_{\theta_d}(h_c) + \\ & \mathbb{E}_{h_o \sim \mathrm{P}(h_o)} \log\left(1 - D_{\theta_d}(h_o)\right) \Big]\end{aligned} \tag{4}$$

As elaborated in [4], Eq. 4 can not be optimized directly. We adopt an alternating optimization between the discriminator and the occluded network. We decouple the minmax objective and obtain the learning objective of the occluded network and discriminator. The learning objectives of these two components are reformulated into loss functions. The loss function of the discriminator is shown

in Eq. 5.

$$\ell_d(\theta_d) = -\log D(h_c) - \log(1 - D(h_o)) \tag{5}$$

Follow the suggestion in [4], it is better for the occluded network to minimize $-\log D(h_o)$ instead of minimizing $\log(1 - D(h_o))$ in order to avoid flat gradients. The loss function of the occluded network is shown in Eq. 6.

$$\ell_{\text{adv}}(\tilde{\theta}_o) = -\log D(h_o) \tag{6}$$

4.3.2 Reconstructed Loss. Although adversarial loss is imposed to ensure the statistical similarity between learned occluded features and non-occluded features, it is still unclear that whether the learned occluded features contain the crucial facial information for expression recognition. One of the most important properties of the occluded features is that we can recover the corresponding non-occluded images from the learned occluded features. To achieve this, we build a decoder network $\hat{x} = F_o^{'}\left(h_o, \theta_o^{'}\right)$ upon the occluded feature representations as a brunch of the occluded network. We introduce the reconstructed loss and hope that the decoder network can recover non-occluded images from occluded images. Concretely, the reconstructed loss is defined as the mean squared error between the reconstructed image and the corresponding non-occluded image as shown in Eq. 7.

$$\ell_{\text{rec}}(\hat{x}, x_c) = \|\hat{x} - x_c\|^2 \tag{7}$$

4.4 Overall Loss Function

We aggregate the aforementioned guidance of privileged information including Eq. 1, 3, 6 and 7, and define the overall loss function of the occluded network as follows:

$$\mathcal{L}(\theta_o) = \ell_{\text{sup}} + \lambda_1 \ell_{\text{sim}} + \lambda_2 \ell_{\text{LIR}} + \lambda_3 \ell_{\text{adv}} + \lambda_4 \ell_{\text{rec}} \tag{8}$$

where λ_1, λ_2, λ_3 and λ_4 are hyper-parameters, which balances the tradeoff among all of the related losses.

Among these losses, ℓ_{LIR} and ℓ_{rec} require paired occluded and non-occluded images. ℓ_{sim} only requires occluded and non-occluded images with the same emotion category, and ℓ_{rec} can utilize complete unpaired images. The flexibility of ℓ_{sim} and ℓ_{adv} make our method applicable in practice.

4.5 Optimization

The occluded network plays the role of "generator" in the proposed framework. Therefore, the optimization procedure is similar to the original GAN framework [4]. Algorithm 1 outlines the learning procedure of the proposed method.

Algorithm 1 The learning algorithm of the proposed method.

Require: The training set $\mathcal{D}_{\text{train}} = \left\{x_o^{(i)}, x_c^{(i)}, y^{(i)}\right\}_{i=1}^{N}$, the number of steps of updating discriminators K_1, the number of steps of updating occluded network K_2, learning rate of the discriminators η_1, learning rate of the occluded network η_2, batch size m and the number of training epochs T.

Ensure: The occluded network F_o.

