Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism

Yong Li, *Student Member, IEEE*, Jiabei Zeng, *Member, IEEE*, Shiguang Shan, *Member, IEEE*, and Xilin Chen, *Fellow, IEEE*

***Abstract*— Facial expression recognition in the wild is challenging due to various unconstrained conditions. Although existing facial expression classifiers have been almost perfect on analyzing constrained frontal faces, they fail to perform well on partially occluded faces that are common in the wild. In this paper, we propose a convolution neutral network (CNN) with attention mechanism (ACNN) that can perceive the occlusion regions of the face and focus on the most discriminative un-occluded regions. ACNN is an end-to-end learning framework. It combines the multiple representations from facial regions of interest (ROIs). Each representation is weighed via a proposed gate unit that computes an adaptive weight from the region itself according to the unobstructedness and importance. Considering different RoIs, we introduce two versions of ACNN: patch-based ACNN (pACNN) and global–local-based ACNN (gACNN). pACNN only pays attention to local facial patches. gACNN integrates local representations at patch-level with global representation at image-level. The proposed ACNNs are evaluated on both real and synthetic occlusions, including a self-collected facial expression dataset with real-world occlusions, the two largest in-the-wild facial expression datasets (RAF-DB and AffectNet) and their modifications with synthesized facial occlusions. Experimental results show that ACNNs improve the recognition accuracy on both the non-occluded faces and occluded faces. Visualization results demonstrate that, compared with the CNN without Gate Unit, ACNNs are capable of shifting the attention from the occluded patches to other related but unobstructed ones. ACNNs also outperform other state-of-the-art methods on several widely used in-the-lab facial expression datasets under the cross-dataset evaluation protocol.**

***Index Terms*— Facial expression recognition, occlusion, CNN with attention mechanism, gate unit.**

I. Introduction

FACIAL expression recognition (FER) has received significant interest from computer scientists and psychologists over recent decades, as it holds promise to an abundance of applications, such as human-computer interaction, affect analysis, and mental health assessment. Although many facial expression recognition systems have been proposed and implemented, majority of them are built on images captured in controlled environment, such as CK+ [1], MMI [2], Oulu-CASIA [3], and other lab-collected datasets. The controlled faces are frontal and without any occlusion. The FER systems that perform perfectly on the lab-collected datasets, are probable to perform poorly when recognizing human expressions under natural and un-controlled conditions. To fill the gap between the FER accuracy on the controlled faces and uncontrolled faces, researchers make efforts on collecting large-scale facial expression datasets in the wild [4], [5]. Despite the usage of data from the wild, facial expression recognition is still challenging due to the existence of partially occluded faces. It is non-trivial to address the occlusion issue because occlusions varies in the occluders and their positions. The occlusions may caused by hair, glasses, scarf, breathing mask, hands, arms, food, and other objects that could be placed in front of the faces in daily life. These objects may block the eye, mouth, part of the cheek, and any other part of the face. The variability of occlusions cannot be fully covered by limited amounts of data and will inevitably lead the recognition accuracy to decrease.

To address the issue of occlusion, we propose a Convolution Neural Network with attention mechanism (ACNN), mimicing the way that human recognize the facial expression. Intuitively, human recognizes the facial expressions based on certain patches of the face. When some regions of the face are blocked (e.g., the lower left cheek), human may judge the expression according to the symmetric part of face (e.g., the lower right cheek), or other highly related facial regions (e.g., regions around the eyes or mouth). Inspired by the intuition, ACNN automatically perceives the blocked facial patches and pays attention mainly to the unblocked and informative patches. Fig. 1 illustrates the main idea of the proposed method. Each Gate Unit in ACNN learns an adaptive weight by the unobstructed-ness or importance. As can be seen in Fig. 1, the last three visualized patches are blocked by the baby's hand and thus they have low unobstructed-ness (α_p). Then, the weighed representations are concatenated and used in the

Manuscript received May 28, 2018; revised September 27, 2018 and November 11, 2018; accepted December 5, 2018. Date of publication December 14, 2018; date of current version February 13, 2019. This work was partially supported by National Key R&D Program of China (grant NO.2017YFB1002802), Natural Science Foundation of China (grants 61702481 and 61702486), and External Cooperation Program of CAS (grant GJHZ1843). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaochun Cao. *(Corresponding author: Jiabei Zeng.)*

Y. Li and X. Chen are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yong.li@vipl.ict.ac.cn; xlchen@ict.ac.cn).

J. Zeng is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jiabei.zeng@vipl.ict.ac.cn).

S. Shan is with the Key Laboratory of Intelligent Information Processing, Center for Excellence in Brain Science and Intelligence Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: sgshan@ict.ac.cn).

Digital Object Identifier 10.1109/TIP.2018.2886767

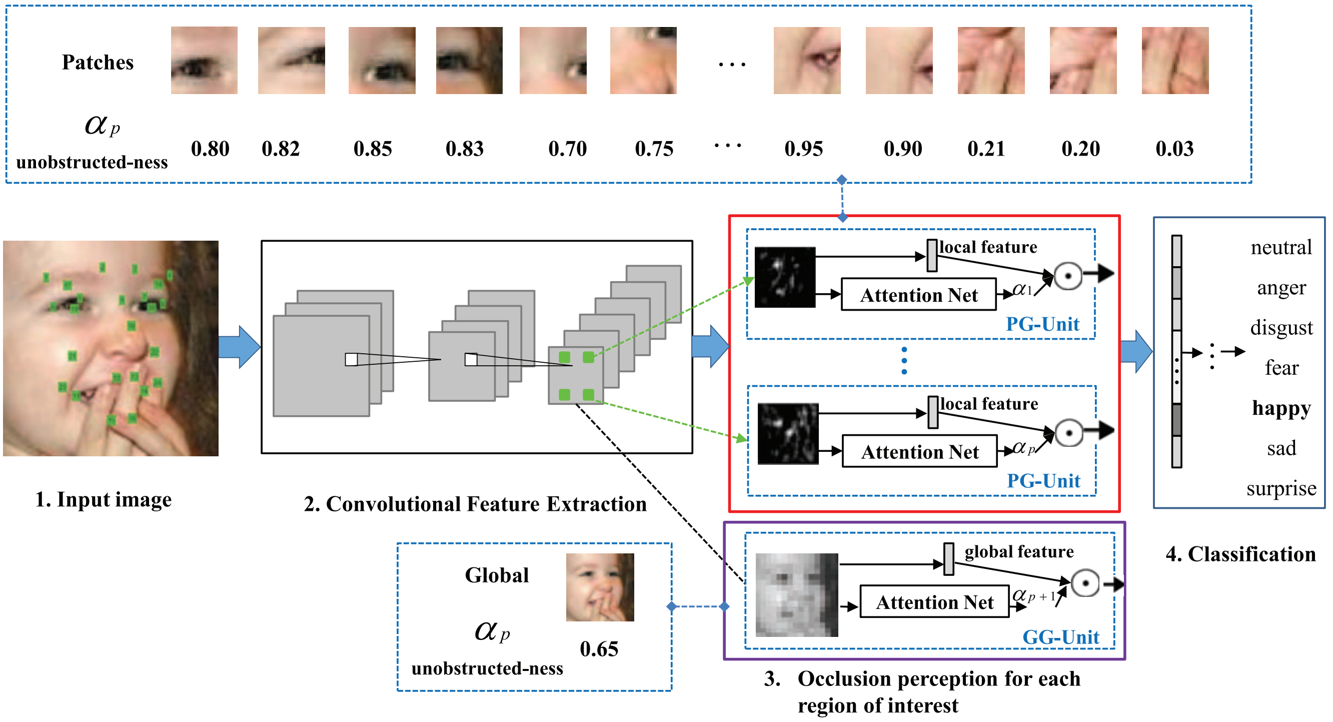


Fig. 1. Illustration of the proposed ACNN for occlusion-aware facial expression recognition. ACNN can be categorized in two versions: pACNN and gACNN. During Part 3, pACNN extracts 24 regions of interest from the intermediate feature maps. Then as can be seen in red rectangle, for each patch region, a specific Patch-Gated Unit (PG-Unit) is learnt to weigh the local representations according to the region's "*unobstructed-ness*" (to what extent the patch is occluded). Then, the weighed representations are concatenated and passed to the classification part. gACNN integrates weighed local representations with global representation (purple rectangle). The global representation is encoded and weighed via a Global-Gated Unit (GG-Unit).

classification part. Thus ACNN is able to focus on distinctive as well as unobstructed regions in facial image.

