Facial Expression Recognition in the Wild: A Cycle-Consistent Adversarial Attention Transfer Approach

Feifei Zhang[1,2], Tianzhu Zhang[2,3], Qirong Mao[1], Lingyu Duan[4], Changsheng Xu[2,3]
[1]School of Computer Science and Communication Engineering, Jiangsu University, China
[2]National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
[3]University of Chinese Academy of Sciences, [4]Institute of Digital Media, Peking University, China
{susanzhang,mao_qr}@ujs.edu.cn,{tzzhang,csxu}@nlpr.ia.ac.cn,lingyu@pku.edu.cn

ABSTRACT

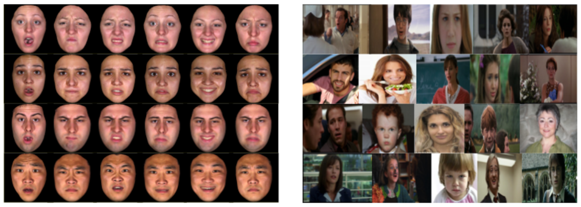
Facial expression recognition (FER) is a very challenging problem due to different expressions under arbitrary poses. Most conventional approaches mainly perform FER under laboratory controlled environment. Different from existing methods, in this paper, we formulate the FER in the wild as a domain adaptation problem, and propose a novel auxiliary domain guided Cycle-consistent adversarial Attention Transfer model (CycleAT) for simultaneous facial image synthesis and facial expression recognition in the wild. The proposed model utilizes large-scale unlabeled web facial images as an auxiliary domain to reduce the gap between source domain and target domain based on generative adversarial networks (GAN) embedded with an effective attention transfer module, which enjoys several merits. First, the GAN-based method can automatically generate labeled facial images in the wild through harnessing information from labeled facial images in source domain and unlabeled web facial images in auxiliary domain. Second, the class-discriminative spatial attention maps from the classifier in source domain are leveraged to boost the performance of the classifier in target domain. Third, it can effectively preserve the structural consistency of local pixels and global attributes in the synthesized facial images through pixel cycle-consistency and discriminative loss. Quantitative and qualitative evaluations on two challenging in-the-wild datasets demonstrate that the proposed model performs favorably against state-of-the-art methods.

KEYWORDS

facial expression recognition, domain adaptation, attention transfer, generative adversarial networks, emotional cue extraction

ACM Reference Format:

Feifei Zhang, Tianzhu Zhang, Qirong Mao, Lingyu Duan, Changsheng Xu. 2018. Facial Expression Recognition in the Wild: A Cycle-Consistent Adversarial Attention Transfer Approach. In 2018 ACM Multimedia Conference (MM '18), Oct. 22–26, 2018, Seoul, Republic of Korea. ACM, NY, NY, USA, 8 pages. https://doi.org/10.1145/3240508.3240574



(a) facial images under laboratory (b) facial images in the wild

Figure 1: Facial expression recognition under laboratory controlled environment (a) and in the wild (b). The FER in the wild is more challenging due to arbitrary pose variations, spontaneous expressions, and illumination changes.

1 INTRODUCTION

As an essential way of human emotional behavior understanding, facial expression recognition (FER) has drawn a great deal of attention from the multimedia research community in recent years. The FER has tremendous impact on a wide-range of applications including psychology, medicine, security, digital entertainment, and driver monitoring, which make it become a core component in the next generation of artificial intelligence [5, 6, 16, 41, 54].

The FER aims to analyze and classify a given facial image into several emotion types, such as, fear, angry, disgust, sad, happy and surprise [9]. To achieve this goal, numerous methods [24, 31, 34] have been proposed for the FER during the past several years. However, most of the existing methods [10, 45, 59] recognize facial expressions on the datasets recorded under laboratory controlled environment, which is far away from real-world scenarios. As shown in Figure 1(a), the facial images are usually captured under laboratory conditions with controlled poses, illuminations, and deliberately acted expressions. However, in the practical scenarios as shown in Figure 1(b), extensive complicated environments would significantly degrade the performance of the recognition methods that deserve working well for the FER on the well-designed datasets. As a result, the FER in the wild is largely unexplored. Different from existing methods, we focus on the FER in the wild, which is to perform the FER by identifying or authorizing individual expressions with facial images captured in real scenarios without any controlled conditions. Therefore, it is more challenging and more applicable.