1: Pretrain the occluded network F_o with $\left\{x_o^{(i)}, y^{(i)}\right\}_{i=1}^{N}$.
2: Pretrain the non-occluded network F_c with $\left\{x_c^{(i)}, y^{(i)}\right\}_{i=1}^{N}$.
3: Fix the parameter θ_c.
4: **for** $t = 1$ to T **do**
5: **for** $k = 1$ to K_1 **do**
6: Randomly sample a mini-batches of occluded and non-occluded images $\left\{x_o^{(i)}, x_c^{(i)}\right\}_{i=1}^{m}$ from $\mathcal{D}_{\text{train}}$.
7: Perform forward propagation with $\left\{x_o^{(i)}, x_c^{(i)}\right\}_{i=1}^{m}$ and fetch $\left\{h_o^{(i)}, h_c^{(i)}\right\}_{i=1}^{m}$.
8: Update D by $\theta_d := \theta_d - \eta_1 \frac{\partial \ell_d(\theta_d)}{\partial \theta_d}$.
9: **end for**
10: **for** $k = 1$ to K_2 **do**
11: Randomly sample a mini-batches of occluded and non-occluded images $\left\{x_o^{(i)}, x_c^{(i)}\right\}_{i=1}^{m}$ from $\mathcal{D}_{\text{train}}$.
12: Update F_o by $\theta_o := \theta_o - \eta_2 \frac{\partial \mathcal{L}(\theta_o)}{\partial \theta_o}$.
13: **end for**
14: **end for**

5 EXPERIMENT

5.1 Experimental Condition

To the best of our knowledge, there is only one facial expression database with realistic occlusions, i.e., FED-RO database [8]. Occluded images should be synthesised on other facial expression databases. Follow the experimental condition in [8], both within-database and cross-database experiments are conducted on seven benchmark databases including RAF-DB [6], AffectNet [10], CK+ [9], MMI [11], Oulu-CASIA [18], SFEW [2] and FED-RO. Among these databases, experiments on two databases (MMI, Oulu-CASIA and SFEW) can not be conducted because selection of peak frames is not public. Therefore, we conduct experiments on the remaining four benchmark databases. The details are described as follows.

The Real-world Affective Faces Database (RAF-DB) is a facial expression database in the wild with around 30K images which are labeled by 40 annotators. In our experiment, we only use images of 7 classes of basic emotions, containing 12,271 images as training set and 3,068 images as test set.

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 seven discrete facial expressions and the intensity of valence and arousal. In our experiment, only images with basic emotions are used, including 280,000 images as training set and 3500 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. In our experiment, we collect onset frames labelled as neutral category and apex frames with six basic emotions. As a result, 636 facial images are collected and 10-fold cross validation strategy is adopted.

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.* in order to evaluate their proposed gACNN method [8]. In total, 400 facial

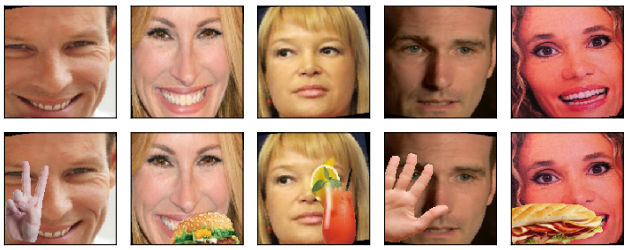

Figure 2: Examples of the synthesized occluded facial images on the AffectNet database.

images with various occlusion are downloaded from the Internet and labelled with seven emotions by three people. Since the size of the FED-RO database is not large, it is only used for cross-database evaluation.

To mimic the scenario in real world, occluded images are artificially synthesised by adding occluding objects at random locations of the faces on all databases except the FED-RO database. The occluding objects we used include food, hands and drinks. Each kind of occluding object has different templates. Figure 2 displays some examples of occluded images on the AffectNet database using this approach. The original facial image and its paired synthesized occluded image make up one training instance. This paired condition is not mandatory.

In our experiments, facial images are resized to 88×88 on the RAF-DB database, 224×224 on the AffectNet database, 48×48 on the CK+ database and 144×144 on the SFEW database. We use ResNet-50 [5] as the architecture of the occluded and non-occluded networks. The last layer of 1000 units in the original ResNet-50 is replaced by the fully connected layers with 7 units. The pretrained weights of ResNet-50 on Imagenet are used for initialization. We build a five-layer CNN as discriminator upon the hidden feature space corresponding to the feature map with a size of $28 \times 28 \times 512$. A decoder network with three deconvolutional layers is also built upon this feature space. The learning rates of the occluded network and discriminator are 2e-5 and 1e-4 respectively. Four hyper-parameters in loss function are determined by grid search.