Considering different facial regions of interest, we propose two versions of ACNN: (1) pACNN crops patches of interest from the last convolution feature maps according to the positions of the related facial landmarks. Then for each patch, a Patch-Gated Unit (PG-Unit) is learned to weigh the patch's local representation by its unobstructed-ness that is computed from the patch itself. (2) gACNN integrates local and global representations concurrently. Besides local weighed features, a Global-Gated Unit (GG-Unit) is adopted in gACNN to learn and weigh the global representation.

A preliminary version of this work appeared as [6]. In this paper, we provide technical details of facial region decomposition, present extended results with more comparisons and on more datasets, and release a facial expression dataset in the presence of real occlusions. The contributions of this work are summarized as follows:

1) We propose a convolutional neural network with attention mechanism (ACNN) to recognize facial expressions from partially occluded faces. ACNN can automatically perceives the occluded regions of the face and focus on the most informative and un-blocked regions.
2) Visualized results show that Gate-Unit (the crucial part of ACNN) is effective in perceiving the occluded facial patches. For pACNN, PG-Unit is capable of learning a low weight for a blocked region and a high weight for an unblocked and informative one. With the integration of PG-Unit and GG-Unit, gACNN gains further improvement on FER performance under occlusions.
3) Experimental results demonstrate the advantages of the proposed ACNNs over other state-of-the-art methods on two large in-the-wild facial expression datasets and several popular in-the-lab datasets, under the settings with either partially occluded or non-occluded faces.
4) We collected and labelled a facial expression dataset in the presence of real occlusions (FED-RO). To the best of our knowledge, It is the first facial expression dataset in the presence of real occlusions.

II. Related Work

We review the previous work considering two aspects that are related to ours, i.e., the similar tasks (facial analysis with occluded faces) and related techniques (attention mechanism).

A. Methods Towards Facial Occlusions

For facial analysis tasks, occlusion is one of the inherent challenges in the real world facial expression recognition and other facial analysis tasks, e.g., facial recognition, age estimate, gender classification, etc. Previous approaches that address facial occlusions can be classified into two categories: holistic-based or part-based methods.

Holistic-based approaches treat the face as a whole and do not explicitly divide the face into sub-regions. To address the occlusion problems, they usually improve the robustness of the features through designated regularization, e.g., L_1-norm [7]. This idea is also suitable for non-facial occlusions, for example, Osherov and Lindenbaum [8] proposed to mutually re-weight L_1 regularization in an end-to-end framework to deal with arbitrary occlusions in object recognition. Another holistic way is to learn a generative model that can reconstruct a complete face from the occluded one [9]. The generative methods rely on the training data with varied occlusion conditions. Specially, Kotsia *et al.* [10] analysed how partial occlusions affects FER performance in detail and concluded that in general, mouth occlusion causes a greater decrease in FER than the equivalent eyes one.

Part-based methods explicitly divide the face into several overlapped or non-overlapped segmentations. To determine the patches on the face, existing works either divide the facial image into several uniform parts ([11]–[13]), or get the patches around the facial landmarks ([14], [15]), or get the patches by a sampling strategy [16], or explicitly detect the occluders [13], [17], [18]. Then, the part-based methods detect and compensate the missing part ([18]–[21]), or re-weight the occluded and non-occluded patches differently [13], [14], or ignore the occluded parts ([16], [17]).

ACNNs differ from previous part or holistic based methods in two ways. One, ACNNs need not explicitly handle occlusions (e.g., detecting), which avoid propagating detecting/inpainting error afterwards. Two, ACNNs unify representation learning and occlusion patterns encoding in an end to end CNN. The two tasks promote mutually during training, while previous methods usually depend on two or more discrete steps.

It is worth mentioning that previous work on person reidentification ([22]) gender classfication ([23]) and FER ([24]) has also adopted methods of unifying global and local features. gACNN differs from these methods by embedding an end to end trainable Gate Unit. The Gate Unit can not only learn occlusion patterns from data and encode them with model weights, but also weigh different patches for image without occlusion. Thus ACNNs are expected to achieve better FER performance on both occluded and non-occluded facial images.

B. CNN With Attention

Human has the ability to orientate rapidly towards salient objects in a cluttered visual scene [25], i.e., we are able to direct our gaze rapidly toward objects of interest in the scene. Recently this kind of attention mechanism has been successfully applied in many computer vision tasks, including fine-grained image recognition [26], image caption [27], person re-identification [28], visual question answering [29]. Usually attention can be modeled as a region sequence in an image. An RNN/LSTM model is adopted to predict the next attention region based on current attention region's location with visual features. References [29] and [27] employed this framework for visual question answering and image caption respectively.

Moreover, Zheng *et al.* [26] adopted channel grouping subnetwork to cluster different convolutional feature maps into part groups according to peak responses of maps, which do not need part annotations but is not suitable for FER in the presence of arbitrary occlusions. For false responses caused by occluders will inevitably disturb channels clustering. Zhao *et al.* [28] estimated multiple 2-dimensional attention maps, they have equal spatial size of convolutional feature maps to weight. This approach is straightforward but does not take occlusion patterns into consideration. Juefei-Xu *et al.* [30] adopted training images of multiple blur levels to enforce the attention shift during the learning process. The progressively trained model is robust to occlusions for gender classification. For face recognition in [31], a facial image is firstly partitioned into blocks, then a spatial attention control strategy over the blocks is learned through reinforcement learning.

Attention models allow for salient features to dynamically come to forefront as needed. This is especially beneficial when there are some occlusions or clutter in an image. They also help interpret the results by visualizing where the model attends to for certain tasks. Compared with existing attention models, our approach adopts facial landmarks for region decomposition, which is straightforward and easily implemented. Meanwhile, ACNNs adopt CNN based Gate Unit for occlusions perception and encoding, guiding the model to shift attention to informative as well as unblocked facial regions.

III. Proposed Method

A. Framework Overview

We propose a convolutional neural network with attention mechanism (ACNN) for facial expression recognition with partial occlusions. To address the occlusion issue, ACNN endeavours to focus on different regions of the facial image and weighs each region according to its obstructed-ness (to what extent the patch is occluded) as well as its contribution to FER.

Figure 2 illustrates the framework of the proposed ACNN. As can be seen in Fig. 2, the network takes a facial image as input. The image is fed into a convolutional net (VGG) and is represented as some feature maps. Then, ACNN decomposes the feature maps of the whole face to multiple sub-feature-maps to obtain diverse local patches. Each local patch is encoded as a weighed vector by a Patch-Gated Unit (PG-Unit). A PG-Unit computes the weight of each patch by an Attention Net, considering its obstructed-ness.

Besides the weighed local representations, the feature maps of the whole face are encoded as a weighed vector by a Global-Gated Unit (GG-Unit). The weighed global facial features with local representations are concatenated and serve as a representation of the occluded face. Two fully connected layers are followed to classify the face to one of the emotional categories. ACNNs are optimized by minimizing the softmax loss.

Considering different interest of the local and global regions, we introduce two versions of ACNN: Patch based ACNN (pACNN) and Global-local based ACNN (gACNN). pACNN only contains local attention mechanism. As displayed in Fig. 2, two examples of PG-Units are illustrated