However, it is not easy to perform the FER in the wild. As shown in Figure 1(b), the FER in practical scenarios would suffer from illumination changes, arbitrary pose variations, spontaneous facial expressions, unconstrained background, and many other unpredictable and challenging situations. Besides, most publicly available FER datasets in the wild have insufficient training data. As shown in Table 1, the Static Facial Expressions in the wild (SFEW) dataset [8] contains only 700 images while the EmotioNet [2] has only

MM '18, October 22–26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240574

Table 1: Details of existing benchmarks for the FER.

Dataset	Pose	Expression	Training Samples
SFEW	arbitrary	7	700
EmotioNet	arbitrary	6	1,141
BU-3DFE	143	7	185,900

1,141 images. Therefore, a great and common strategy to conduct the FER in the wild is gathering sufficient annotated facial images in the wild. However, labeling such facial images is labor-intensive and time-consuming. An avenue for overcoming the lack of labeled training data is to adopt transfer learning, which can apply knowledge learnt from one domain to other related domains. For the FER, it is easier to collect a large-scale dataset under laboratory controlled environment. As shown in Table 1, the BU-3DFE [49] has more than $185,000$ labeled images. This inspires us to conduct the FER in the wild by using transfer learning with the large-scale labeled data captured under laboratory controlled environment. Here, we take the dataset with sufficient training samples captured under laboratory as the source domain, and the dataset in the wild with limited samples as the target domain. Then, numerous transfer learning methods can be utilized to boost the performance for the FER in the wild. In the context of deep learning, fine-tuning a deep network pre-trained on the dataset with sufficient training samples [14] or conducting domain adversarial networks [48] are the frequently used strategies to learn task specific deep features. However, since the ratio between the number of learnable parameters and the number of training samples still remains the same, these methods also need sufficient training samples or to be terminated after a relatively small number of iterations [12]. To overcome this issue, attention transfer has been proposed and successfully adopted in several domain adaptation tasks [30, 51], which attempts to transfer attention knowledge from a powerful deeper network that is trained with sufficient training samples to a shallower network that can be trained with limited training data with the goal of improving the performance of the latter. However, it is still challenging to train such a high-quality cross-domain model for the FER in the wild due to the large domain shift in the images. As shown in Figure 1, the facial images captured under laboratory controlled environment only consist of faces without any background, but the facial images in the wild usually have complex background.

To deal with the large domain shift between source domain and target domain for the FER, we can adopt the data in source domain to synthesize facial images as similar as the data in target domain by using generative adversarial networks (GAN) model, which has been proven to generate impressively realistic faces through a two-player game between a generator G and a discriminator D. However, these GAN based methods usually require sufficient input-output image pairs for training, which is not available for FER in the wild with limited samples in the target domain. Fortunately, we can easily collect large-scale web images, which are similar to the facial images in target domain, as auxiliary data by querying widely available commercial search engines. Therefore, this inspires us to resort to the GAN for transforming labeled facial images in source domain so that they look like images captured in the wild (target domain) through harnessing information from the web facial images in the auxiliary domain. The basic idea is shown in Figure 2. For the GAN model, there are many promising image-to-image translation developments [23, 32, 60], but they do not necessarily preserve key attributes, such as expression, pose in the FER task. Although the generated image may “look” like it comes from the auxiliary domain, some crucial semantic information may be lost. For example, a model transforming a labeled facial image in source domain to generate an image in auxiliary domain may lose the expression attributes, which cannot be straightforwardly applied to the FER task. Consequently, the desired model for our task is that it can generate facial images which should simultaneously preserve the expression attributes in source domain and transform helpful content information in auxiliary domain as shown in Figure 2.

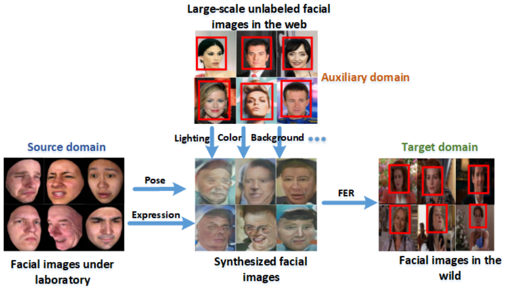

Large-scale unlabeled facial images in the web
Auxiliary domain
Lighting
Color
Background
Source domain
Target domain
Pose
Expression
FER
Facial images under laboratory
Synthesized facial images
Facial images in the wild

Figure 2: The basic idea of the proposed model, which takes the large-scale unlabeled web facial images as the bridge to alleviate the domain gap between source and target domain, and can recognize facial expression in the wild with the help of facial images under laboratory.