We conduct experiments on all databases using eight methods. Firstly, a standard ResNet-50 is trained as baseline using only occluded images. Another ResNet-50 is also trained using non-occluded images and is fixed as guidance later. Secondly, the methods with ℓ_{sim}, ℓ_{LIR} and both of them in label space are trained for comparison. Similarly, the methods with ℓ_{adv}, ℓ_{rec} and both of them in feature space are trained. Finally, our proposed which combines ℓ_{sim}, ℓ_{LIR}, ℓ_{adv} and ℓ_{rec} is trained. The overall accuracy on seven facial expression categories is used as a performance metric.

We also conduct experiemnts to compare the inference speed of our method against the state-of-the-art works i.e., PG-CNN [7], gACNN [8] on AffectNet dataset. The inference of all methods is completed on a NVIDIA TeslaV100 GPU with 32GB memory. During the inference stage, we set the batch size as 8. PG-CNN and aACNN are implemented in caffe, our method is implemented in pytorch, so this experiment is only as a reference. We use frame per seond (Fps) as speed performance metric.

5.2 Experimental Results and Analysis

5.2.1 Analysis of facial expression recognition with synthesized occlusions. The experimental results of facial expression recognition with synthesized occlusions are shown in Table 1 and 2. From the tables, we can find the following observations:

First, adopting one of the four introduced losses leads to an improvement comparing with the baseline using ResNet. Specifically, the accuracies of ResNet + ℓ_{sim}, ResNet + ℓ_{LIR}, ResNet + ℓ_{adv} and ResNet+ ℓ_{rec} are 2.87%, 2.74%, 3.65% and 3.19% higher than that of the ResNet on the RAF-DB database. The experimental results on the AffectNet and CK+ databases show similar trend. The experimental results also demonstrate that the adversarial loss is more effective than similarity constraint, LIR and reconstructed loss, which is 0.78%, 0.91%, 0.46% higher than using only similarity constraint, LIR and reconstructed loss. The adversarial loss uses non-occluded images as privileged information to close the statistical gap between the occluded images and non-occluded images in feature level and trains a more effective classifier.

Second, our proposed method achieves the best performance by using similarity constraint, LIR, adversarial loss and reconstructed loss together. Specifically, the accuracy of our method is 1.56%, 1.69%, 0.78% and 1.24% higher than those of ResNet + ℓ_{sim}, ResNet + ℓ_{LIR}, ResNet + ℓ_{adv} and ResNet+ ℓ_{rec} on the RAF-DB database. Our method utilizes the advantages of four introduced losses at the same time to improve the accuracy of occluded facial expression recognition. Guidance in both feature and label spaces can assist the occluded classifier in learning more robust feature representations and making better predictions.

Third, guidance in feature space is more effective than guidance in label space in most cases. For example, the accuracy of ResNet + ℓ_{adv} + ℓ_{rec} is 0.78% higher than that of ResNet + ℓ_{sim} + ℓ_{LIR} on the RAF-DB database. Since the dimension of feature space is larger than that of label space, the adversarial loss and reconstructed loss play a more significant role than similarity constraint and LIR. Moreover, adversarial loss and reconstructed loss act earlier than similarity constraint and LIR. Thus the occluded features learned with adversarial loss and reconstructed loss lead to a better classifiers.

Fourth, we evaluate the generalization ability of our proposed method by cross-database evaluation, where our proposed model is trained on the RAF-DB or AffectNet database and tested on the CK+ database with synthesized occluded images. The method of ResNet + ℓ_{sim} + ℓ_{LIR} achieves the best performance on the RAF-DB/CK+ database, which is 2.02% higher than ResNet. The method of ResNet+ℓ_{sim}+ℓ_{LIR}+ℓ_{adv}+ℓ_{rec} achieves the best performance on the AffectNet/CK+ databases, which are 1.69% higher than ResNet. From Table 1, we can find that the similarity constraint degrades the performance slightly. It may be caused by the difficulty of choosing a suitable distance metric.