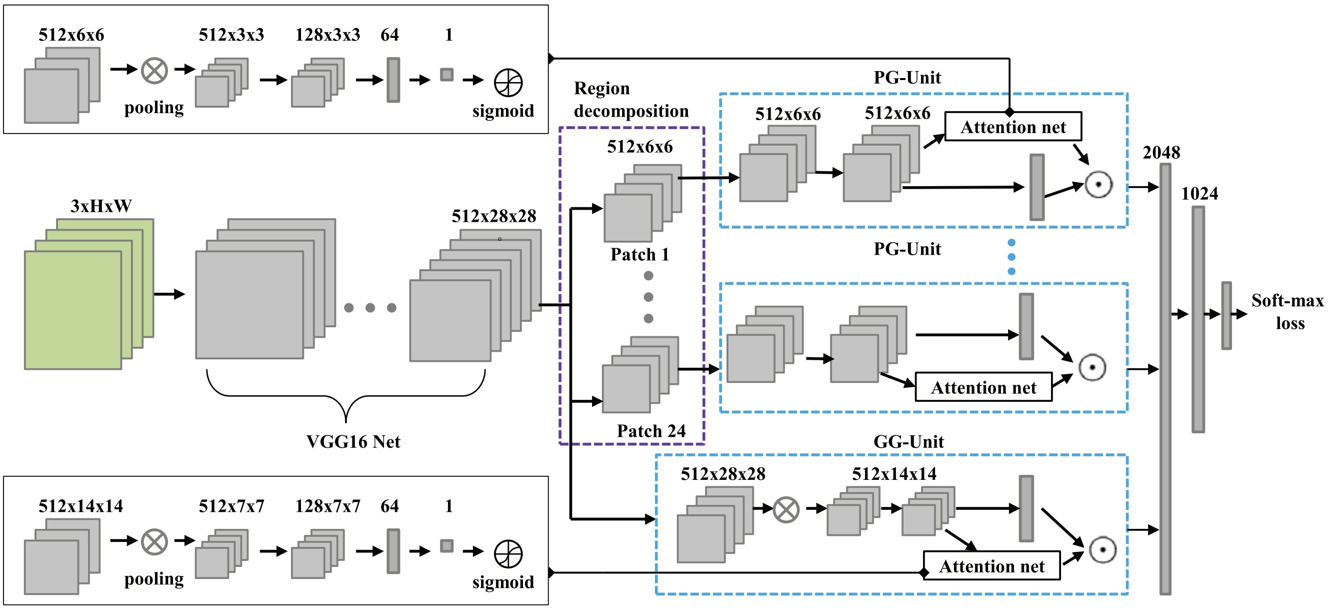


Fig. 2. Framework of the proposed ACNN. ACNN takes a facial image as the input and encodes the image with VGG-16 Net [32]. Then for pACNN the feature maps from the last convolution layer (conv4_2 in VGG16 net) are cropped into 24 local patches through a region decomposition scheme. Each patch is then processed by PG-Unit. Each PG-Unit encodes a patch by a vector-shaped feature and estimates how informative the patch is through an Attention net. For gACNN, the full feature maps are encoded and weighed as a representation vector through a GG-Unit. The softmax loss is attached at the end. Parameters in the overall network are learned by minimizing the softmax loss.

in the top two blue dotted rectangles. gACNN combines part-based with holistic-based attention methods. The GG-Unit in gACNN is illustrated in the bottom blue dotted rectangle.

B. Patch Based ACNN (pACNN)

Classifying facial expressions into different categories requires capturing regional distortions of facial muscles. Inspired by this intuition, pACNN is designed to focus on local discriminative and representative patches. pACNN contains two key schemes: region decomposition and occlusion perception. We present details of them as below.

1) Region Decomposition: Facial expression is distinguished in specific facial regions, because the expressions are facial activities invoked by sets of muscle motions. Localizing and encoding the expression-related parts is of benefit to recognize facial expression [12]. Additionally, dividing the face into multiple local patches helps to find the position of occlusions [14], [18].

To find the typical facial parts that are related to expression, We first detect 68 facial landmark points by the method in [33] and then, based on the detected 68 points, we select or re-compute 24 points that cover the informative regions of the face, including the eyes, nose, mouth, cheeks. Then we extract the patches according to the positions of each subject's facial landmarks. Fig. 3 shows the selection of facial patches. The details are as follows.

a) We pick 16 points from the original 68 facial landmarks to cover each subject's eyebrows, eyes, nose, mouth. The selected points are indexed as 19, 22, 23, 26, 39, 37, 44, 46, 28, 30, 49, 51, 53, 55, 59, 57 in Fig. 3 (b).

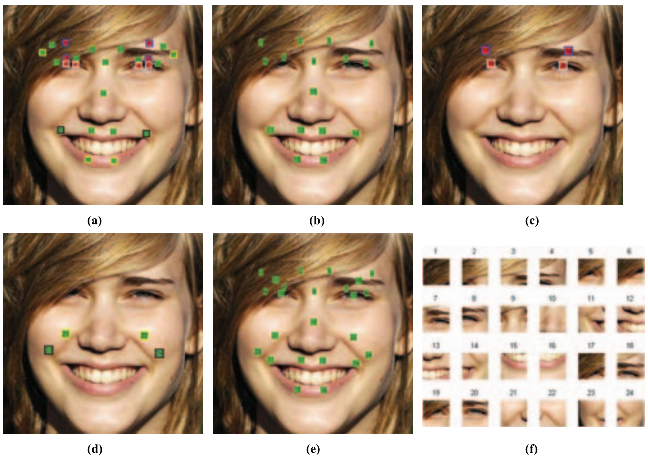


Fig. 3. Region decomposition of the face. (a) denotes totally 30 landmarks within original 68 facial landmarks. These points are involved in point selection. (b) shows 16 points that we pick to cover the facial regions on or around eyes, eyebrows, nose, mouth. (c) illustrates four points that are re-computed to better cover eyes and eyebrows. (d) displays four points that are re-computed to cover facial cheeks. (e) shows the selected 24 facial landmarks, around which the patches in (f) are cropped. Better viewed in color and zoom in.

b) We add one informative point for each eye and eyebrow. We pick four point pairs (red points in Fig.3a) around the eyes and eyebrows, then compute midpoint of each point pair as delegation. It is because we conduct patch extraction on convolutional feature maps rather than on the input image, adjacent facial points on facial images will coalesce into a same point on feature maps. The final midpoints are illustrated in Fig. 3 (c).

c) As facial cheeks are not directly covered by facial landmarks, we pick two point pairs, then compute midpoints of them. Indexes of the two point pairs are (18, 59), (27, 57). We then select two facial points that has fixed offset relative to the mouth corners. For the left mouth corner, the coordinate of target point can be calculated as $(x, y) = (x_{left} - 16, y_{left} - 16)$. For the right mouth corner, the target point coordinate is $(x, y) = (x_{right} + 16, y_{right} - 16)$.

The selected patches are defined as the regions taking each of the 24 points as the center. It is noteworthy that face alignment method in [33] is robust to occlusions, which is important for precise region decomposition.

As can be seen in the overall framework (Fig. 2), the patch decomposition operation is conducted on the feature maps from convolution layers rather than on the original image. This is because sharing some convolutional operations can decrease the model size and enlarge the receptive fields of subsequent neurons. Based on the $512 \times 28 \times 28$ feature maps as well as the 24 local region centers, we get a total of 24 local regions, each with a size of $512 \times 6 \times 6$.

2) Occlusion Perception With Gate Unit: We embed the Patch-Gated Unit in the pACNN to automatically perceives the blocked facial patches and pay attention mainly to the unblocked and informative patches. The detailed structure of PG-Unit is illustrated in the top two blue dashed rectangle in Fig. 2. In each patch-specific PG-Unit, the cropped local feature maps are fed to two convolution layers without decreasing the spatial resolution, so as to preserve more information when learning region specific patterns. Then, the last feature maps are processed in two branches. The first branch encodes the input feature maps as the vector-shaped local feature. The second branch consists of an Attention Net that estimates a scalar weight to denote the importance of the local patch. The local feature are then weighed by the computed weight.

Mathematically speaking, let us suppose $\mathbf{p}_i$ denotes the input $512 \times 6 \times 6$ feature maps of the i-th patch. $\tilde{\mathbf{p}}_i = \tilde{\phi}(\mathbf{p}_i)$ denote the last $512 \times 6 \times 6$ feature maps ahead of the two branches (top blue dashed rectangle in Fig. 2). The i-th PG-Unit takes the feature maps $\tilde{p}_i$ as the input, learns the local specific facial feature ψ_i:

$$\psi_i = \psi(\tilde{\mathbf{p}}_i) \tag{1}$$

And a corresponding weight α_i:

$$\alpha_i = \mathcal{I}_i(\tilde{\mathbf{p}}_i) \tag{2}$$

ψ_i is a vector that represents the un-weighed feature. α_i is a scalar that represent the patch i's importance or "unobstructed-ness". $\mathcal{I}(\cdot)$ means the operations in the Attention Net, consisting a pooling operation, one convolution operation, two inner productions, and a sigmoid activation. The sigmoid activation forces the output α_i ranges in [0, 1], where 1 indicates the most salient unobstructed patch and 0 indicates the completely blocked patch.

Finally, the i-th PG-Unit then uses α_i to weight the local feature ψ_i, and outputs its weighed feature ϕ_i:

$$\phi_i = \alpha_i \cdot \psi_i, \tag{3}$$

Under the attention mechanism in the proposed Gate-Unit, each cropped patch is weighed differently according to its occlusion conditions or importance. Through the end-to-end training of the overall pACNN, these PG-Units can automatically learn low weights for the occluded parts and high weights for the unblocked and discriminative parts.