Inspired by the above discussions, we propose a novel Cycle-consistent adversarial Attention Transfer (CycleAT) model with the guiding of auxiliary domain, which is able to not only synthesize labeled facial images in the wild, but also conduct FER in the wild. Specifically, we first train an auxiliary domain guided cycle-consistent GAN to generate labeled facial images in the wild, which can be further divided into two stages. At stage-I, we use a pixel cycle-consistency constraint in the facial image generator to preserve the local structural information of the facial images in source domain, which can guarantee the quality of generated images. At stage-II, we add a classifier f_W into the generator by exploiting the global attribute-level consistency, which can preserve the expression attributes in the generated facial images, thus ensuring the label consistency. Then, a target classifier f_T is trained by taking advantage of the sufficient labeled generated facial images. Unlike previous approaches that distill knowledge through class probabilities [19], we do so by attention transfer strategy, which is helpful to train an effective classifier with limited training samples.

The major contributions of this work can be summarized as follows. (1) We propose an auxiliary data guided learning model for simultaneous facial image synthesis and FER in the wild by harnessing information from labeled facial images captured under laboratory controlled environment and unlabeled web facial images from the Internet. (2) The class-discriminative spatial attention maps from the classifier in source domain are leveraged to boost the performance of the classifier in target domain. (3) The local and global structural consistency of the synthesized facial images has been effectively enforced through pixel cycle-consistency and discriminative loss. (4) The proposed model achieves state-of-the-art results on the SFEW [8] and EmotioNet [2] datasets for facial expression recognition in the wild.

2 RELATED WORK

In this section, we mainly discuss methods related to FER, domain adaptation, and generative adversarial network.

Facial Expression Recognition. Extensive efforts have been devoted to facial expression analysis [5, 53, 55, 61]. Most of existing approaches on the FER study the expressions of six basic emotions including happy, sad, surprise, fear, angry and disgust because of their marked reference representation in our affective lives and the availability of the relevant training data [52, 64, 65]. Generally, most existing FER methods mainly include two stages, i.e., feature extraction and expression recognition. In the first stage, features are extracted from facial images to characterize facial appearance/geometry changes caused by activation of the expression. According to whether the features are extracted by manually designed descriptors or by deep learning models [11, 56–58], they can be categorized into engineered features [10, 39, 63] and learning-based features [16, 22, 24, 31, 54]. For the engineered features, it can be further divided into texture-based local features, geometry-based global features, and hybrid features. The texture-based features mainly include HOG [14], SIFT [63], Histograms of LBP [66], Haar features [44], and Gabor wavelet coefficients [46]. The geometry-based global features are mainly based on the landmark points around eyes, mouth, and noses [37, 38]. The hybrid features usually refer to the features that combine two or more of the engineered features [10]. For the learning-based features, most methods are based on deep neutral networks [31, 36]. After feature extraction, in the second stage, the extracted features are fed into a supervised classifier[18, 22, 36] to train an expression recognizer for FER. Although a handful of methods on the FER have been proposed, most of these studies are conducted in "lab-controlled" environment, i.e., the faces are captured under laboratory conditions and the expressions are deliberately acted. Different from existing methods, we mainly focus on the FER in the wild, which is more challenging because the facial images are usually collected from real scenarios with spontaneous expressions and poses, and complex backgrounds.

Domain Adaptation. Domain adaptation is a very active research area and has been widely studied for multimedia data analysis. Its major issue is how to deal with domain shift, which refers to the situation where data distribution differs between source and target domain, causing the classifier learnt from source domain to perform poorly on target domain. A large number of domain adaptation algorithms have been proposed to address this issue [13, 30, 43, 48], and the key focus is to learn domain-invariant feature representations. For example, Bousmails et al. [4] adopt domain separation networks that explicitly model the unique characteristics for each domain, so that the invariance of the shared feature representation can be improved. Timnit et al. [13] study fine-grained domain adaptation to overcome the domain shift between easily acquired annotated images and the real world. They show that while the first layers of a Convolution Neural Network (CNN) can learn general features, the features from the last layers are more specific and less transferable. Therefore the CNN needs to be fine-tuned on sufficient labeled target data to achieve domain adaptation. Very recently, attention maps have been studied as a mechanism to transfer knowledge [51]. Its basic idea is as follows. Assume that we have a test image x, a target expression k, a trained CNN model, and the corresponding J feature maps A^j of a CNN layer. The image x is first forwardly propagated through the trained CNN model, then a spatial attention map is constructed by computing statistics of the feature maps across all the J channel dimension:

$$F(A) = \sum\nolimits_{j=1}^{J} |A^j| \tag{1}$$

The key focus of this work is to learn knowledge transfer from a deeper network to a shallower network within the same domain. More details of the activation-based attention transfer can be found in [51]. Li et al. [30] extend this work for cross-domain knowledge transfer from web images to videos.