5.2.2 Analysis of facial expression recognition with realistic occlusions. The experimental results of facial expression recognition with realistic occlusions are shown in Table 3. We merged all training set of RAF-DB and AffectNet databases for training, and perform

Table 1: Experimental results of facial expression recognition with synthesized occlusions on the RAF-DB, AffectNet, CK+ and SFEW databases. († denotes models trained on the RAF-DB database. ‡ denotes models trained on the AffectNet database.)

Methods	RAF-DB	AffectNet	†CK+	‡CK+
PG-CNN [7]	78.05	52.47	79.49	86.27
gACNN [8]	80.54	54.84	79.49	88.17
ResNet	77.54	54.54	78.46	88.21
ResNet + ℓ_{sim}	80.41	56.02	77.67	88.05
ResNet + ℓ_{LIR}	80.28	55.11	78.47	88.52
ResNet + ℓ_{sim} + ℓ_{LIR}	80.57	56.09	75.79	89.62
ResNet + ℓ_{adv}	81.19	55.94	79.09	88.68
ResNet + ℓ_{rec}	80.73	55.77	80.03	89.46
ResNet + ℓ_{adv} + ℓ_{rec}	81.35	56.11	**80.66**	89.62
Ours: ResNet+ℓ_{sim}+ℓ_{LIR}+ℓ_{adv}+ℓ_{rec}	**81.97**	**56.42**	79.21	**89.90**

Table 2: 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, 24×24 respectively.)

Methods	R8	R16	R24	eye occluded	mouth occluded
RGBT [17]	92.0	82.0	62.5	88.0	30.3
WLS-RF [1]	92.2	86.4	74.8	87.9	72.7
PG-CNN [7]	96.58	95.70	92.86	96.50	**93.92**
gACNN [8]	96.58	95.97	**94.82**	96.57	93.88
ResNet	95.91	95.13	92.30	95.13	92.30
ResNet + ℓ_{sim}	96.23	95.44	92.61	95.28	92.77
ResNet + ℓ_{LIR}	96.70	95.59	92.77	95.44	92.45
ResNet + ℓ_{sim} + ℓ_{LIR}	97.48	96.54	92.92	95.91	92.92
ResNet + ℓ_{adv}	96.54	95.91	93.08	95.91	92.61
ResNet + ℓ_{rec}	95.75	95.13	92.30	95.60	92.45
ResNet + ℓ_{adv} + ℓ_{rec}	97.17	96.38	93.40	96.38	93.08
Ours: ResNet+ℓ_{sim}+ℓ_{LIR}+ℓ_{adv}+ℓ_{rec}	**97.80**	**96.86**	94.03	**96.86**	93.55

Table 3: Experimental results of facial expression recognition with realistic occlusions on the FED-RO database. The training set on the RAF-DB and AffectedNet databases are merged for training.

Methods	Accuracy(%)
PG-CNN [7]	64.25
gACNN [8]	66.50
ResNet	64.75
ResNet + ℓ_{sim}	67.25
ResNet + ℓ_{LIR}	67.75
ResNet + ℓ_{sim} + ℓ_{LIR}	68.00
ResNet + ℓ_{adv}	68.75
ResNet + ℓ_{rec}	68.25
ResNet + ℓ_{adv} + ℓ_{rec}	69.25
Ours: ResNet+ℓ_{sim}+ℓ_{LIR}+ℓ_{adv}+ℓ_{rec}	**69.75**

PG-CNN [7] is identical to pACNN in [8].

cross-database evaluation on the FED-RO database. From Table 3, we can find the following observations:

First, our proposed method outperforms other methods in terms of classification accuracy. Specifically, the accuracy of our method is 5.00%, 1.75% and 0.50% than ResNet, ResNet + ℓ_{sim} + ℓ_{LIR} and ResNet + ℓ_{adv} + ℓ_{rec} respectively. We also find that the combination of adversarial loss and reconstructed loss used in feature space plays a more important role than the combination of similarity constructed loss and LIR loss used in label space. Such observations are consistent with experiments under synthesized occlusions.

Second, we also investigate per expression category classification performance on the FED-RO database. The confusion matrix based on our method is shown in Figure 3. As can be seen, our method achieves the highest and lowest accuracies on happiness and disgust category respectively.