C. Global-Local Based ACNN (gACNN)

pACNN is efficient to learn local facial representations with attention mechanism because it incorporates prior knowledge of facial expression. However, those facial patches in pACNN may ignore some complementary information displayed in facial images. Integration with global representation is expected to lead better FER performance in the presence of occlusions.

1) Integration With Full Face Region: Besides focusing on the local facial patches, gACNN takes global face region into consideration. On the one hand, the Global-Local Attention method help to infer local details and global context cues from image cocurrently [34]. On the other hand, gACNN can be viewed as a type of ensemble learning, which seeks to promote diversity among the learned features. The feature maps of the whole face are encoded from conv4_2 to conv5_2 in VGG16 net. Based on the $512 \times 28 \times 28$ feature maps, we obtain the encoded region with the size of 512x14x14.

2) Global-Gated Unit (GG-Unit): We further embed the GG-Unit in gACNN to automatically weigh the global facial representation. The detailed structure of GG-Unit is displayed in the downmost blue dashed rectangle in Fig. 2. Among the two branches in GG-Unit, The first branch encodes the input feature maps as the vector-shaped global representation. The second branch consists of an Attention Net that learns a scalar weight to denote the contribution of the global facial representation. The global representation is then weighed by the computed weight.

D. The Impact of Landmark Misalignment on ACNNs

The proposed ACNNs rely on the detected landmarks. It cannot be neglected that facial landmarks will suffer misalignment in the presence of severe occlusions. The proposed ACNNs are not sensitive to the landmark misalignment. We describe the reasons as below.

Firstly, we detected the facial landmarks by the method proposed in [33], which is robust against facial occlusions to some degree. As illustrated in Fig. 4, facial landmarks are quite accurate in the presence of partial occlusions (e.g., the first, second, third, fourth, fifth column). Large misalignments emerge when side face confronted with severe occlusions (sixth column in Fig. 4).

Secondly, The extracted patches are not sensitive to landmark misalignment. It is because our method extracts patches on convolutional feature maps other than on the input image. The spatial dimensions of convolutional feature maps are

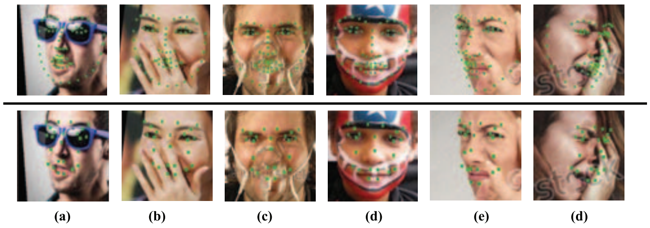


Fig. 4. Illustration of the original facial landmarks (first row) and the selected 24 points (second row). The facial images in first, second, third, fourth column are occluded on eyes, nose, mouth, cheeks respectively, their facial landmarks are quite accurate. The landmarks still suffer little misalignments for side face displayed in fifth column. In sixth column, The facial landmarks show large deviations under severe occlusion.

1/8 of input facial images. 8 pixels misalignments of facial landmarks will only induce 1 step deviation on related extracted patches.

Finally, The misaligned facial landmarks have no effect on full face representation in gACNN.

IV. Experiment

In this section, the experimental evaluations of ACNNs are presented. Before showing the results, we will firstly describe the experimental settings, including datasets, synthesis of occluded images, the newly collected facial expression dataset and our implementation details. Then, we compare our method with the state-of-the-art FER methods and methods with attention mechanism. Finally, an ablation analysis of the ACNNs is provided.

A. Experimental Setup

1) Datasets: We evaluated the methods on both in-the-wild datasets (RAF-DB [4], AffectNet [5], SFEW [35]) and in-the-lab datasets (CK+ [1], MMI [2], and Oulu-CASIA [3]). **RAF-DB** contains 30,000 facial images annotated with basic or compound expressions by 40 trained human coders. In our experiment, only images with basic emotions were used, including 12,271 images as training data and 3,068 images as test data. **AffectNet** is the largest dataset with annotated facial emotions. It contains about 400,000 images manually annotated for the presence of seven discrete facial expressions and the intensity of valence and arousal. We only used the images with neutral and 6 basic emotions, containing 280,000 training samples and 3,500 test samples. **The Extended Cohn-Kanade dataset (CK+)** contains 593 video sequences recorded from 123 subjects. We selected the first and last frame of each video sequence as the neural and target expressions, which resulted in 634 images. **MMI** dataset includes more than 30 subjects of both genders. There are 79 sequences of each subject. Each begins and ends with neural facial expression. We extracted the neutral and peak frames from each sequence, which resulted in 7348 images. **Oulu-CASIA** dataset contains six prototypic expressions from 80 people between 23 to 58 years old. We selected peak and neutral frames from sequences captured in normal illumination, which resulted in 9431 images. **SFEW** The Static Facial Expressions in the Wild (SFEW) dataset [35] is created by selecting static frames from Acted Facial Expressions in the Wild (AFEW) [36]. This dataset contains 95 subjects in total and covers unconstrained facial expressions, different head poses, ages, and occlusions. In total there are 663 well-labelled usable images.

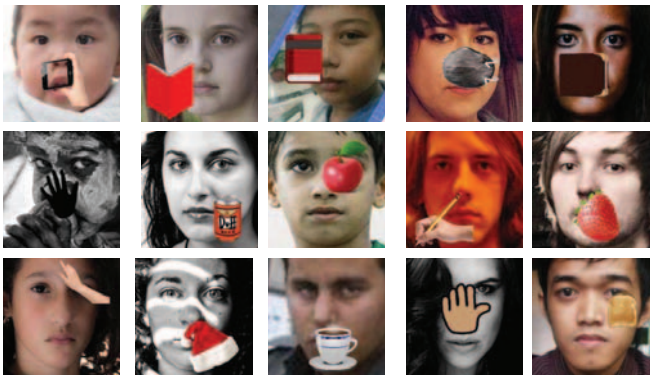


Fig. 5. Examples of the synthesized occluded facial images from RAF-DB dataset. The occluders are various in color, shape, and positions.

2) Synthesis of Occluded Images: We synthesized the occluded images by manually collecting about 4k images as masks for generating occluders. These mask images were collected from search engine using more than 50 keywords, such as beer, bread, wall, hand, hair, hat, book, cabinet, computer, cup etc. All the items were selected due to their high frequency of occurences as obstructions in facial images. Since Benitez-Quiroz *et al.* [37] verified that small local occluders take no affects on current FER algorithms, we heuristically restrain the size of the occluders S satisfying $S \in [96, 128]$, which is smaller or equal to half size of expression images. Fig. 5 shows some occluded examples derived from RAF-DB dataset. These artificial synthesised images are various in occlusion patterns and can better reflect occluder distribution in wild condition.

3) Facial Expression Dataset With Real Occlusion: Some initial efforts reported on FER in real-life with natural occlusions are limited to occlusions arising from sunglasses, medical mask [38], hands [14], [39], [40] or hair [21]. These results are primarily used for validating the performance of the system on sampled occluded images. Until now no thorough FER evaluation has been reported on a facial expression dataset in the presence of real occlusions [41].

To address the problems mentioned, we collected and annotated a facial expression dataset with real occlusions (FED-RO) in the wild for evaluation. To the best of our knowledge, it is the first facial expression dataset in the presence of real occlusions in the wild. We collected this dataset by mining Bing & Google search engine (with appropriate licenses) for occluded images. We adopted search queries such as "smile+face+glasses", "smile+face+beard", "disguise+face+eating", "sad+man+respirator", "neutral+child+drinking", "surprise+girl+cellphone" etc. Then each image was carefully labelled by three people. We removed the images labelled with inconsistent facial expression categories. To ensure the images in FEO-RO are not included in RAF-DB