Generative Adversarial Network. The Generative Adversarial Network (GAN) is introduced in [15]. The goal is to train generative models through an objective function that implements a minimax two-player game between a discriminator D (a function aiming to tell apart real from fake input data) and a generator G (a function that is optimized to generate input data from noise that 'fools' the discriminator). Through this game, the generator and discriminator can both improve themselves. Concretely, D and G play the game with a value function $V(D, G)$:

$$\begin{aligned} \min_G \max_D V(D,G) =& E_{x\sim p_d(x)}[\log D(x)]+ \\ & E_{z\sim p_z(z)}[\log(1-D(G(z)))] \end{aligned} \tag{2}$$

The two parts, G and D, are trained alternatively. Researchers have successfully applied GAN-based approaches to various applications such as image generation [7, 32, 62], object detection [29] and image classification [3, 47]. Recent models [40, 60] adopt conditional GAN [35] for image-to-image translation problems, but they require input-output image pairs for training, which is in general not available in domain adaptation problems. More recent methods focus on incorporating consistent constraints in the image generation process by translating images back to their original domains while ensuring that they remain identical to their original versions. For example, Zhu et al. [67] produce compelling image translation results by Cycle-Consistent Adversarial Network (CycleGAN) such as generating photorealistic images from impressionism paintings or transforming horses into zebras at high resolution using the cycle-consistency loss. Our motivation comes from such findings about the effectiveness of the cycle-consistency loss, but differs from them. We propose an auxiliary domain guided cycle-consistent adversarial attention transfer learning model for simultaneous facial image synthesis and FER in the wild. On the one hand, we can distill knowledge from the sufficient facial images in source domain through attention transfer. On the other hand, labeled facial images are synthesized by harnessing information from large-scale unlabeled web facial images in auxiliary domain.

3 PROPOSED METHOD

In this section, we first give a brief overview of the proposed CycleAT model. We then describe the learning process and show the difference with existing models.

3.1 Cycle-Consistent Adversarial Attention Transfer for FER

As previewed in Section 1, it is an onerous task to collect large-scale facial expression images in the wild with correct labels. However, it is easier to build such dataset under laboratory controlled environment. Therefore, we formulate the FER in the wild as a domain

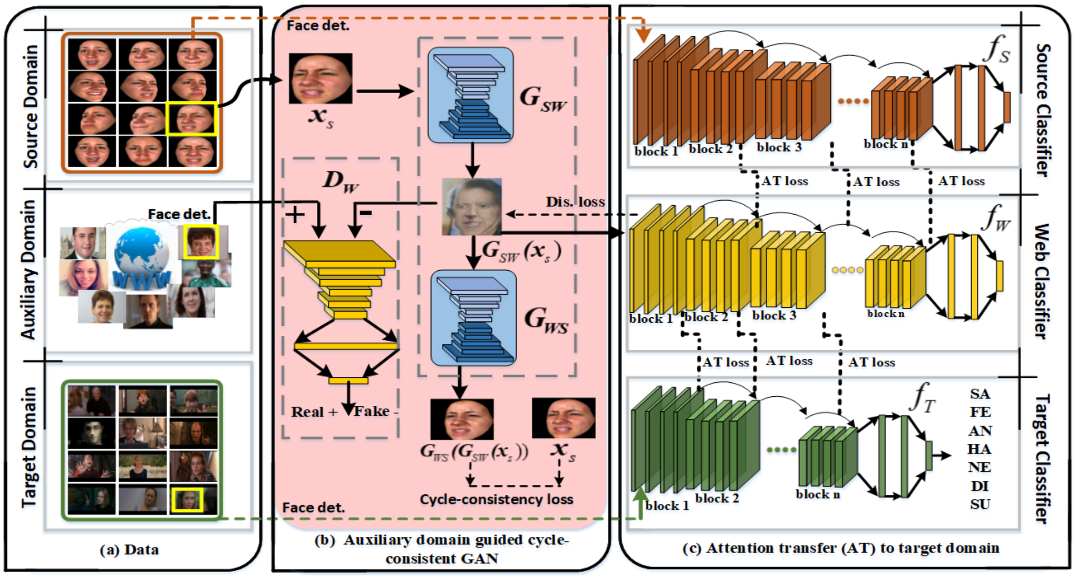


Figure 3: The architecture of the proposed cycle-consistent adversarial attention transfer model for simultaneous facial image synthesis and FER in the wild. The part (a) shows data from source, auxiliary, and target domain. The part (b) is for cycle-consistent adversarial image synthesis guided by auxiliary data. Here, for simplicity, we only show the generated process from source domain to auxiliary domain. It is the same for the process from auxiliary domain to source domain. The part (c) shows the attention transfer module.