5.2.3 Analysis of facial expression recognition with realistic occlusions. The experimental results of facial expression recognition with realistic occlusions are shown in Table 3. We merged all training set of RAF-DB and AffectNet databases for training, and perform cross-database evaluation on the FED-RO database. From Table 3, we can find the following observations:

First, our proposed method outperforms other methods in terms of classification accuracy. Specifically, the accuracy of our method is 5.00%, 1.75% and 0.50% than ResNet, ResNet + ℓ_{sim} + ℓ_{LIR} and ResNet + ℓ_{adv} + ℓ_{rec} respectively. We also find that the combination of adversarial loss and reconstructed loss used in feature space

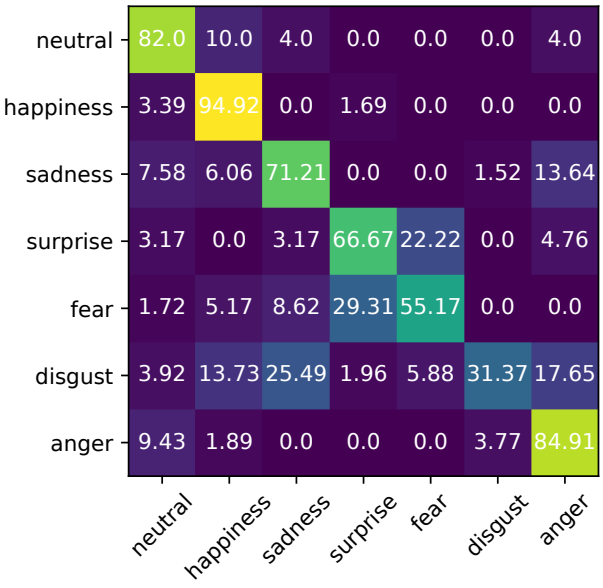

Figure 3: Confusion matrix on the FED-RO database. (Row indexes represent the ground truth labels while column indexes represent the predictions.)

Table 4: Inference speed comparison results of occluded facial expression recognition.

Methods	Fps
PG-CNN [7]	1.33
gACNN [8]	1.19
Ours: ResNet+ℓ_{sim}+ℓ_{LIR}+ℓ_{adv}+ℓ_{rec}	**155.3**

plays a more important role than the combination of similarity constructed loss and LIR loss used in label space. Such observations are consistent with experiments under synthesized occlusions.

Second, we also investigate per expression category classification performance on the FED-RO database. The confusion matrix based on our method is shown in Figure 3. As can be seen, our method achieves the highest and lowest accuracies on happiness and disgust category respectively.

5.3 Comparison with Related Works

Previous methods like RGBT [17] and WLS-RF [1] only conducted experiment on the CK+ database as shown in Table 2. For the state-of-the-art works i.e., PG-CNN [7], gACNN [8], we compare our method with them on four databases as shown in Table 1, 2 and 3.

As shown in the Table 1, our method achieves better performance on the RAF-DB and AffectNet databases. Specifically, the accuracy of our method is 3.92% and 1.43% higher than those of PG-CNN and gACNN on the RAF-DB database, and is 3.95% and 1.58% higher than those of PG-CNN and gACNN on the AffectNet database. On the CK+ database, our method achieves better

Figure 4: Comparion between occluded facial images and reconstructed facial images on the RAF-DB database. The first row represents original non-occluded images. The second row represents synthesized occluded images. The third row represents the outputs of decoder network.

accuracies than those of RGBT and WLS-RF among all types of occlusions. Our method also achieves better accuracies than those of PG-CNN and gACNN under three types of occlusions. RGBT adopts limited hand-crafted features based on the templates of a specific database. The whole framework of WLS-RF is not trained in an end-to-end manner and there is a lack of guidance of non-occluded facial images. PG-CNN exploits a specific structure names Patch Gated Unit to extract local features from facial regions of the convolutional feature maps. gACNN exceeds PG-CNN by introducing Global Gated Unit to complement the global information of facial images for FER. These two methods also ignore the potential assist of non-occluded images totally. While our method uses original non-occluded images as privileged information to construct better occluded feature representations and a more robust classifier. Therefore, better performance can be achieved by our method during testing phase.