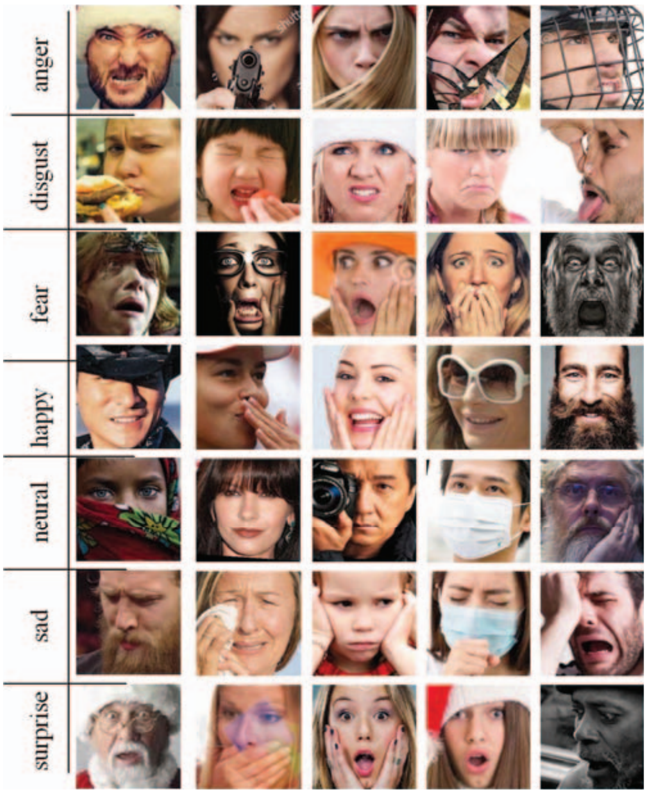


Fig. 6. Some example images picked from FED-RO. The occluders are various in occlusion types, positions, ratio etc.

or AffectNet dataset, we have taken following steps to filter out repeated facial images:

1) We conducted face detection, alignment, cropping for each image in FED-RO, RAF-DB, AffectNet dataset.
2) We extracted features from a VGG16 network (pre-trained on RAF-DB and AffectNet datasets) for each facial image to get a 4096 dimensional feature.
3) We calculated cosine similarity score for each image pair (i.e., one image come from FED-RO, the other image come from RAF-DB or AffectNet dataset). If the similarity score was larger than a pre-defined threshold, the image pair would be checked by human. If the two images were the same, we would drop the corresponding image in FED-RO. We set the threshold as 0.01.

Finally, FED-RO contains 400 images in total. The images are categorized into seven basic expressions (i.e., neutral, anger, disgust, fear, happy, sad, surprise). Fig.6 shows some example images in FED-RO. Occlusion patterns in FED-RO are diverse in color, shape, position and occlusion ratio. Table I illustrates total number of images for each expression category in FED-RO. Obviously there exists little imbalance among different categories.

4) Implementation Details: We implemented ACNNs using Caffe deep learning framework [42]. We adopted VGG-16 [32] as the backbone network for ACNNs due to its simple structure and excellent performance in object classification. We only chose the first nine convolution layers as the feature maps extractor. For pACNN, the extracted feature maps were decomposed into local regions and attached 24 PG-Units. For gACNN the full feature maps were encoded as a whole and attached a GG-Unit. The pre-trained model based on ImageNet dataset was used for initializing the model. All the datasets were mixed with their modifications with synthesized facial occlusions with 1:1 ratio except for SFEW dataset, for the reason that facial images in SFEW dataset already contain some occlusions. We adopted a batch-based stochastic gradient descent method to optimize the model. The base learning rate was set as 0.001 and was reduced by polynomial policy with gamma of 0.1. The momentum was set as 0.9 and the weight decay was set as 0.0005. The training of models was completed on a Titan-X GPU with 12GB memory. During the training stage, we set the actual batch size as 128 and the maximum iterations as 50K. It took about 2 days to finish optimizing the model.

TABLE I

TOTAL NUMBER OF IMAGES FOR EACH CATEGORY IN FED-RO

Dataset	neutral	anger	disgust	fear	happy	sad	surprise
FED-RO	50	53	51	58	59	66	63

TABLE II

TEST ACCURACY (%) ON RAF-DB AND AFFECTNET DATASETS. (*Clean*: ORIGINAL IMAGES. *occ.*: SYNTHETICALLY OCCLUDED IMAGES)

Methods	RAF-DB(clean/occ.)	AffectNet(clean/occ.)
VGG16 [32]	80.96/75.26	51.11/46.48
DLP-CNN [28]	80.89/76.29	54.47/51.07
GAN-Inpainting [44]	81.87/77.86	52.97/49.71
pCNN	81.64/76.09	53.9/50.32
pACNN (proposed)	83.27/78.05	55.33/52.47
gCNN	83.05/79.01	53.78/50.44
gACNN (proposed)	**85.07/80.54**	**58.78/54.84**

pCNN denotes CNN with region decomposition
gCNN denotes CNN with region decomposition and global representation

5) Evaluation Metric: We report FER performance on both non-occluded and occluded images of all datasets. For both occluded and non-occluded FER scenarios we adopt the overall accuracy on seven facial expression categories (i.e., six prototypical plus neural category) as a performance metric. In addition, we also report confusion matrix on FED-RO to show the discrepancies between the expressions. Both cross-dataset evaluation and 10-fold evaluation within dataset are used in our experiments.

B. Experiment on Artificial Occlusions

1) Comparison With Other Attention Models: We compare ACNNs with DLP-CNN ([28]). DLP-CNN estimates K spatial maps for attention parts generation. The hyper-parameter K is fine-tuned to the best in our experiments. Table II reports the results of pACNN, gACNN and DLP-CNN on RAF-DB and AffectNet datasets. Compared with DLP-CNN, pACNN achieves better FER performance. It is because the PG-Unit in pACNN enable the model to focus on local discriminative patches. From RAF-DB to AffectNet dataset, the performance

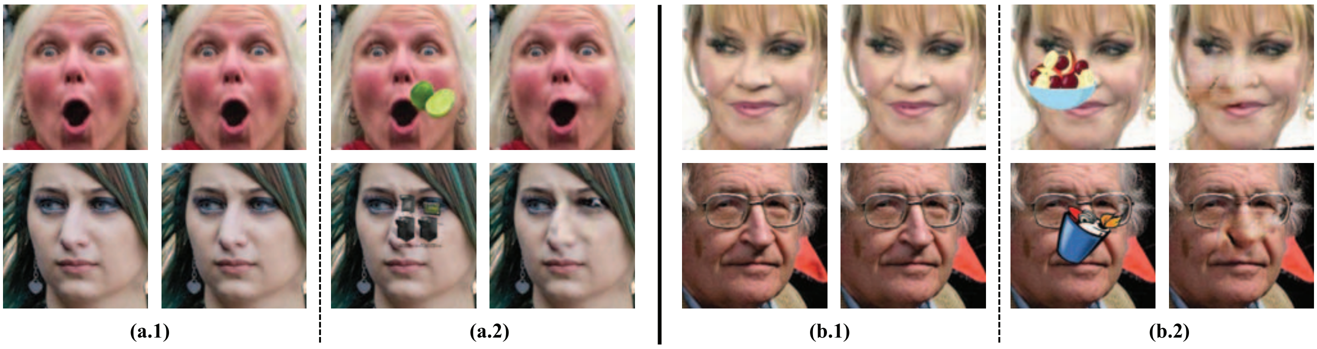


Fig. 7. Inpainting results on test sets of RAD-DB and AffectNet datasets. Sub-figure a.1 and a.2 come from RAF-DB, subfigure b.1 and b.2 are picked from AffectNet dataset. For each sub-figure, the left and right image denote the input and inpainting result respectively. Better viewed in color and zoom in.

gap between pACNN and DLP-CNN becomes narrow because of significant increase in training data.

Compared with DLP-CNN, gACNN outperforms by a big margin (5.17% for RAF-DB and 7.91% for AffectNet) on non-occluded images. It is because the model with local-global attention mechanism can better capture subtle muscle motions than the model with only global attention. gACNN exceeds DLP-CNN on occluded datasets with the help of Gate Unit, which encodes occlusion patterns in model weights and enable the model to pay attention to unblocked as well as distinctive facial regions.

2) Comparison With Other Methods Handling FER With Synthetic Occlusion: We compare ACNNs with state-of-the-art methods WLS-RF [14] and RGBT [16]. WLS-RF adopted multiply weighed random forests and RGBT converted a set of Gabor based part-face templates into template match distance features for FER with occlusions. We followed the same occlusion protocol of WLS-RF and RGBT and evaluated performance on model trained by AffectNet dataset.

Table III shows the comparisons. The overall performance of ACNNs is significantly better than that of WLS-RF and RGBT. ACNNs show little accuracy degradation under all the occlusion conditions, while WLS-RF and RGBT suffer notable performance degradation when mouth occluded or occlusion with R24 pattern. In detail, pACNN shows maximal (4.30%) FER performance degradation under R24 occlusion pattern. gACNN shows 2.61% performance degradation under mouth occlusion. It is noteworthy that ACNNs suffer little from the eyes occlusion. The experimental evaluation shows mouth occlusion impacts FER performance more than the eyes one, which is consistent with the conclusions of [10]. The proceeds of ACNNs are due to Gate Unit as well as large amount of training data in AffectNet dataset.