We also compare the generalization ability of our method with related works by cross-database experiment. Experimental results with realistic occlusions in Table 3 show that our method outperforms related works. Specifically, our method achieves better accuracy than PG-CNN and gACNN by 5.5% and 3.25% respectively. It demonstrates that the occluded feature learned by using non-occluded image as privileged information during training phase can contain more important information for facial expression recognition with realistic occlusions.

We compare the inference speed of our method with the state-of-the-art works. As shown in the Table 4, our method achieves faster inference speed than PG-CNN and gACNN. Specifically, our method is 116.7 times faster than PG-CNN and 130.5 times faster than gACNN respectively. In the train and test stage, PG-CNN and gACNN all use 24 Patch Gated Unit to extract local features from 24 local regions which is time-consuming. However our method only use pretrained standard ResNet-50 to infer images which greatly reduces the testing time. Because our method, gACNN and PG-CNN use differnent deep learning framework, this experimentt is only as a reference.

5.4 Evaluation of Reconstructed Loss

As elaborated in Section 4.3.2, we introduce a decoder network which is built upon the feature space and uses reconstructed loss in the overall loss. It is a very crucial property of occluded features that we can restore it to corresponding non-occluded images. To further evaluate the effectiveness of the reconstructed loss, we visualize the outputs of the decoder network. Figure 4 displays the visualization of the original non-occluded images, occluded images and reconstructed images on the RAF-DB database. As can be seen, although the reconstructed images look somewhat blurry, we can still recognise the expression category from them. It demonstrates that the learned occluded features contain the crucial facial information for facial expression recognition.

6 CONCLUSION

In this paper, we propose an occluded facial expression recognition method that utilizes non-occluded images as privileged information to enhance the occluded classifier. Two deep neural networks are first trained with occluded and non-occluded images. The pre-trained non-occluded network is fixed and used as guidance to make full use of the non-occluded images. During the fine-tune process of the occluded network, the similarity constraint and loss inequality regularization are introduced in label space. The adversarial learning and a decoder network are also introduced to further regularize the learned occluded features. Our method achieves the state-of-the-art performance on the task of occluded facial expression recognition on four benchmark databases, demonstrating that our method can leverage the non-occluded images effectively and enhance the performance of occluded facial expression recognition.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 91748129, and Grant 61727809, and in part by the Project from Anhui Science and Technology Agency under Grant 1804a09020038.

REFERENCES

[1] Arnaud Dapogny, Kevin Bailly, and Séverine Dubuisson. 2018. Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. *International Journal of Computer Vision* 126, 2-4 (2018), 255–271.

[2] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2106–2112.

[3] Jason Farquhar, David Hardoon, Hongying Meng, John S Shawe-taylor, and Sandor Szedmak. 2006. Two view learning: SVM-2K, theory and practice. In *Advances in neural information processing systems*. 355–362.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

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

[6] Shan Li, Weihong Deng, and JunPing Du. 2017. 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*. 2852–2861.

[7] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2018. Patch-Gated CNN for occlusion-aware facial expression recognition. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2209–2214.

[8] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2019. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing* 28, 5 (2019), 2439–2450.

[9] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. 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, 94–101.

[10] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985* (2017).

[11] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. 2005. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*. IEEE, 5–pp.

[12] Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural networks* 22, 5-6 (2009), 544–557.

[13] Ziheng Wang and Qiang Ji. 2015. Classifier learning with hidden information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4969–4977.

[14] Huiyuan Yang, Umur Ciftci, and Lijun Yin. 2018. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2168–2177.

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

[16] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. 2018. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3359–3368.

[17] Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran. 2014. Random Gabor based templates for facial expression recognition in images with facial occlusion. *Neurocomputing* 145 (2014), 451–464.

[18] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. 2011. Facial expression recognition from near-infrared videos. *Image and Vision Computing* 29, 9 (2011), 607–619.