3) Comparison With Inpainting Methods: We also perform a comprehensive comparison with inpainting methods for FER in the presence of occlusions. Recently several generative avdersarial network (GAN) based methods have achieved high visual fidelity on image inpaining and completion, while many of them requires prior knowledge about occlusions (e.g., accurate positions of occluders) [44]–[47], or occlusions should be structured [48]. Those methods are not applicable to FER in the presence of arbitrary occlusions. Instead we adopted the method proposed in [43] for comparison. Isola *et al.* [43] adopted U-Net architecture as the generator, and utilized the loss function: $\mathcal{L} = \min_G \max_D \mathcal{L}_{GAN}(G, D) + \lambda \mathcal{L}_{L1}(G)$. The generator is tasked to not only fool the discriminator but also be near the ground truth output (the facial expression without artificial occlusions). λ denotes weight of L_1 loss and was tuned to achieve best visual plausible outputs. Since expression is a subtle property of faces that requires good representations of detailed local features, it remains unknown whether adversarial generative method can recover corrupt local regions in facial expression images.

TABLE III

10-Fold Test Accuracy (%) on CK+ Dataset With Synthetic Occlusions. (R8, R16, R24 Denote the Size of the Occlusion as 8×8, 16×16, 24×24. The Full-Image Size Is 48×48)

Occlusion	pACNN	gACNN	WLS-RF [40]	RGBT [16]
non-occlusion	**97.03**	96.4	94.3	94.4
R8	**96.58**	**96.58**	92.2	92.0
R16	95.70	**95.97**	86.4	82.0
R24	92.86	**94.82**	74.8	62.5
eyes occluded	96.50	**96.57**	87.9	88.0
mouth occluded	**93.92**	93.88	72.7	30.3

Fig. 7 illustrates some inpainting results of test images in RAF-DB and AffectNet datasets. It is obvious that GAN based inpainting method can successfully recover an occluded facial image to a great degree. At the same time we can observe little distortions in the generated facial images if the occluded facial regions contain subtle textures (e.g., right eye in the lower part of sub-graph a.2).

Table II illustrates the performance of the generative inpainting method. The inpainting method achieves slightly better performance than VGG16 on both RAF-DB and AffectNet datasets, still far behind pACNN as well as gACNN. It is because although generated facial images contain less visually occlusions, the inevitable distortions as well as distance from the original latent image manifold limit more FER performance enhancement [47].

4) Cross Dataset Evaluation: We evaluated the generalization ability of ACNNs under the cross-dataset evaluation protocol. In our experiments, ACNNs were trained on

TABLE IV

CROSS DATASET EVALUATION (ACCURACY%) ON IN-THE-LAB DATASETS (*Clean*: ORIGINAL IMAGES. *occ.*: SYNTHESIZED OCCLUDED IMAGES)

method	CK+ (clean/occ.)	MMI (clean/occ.)	Oulu-CASIA (clean/occ.)	SFEW
Molla et al. [50]	64.2 / –	55.6 / –	– / –	39.8
Mayer et al. [51]	60.8 / –	60.3 / –	– / –	–
Zhang et al. [52]	61.2 / –	**66.9** / –	– / –	–
	model trained on AffectNet dataset			
pCNN	89.27 / 85.33	66.94 / 61.26	54.77 / 51.05	48.02
pACNN	90.38 / 86.27	68.92 / 63.94	57.93 / 54.18	49.75
gCNN	88.01 / 84.70	68.00 / 63.18	56.14 / 52.61	**52.59**
gACNN	**91.64 / 88.17**	**70.37 / 65.48**	**58.18 / 55.42**	51.72
	model trained on RAF-DB			
pCNN	79.81 / 76.02	57.02 / 53.70	49.83 / 46.98	45.07
pACNN	80.28 / 79.49	55.61 / 53.44	50.04 / 47.15	44.33
gCNN	**81.86 / 79.81**	58.31 / 55.17	49.05 / 45.89	47.04
gACNN	81.07 / 79.49	**59.51 / 55.28**	**50.31 / 47.76**	**51.47**

RAF-DB or AffectNet dataset and evaluated on CK+, MMI, Oulu-CASIA, SFEW dataset with or without synthetic occlusions. Table IV shows the results compared with other FER methods.

Among the compared experiments, Mollahosseini *et al.* [49] adopted an inception based CNN and provided the average cross-dataset recognition accuracy. Mayer *et al.* [50] and Zhang *et al.* [51] reported the highest cross-dataset results, which were both trained on MMI and evaluated on CK+ or vice versa.

As can be seen from Table IV, ACNNs achieve better performance than other methods with few exceptions. For model trained on AffectNet, gACNN exceeds [49]–[51] by at least 42.7% and 5.18% on CK+ and MMI dataset respectively. It suggests that ACNNs trained on AffectNet dataset can generalize better than ACNNs trained on RAF-DB due to a larger amount of training data.

The benefit of Gate Unit on SFEW dataset is not as consistent as on other datasets. It is because facial images in SFEW are diverse in large head pose, which causes some misalignments in facial landmark points.

C. Experiment on Realistic Occlusions

We merged all training images in RAF-DB and AffectNet datasets for training, and evaluated performances of VGG16 [32], ResNet-18 [52], and their related CNNs and ACNNs on our manually collected FED-RO. We also adopted DLP-CNN [28], GAN-Inpainting [43] for evaluations.

Table V illustrates FER performance on FED-RO. ACNNs achieve the best average FER performance among all network structures. In detail, pACNN and gACNN outperform VGG16 by 6.77% and 8.51% respectively. ResNet-18 achieves comparable performance with VGG16. The slight performance improvements are reasonable.

Results in Table V illustrate that DLP-CNN achieves comparable performance with VGG16, and that GAN-Inpainting behaves the worst among all methods. The poor performance of GAN-Inpainting can be explained in two aspects. Firstly, it is difficult for the model to recover missing facial parts in the presence of arbitrary real occlusions. Secondly, the model has a weak ability to focus on unblocked & distinctive facial regions.

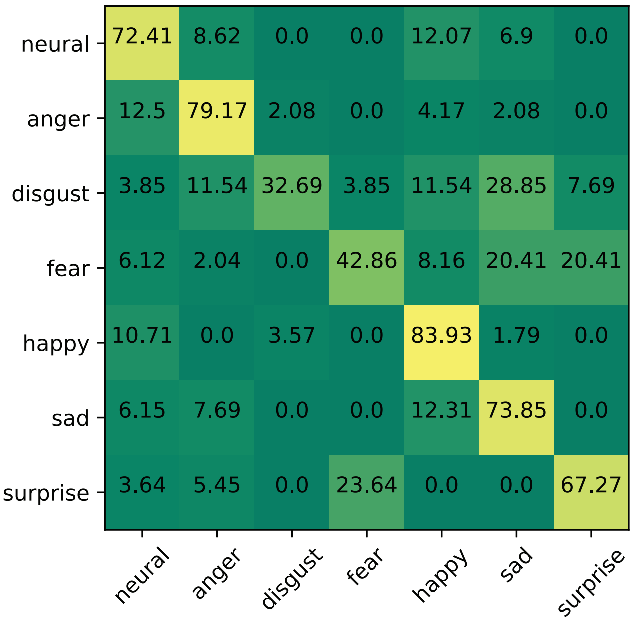


Fig. 8. Confusion matrix based on gACNN for FED-RO. We merged RAF-DB and AffectedNet dataset for training.

To investigate per expression category classification performance on FED-RO, we also reported confusion matrix of gACNN in Fig. 8. It is clear that gACNN achieves highest and lowest classification accuracy on *happy* and *disgust* category respectively. Easily confused expression categories are *fear* and *surprise*, *disgust* and *sad*, *fear* and *sad*.

We carefully browsed all predictions on FED-RO and displayed some representative failed examples in Fig. 9. Although gACNN is robust to most occlusions, it still suffers from extremely heavy facial occlusions and novel occluders. The former can be explained that severe occlusions will inevitably cause large misalignment in facial landmarks. Thus gACNN is unable to focus on preplanned local facial patches. The latter can be explained that Gate Unit in gACNN may generate a weight that is not so adaptive when meets a rarely seen occluder (e.g., white beard in Fig. 9). One possible way to handle such hard examples is to take the context [53] or human body gestures [54] within images into consideration, which we will explore in future work.

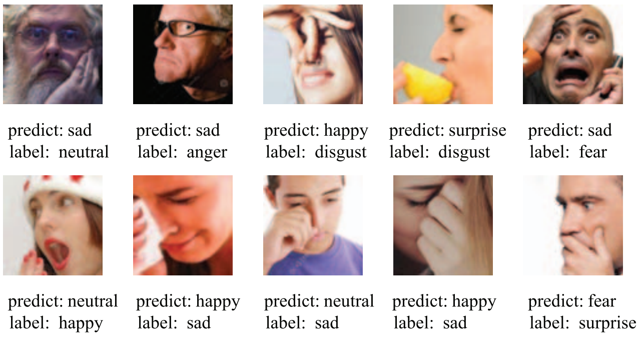
D. Ablation Analysis

Both diverse representations learning and Gate Unit help ACNNs gain improvements on FER. We conducted a quantitative evaluation of these two components in order to better understand our method. For a more detailed analysis of the FER results, we also explored how different network structures affect the attention maps of the clean and occluded images.

TABLE V
ACCURACY (%) COMPARISON ON FED-RO

method	base	pCNN	gCNN	pACNN	gACNN	DLP-CNN	GAN-Inpainting
VGG16	60.15	59.00	63.18	64.22	**65.27**	60.31	58.33
ResNet-18	64.25	61.75	63.50	64.25	**66.50**		

base denotes original network structure


predict: sad label: neutral

predict: sad label: anger

predict: happy label: disgust

predict: surprise label: disgust

predict: sad label: fear

predict: neutral label: happy

predict: happy label: sad

predict: neutral label: sad

predict: happy label: sad

predict: fear label: surprise

Fig. 9. Some example images in FED-RO that gACNN failed to predict the correct expression categories. It suggests that gACNN is vulnerable to large head pose, extremely heavy facial occlusion, as well as novel occlusions.

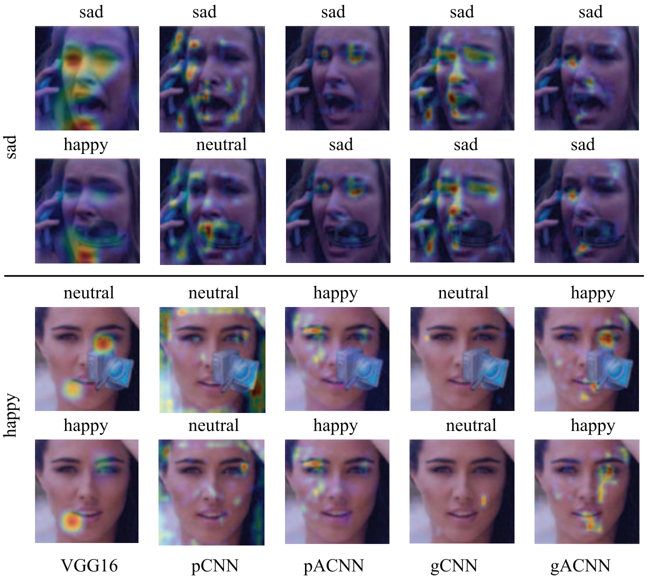


Fig. 10. Attention maps of several test images (i.e., the first and fourth row) and their modifications with artificial facial occlusions (i.e., the second and third row). The images' expression labels are displayed on the leftmost. Different methods' predictions are displayed directly above the corresponding images. A deep red denotes high attention. Better viewed in color and zoom in.

1) VGG-16 VS pCNN, pCNN VS gCNN: We compared VGG-16 and pCNN (CNN without Gate Unit) to verify benefit of region decomposition. As listed in Table II, pCNN exceeds VGG-16 on both original and occluded images. The promotions of pCNN suggest that globally encoded representation has fallen behind in reflecting subtle muscle motions compared with locally learned patterns.

With the help of global and local representations, gCNN achieves better and comparable performance than pCNN on RAF-DB and AffectNet datasets respectively. It is because features derived from gCNN contain both global facial information and detailed properties of local facial muscles. Thus the influence of occlusions can be suppressed to a larger degree.

2) pCNN VS pACNN, gCNN VS gACNN: We conducted two groups of experiments (i.e., pCNN vs pACNN, gCNN vs gACNN) to verify benefit of Gate Unit. As displayed in Table II, total improvements of pACNN on RAF-DB and AffectNet datasets are 1.99%, 2.58% and 3.65%, 4.27% respectively. As a comparison, gACNN exceeds gCNN by 2.43%, 1.93%, 9.29%, 8.72%. This is because Gate-Unit enables the model to attend to most related local patches, and shift attention to other related local parts when original ones are occluded. Compare with pACNN, gACNN achieves a higher accuracy with the help of globally weighed feature, which can supply essential context information ignored by the local patches in pACNN. Similar performance improvements can be found in Table IV, where gACNN outperforms pACNN on CK+, MMI, Oulu-CASIA, SFEW datasets. It is noteworthy that ACNNs also outperform other network structure on FED-RO (Table V).

To invertigate how Gate Unit influence attention maps for different network structure, we visualized the attention maps of VGG16, pCNN, pACNN, gCNN and gACNN using the method in [55]. Selvaraju *et al.* [55] adopted the gradient-weighed class activation mapping (Grad-CAM) to visualize a wide variety of CNN model families. To visualize the attention map of PG-CNN, we firstly derived the Grad-CAMs of conv4_2 (in VGG16 net) and its cropped local patches, then scaled these local Grad-CAMs according to the attention weights generated by related Gate Units. At last we projected these local weighed Grad-CAMs back to conv4_2 according to their original coordinates. Attention maps of pCNN, gCNN, gACNN are derived in the same manner.

Fig. 10 shows the attention maps. VGG16 relies on relatively large facial regions and tends to be vulnerable to occlusions. Compared with pCNN, pACNN is capable of discovering local discriminative patches and is much less sensitive to occlusions. It is obvious that gCNN fails to focus on discriminative facial regions when expression intensity decreases (e.g., occluded *happy* face in Fig. 10). Among all the different network structures, pACNN and gACNN are capable of perceiving occlusions and shifting attention from the occluded patches (e.g., right mouth corner in the subfigures for *happy*) to other unobstructed ones (e.g., right eye in the subfigures for *happy*).

Fig. 11 displays images with real occluders picked from test set in RAF-DB and AffectNet datasets. Obviously gACNN outperforms other network structure. Take facial images in the first row for example, gACNN is capable of focusing on facial regions and neglecting occlusion parts under low image resolution. In the second row of Fig. 11, gACNN precisely focuses on discriminative local regions in the presence of extreme self-occlusion. Besides, gACNN outputs correct predictions for both of the facial images.

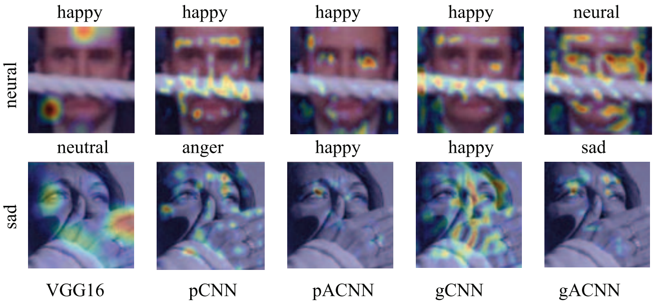


Fig. 11. Attention maps of several test images with real occlusions in RAF-DB. The images' labels are displayed on the leftmost. Different methods' predictions are displayed directly above the corresponding images. Better viewed in color and zoom in.

V. Conclusion

In this work we present CNN with attention mechanism (ACNN) for facial expression recognition in the presence of occlusions. The Gate Unit in ACNN enables the model to shift attention from the occluded patches to other unobstructed as well as distinctive facial regions. Considering that facial expression is distinguished in specific facial regions, we designed a patch based pACNN that incorporates region decomposition to find typical facial parts that are related to expression. We also developed an efficient gACNN to supplement global facial information for FER in the presence of occlusions. Experiments under intra and cross dataset evaluation protocols demonstrated ACNNs outperform other state-of-the-art methods. Ablation analyses show ACNNs are capable of shifting attention from occluded patches to other related ones. For future work, we will study how to generate attention parts in faces without landmarks, as ACNNs rely on robust face detection and facial landmark localization modules.

Acknowledgment

The authors would like to thank Dr. Xin Liu and Prof. Hu Han for discussions.

References

[1] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. CVPRW*, Jun. 2010, pp. 94–101.

[2] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. ICME*, Jul. 2005, p. 5.

[3] G. Zhao, X. Huang, M. Taini, and S. Z. Li, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, 2011.

[4] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. CVPR*, Jul. 2017, pp. 2584–2593.

[5] A. Mollahosseini, B. Hasani, and M. H. Mahoor. (2017). "AffectNet: A database for facial expression, valence, and arousal computing in the wild." [Online]. Available: https://arxiv.org/abs/1708.03985

[6] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated CNN for occlusion-aware facial expression recognition," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2209–2214.

[7] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[8] E. Osherov and M. Lindenbaum, "Increasing CNN robustness to occlusions by reducing filter support," in *Proc. CVPR*, Oct. 2017, pp. 550–561.

[9] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *Proc. CVPR*, Jun. 2011, pp. 2857–2864.

[10] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image Vis. Comput.*, vol. 26, no. 7, pp. 1052–1067, 2008.

[11] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 6, pp. 748–763, Jun. 2002.

[12] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Proc. CVPR*, Jun. 2012, pp. 2562–2569.

[13] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Towards a dynamic expression recognition system under facial occlusion," *Pattern Recognit. Lett.*, vol. 33, no. 16, pp. 2181–2191, 2012.

[14] A. Dapogny, K. Bailly, and S. Dubuisson, "Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection," *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 255–271, 2017.

[15] W. Li, F. Abtahi, and Z. Zhu, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *Proc. CVPR*, Jul. 2017, pp. 6766–6775.

[16] L. Zhang, D. Tjondronegoro, and V. Chandran, "Random Gabor based templates for facial expression recognition in images with facial occlusion," *Neurocomputing*, vol. 145, pp. 451–464, Dec. 2014.

[17] R. Min, A. Hadid, and J.-L. Dugelay, "Improving the recognition of faces occluded by facial accessories," in *Proc. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 442–447.

[18] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Facial action unit prediction under partial occlusion based on error weighted cross-correlation model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3482–3486.

[19] H. Towner and M. Slater, "Reconstruction and recognition of occluded facial expressions using PCA," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* Springer, 2007, pp. 36–47.

[20] Y. Deng, D. Li, X. Xie, K.-M. Lam, and Q. Dai, "Partially occluded face completion and recognition," in *Proc. ICIP*, Nov. 2009, pp. 4145–4148.

[21] X. Mao, Y. Xue, Z. Li, K. Huang, and S. Lv, "Robust facial expression recognition based on RPCA and AdaBoost," in *Proc. 10th Workshop Image Anal. Multimedia Interact. Services*, May 2009, pp. 113–116.

[22] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[23] M. Afifi and A. Abdelhamed. (2017). "AFIF4: Deep gender classification based on AdaBoost-based fusion of isolated facial features and foggy faces." [Online]. Available: https://arxiv.org/abs/1706.04277

[24] D. M. Vo, A. Sugimoto, and T. H. Le, "Facial expression recognition by re-ranking with global and local generic features," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 4118–4123.

[25] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.

[26] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. ICCV*, Oct. 2017, pp. 5219–5227.

[27] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.

[28] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. ICCV*, Oct. 2017, pp. 3239–3248.

[29] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma, "Structured attentions for visual question answering," in *Proc. ICCV*, vol. 3, Oct. 2017, pp. 1300–1309.

[30] F. Juefei-Xu, E. Verma, P. Goel, A. Cherodian, and M. Savvides, "DeepGender: Occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 136–145.

[31] E. Norouzi, M. N. Ahmadabadi, and B. N. Araabi, "Attention control with reinforcement learning for face recognition under partial occlusion," *Mach. Vis. Appl.*, vol. 22, no. 2, pp. 337–348, 2011.

[32] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[33] J. Zhang, M. Kan, S. Shan, and X. Chen, "Occlusion-free face alignment: Deep regression networks coupled with de-corrupt autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3428–3437.

[34] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4133–4139.

[35] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2106–2112.

[36] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 3, pp. 34–41, Jul./Sep. 2012.

[37] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez. (2017). "EmotioNet challenge: Recognition of facial expressions of emotion in the wild." [Online]. Available: https://arxiv.org/abs/1703.01210

[38] S.-S. Liu, Y. Zhang, K.-P. Liu, and Y. Li, "Facial expression recognition under partial occlusion based on Gabor multi-orientation features fusion and local Gabor binary pattern histogram sequence," in *Proc. 9th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process. (IIH-MSP)*, Oct. 2013, pp. 218–222.

[39] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 699–714, May 2005.

[40] T. D. Nguyen and S. Ranganath, "Tracking facial features under occlusions and recognizing facial expressions in sign language," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–7.

[41] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran. (2018). "Facial expression analysis under partial occlusion: A survey." [Online]. Available: https://arxiv.org/abs/1802.08784

[42] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. (2017). "Image-to-image translation with conditional adversarial networks." [Online]. Available: https://arxiv.org/abs/1611.07004

[44] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. (2018). "Generative image inpainting with contextual attention." [Online]. Available: https://arxiv.org/abs/1801.07892

[45] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, p. 107, 2017.

[46] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5892–5900.

[47] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6882–6890.

[48] Z. Li, Y. Hu, and R. He. (2017). "Learning disentangling and fusing networks for face completion under structured occlusions," [Online]. Available: https://arxiv.org/abs/1712.04646

[49] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. WACV*, Mar. 2016, pp. 1–10.

[50] C. Mayer, M. Eggers, and B. Radig, "Cross-database evaluation for facial expression recognition," *Pattern Recognit. Image Analysis*, vol. 24, no. 1, pp. 124–132, 2014.

[51] X. Zhang, M. H. Mahoor, and S. M. Mavadati, "Facial expression recognition using l_p-norm MKL multiclass-SVM," *Mach. Vis. Appl.*, vol. 26, no. 4, pp. 467–483, 2015.

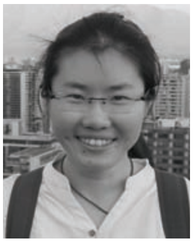
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[53] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1960–1968.

[54] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 64–84, Feb. 2009.

[55] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. (2016). "Grad-CAM: Visual explanations from deep networks via gradient-based localization." [Online]. Available: https://arxiv.org/abs/1610.02391

Yong Li received the B.S. and M.S. degrees in computer science from Zhengzhou University, Henan, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Institute of Computing Technology, Chinese Academy of Sciences, since 2016. He was a Software Engineer with Baidu, Inc., from 2015 to 2016. His research interests include machine learning and affective computing. He specially focuses on facial expression recognition and facial action unit detection.



Jiabei Zeng received the B.S. and Ph.D. degrees from Beihang University, Beijing, China, in 2011 and 2017, respectively. From 2013 to 2015, she was a Visiting Scholar with Carnegie Mellon University. She has been an Assistant Professor with the Institute of Computing Technology, Chinese Academy of Sciences, since 2017. Her research interests include computer vision and affective computing, especially on facial expression analysis.

Shiguang Shan received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He joined ICT, CAS, in 2002, where he has been a Professor, since 2010. He is currently the Deputy Director of the Key Laboratory of Intelligent Information Processing, CAS. He has published over 200 papers in refereed journals and proceedings in the areas of computer vision and pattern recognition. He is also personally interested in brain science and cognitive neuroscience and their interdisciplinary research topics with AI. His research interests include computer vision, pattern recognition, and machine learning. He specially focuses on face recognition related research topics. He was a recipient of the China's State S&T Progress Award in 2005 for his research work and the China's State Natural Science Award in 2015. He has served as the Area Chair for many international conferences, including ICCV 2011, ICPR 2012, ACCV 2012, FG 2013, ICPR 2014, ICASSP 2014, ACCV 2016, ACCV 2018, FG 2018, and BTAS 2018. He is currently an Associate Editor of several international journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, *Computer Vision and Image Understanding*, *Neurocomputing*, and *Pattern Recognition Letters*.

Xilin Chen received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1988, 1991, and 1994, respectively. He was a Professor with the Harbin Institute of Technology from 1999 to 2005. He has been a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), since 2004. He is currently the Director of the Key Laboratory of Intelligent Information Processing, CAS. He has published one book and over 200 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is a Fellow of the IEEE/IAPR/CCF. He is currently an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the *Journal of Visual Communication and Image Representation*, a leading Editor of the *Journal of Computer Science of Technology*, and the Associate Editor-in-Chief of the *Chinese Journal of Computers*. He was a recipient of several awards, including the China's State Scientific and Technological Progress Award in 2000, 2003, 2005, and 2012 for his research work, and the China's State Natural Science Award in 2015.