adaptation problem. Here, the dataset captured under laboratory controlled environment is source domain, and facial images in the wild with limited samples are used as target domain. However, it is challenging to bridge source domain and target domain due to their representation gaps, such as background, color, lighting, poses. To deal with this issue, we resort to large-scale unlabeled facial images in the web as auxiliary domain, and propose an auxiliary data guided cycle-consistent adversarial attention transfer model for simultaneous facial images synthesis and facial expression recognition in the wild.

Figure 3 shows the overall pipeline of the proposed CycleAT model which transfers attention knowledge from the labeled images in source domain to target domain by harnessing information from large-scale unlabeled web facial images in auxiliary domain. Specifically, before passing an image into our model, we first perform face detection using a lib face detection algorithm with 68 landmarks [50]. After the preprocessing, we pre-train a classification model f_S on source domain. Then a facial image generator is learnt between the source domain and auxiliary domain by a GAN-based method with cycle-consistency and discriminative loss. Specifically, we introduce a mapping G_{SW} from source domain to auxiliary images and train it to produce facial images that fool an adversarial discriminator D_W. Then, we introduce another mapping G_{WS} constrained by a cycle-consistency loss to guarantee that the generated image can be mapped back to the original image, which can preserve the local structural information of generated facial images. Furthermore, a classifier f_W is incorporated into the generator by exploiting the global attribute-level consistency, which can preserve the expression attributes in the generated facial images, thus ensuring the label consistency. Finally, a shallower classification model f_T is trained on the target domain by distilling attention knowledge from f_W. Thus, in our model, f_W is like a bridge that can alleviate the domain gap between source domain and target domain.

3.2 Learning

Cross-domain setting: There are three kinds of data used in our method, i.e., source domain, target domain, and auxiliary domain. In the source domain, we are given n_s facial images X_S with associated labels Y_S. Similarly, the target domain consists of n_t facial images X_T with associated labels Y_T. The number of labeled images in the source domain is much larger than the number of labeled images in the target domain, i.e., $n_t \ll n_s$. All the labeled images belong to K expressions. In the auxiliary domain, it contains large-scale web images. Here, it has n_w facial images X_W without labels.

Source domain learning: We start with learning a source model f_S that can perform the classification task on source data. Residual network following the architecture of [51] is adopted here, taking X_S as inputs. The images belong to K categories. We use a typical softmax cross-entropy loss for the source model f_S, which corresponds to

$$L_S(f_S, X_S, Y_S) = \mathrm{E}_{(x_s, y_s)\sim(X_S, Y_S)} - \sum_{k=1}^{K} 1[k = y_s] log(\sigma(f_S^{(k)}(x_s))), \tag{3}$$

where σ denotes the softmax function. While the learnt model f_S will perform well on the source data, typically large domain shift between the source and target domain leads to reduced performance when evaluating on target data. To mitigate the effect of domain shift, we learn to synthesize samples from source to auxiliary images, which share similar content distribution with the target images.

Auxiliary domain learning: In this part, we first introduce a mapping from source to web facial images G_{SW} and train it to produce images that fool an adversarial discriminator D_W. Conversely, the adversarial discriminator attempts to classify the real web data from the source generated data. This corresponds to the loss function

$$
\begin{aligned}
L_{GAN}(G_{SW}, D_W, X_W, X_S) =& \mathrm{E}_{x_w \sim X_W}[log D_W(x_w)] + \\
& \mathrm{E}_{x_s \sim X_S}[log(1 - D_W(G_{SW}(x_s)))]
\end{aligned}
\tag{4}
$$

However, with large enough capacity, a network can map the face images in the source domain to any random permutation of images in the auxiliary domain. As a result, it is undesirable in the FER task, where we have to ensure the quality of the generated faces. Thus, we introduce another mapping from web to source G_{WS} and train it according to the same GAN loss, i.e.,

$$
\begin{aligned}
L_{GAN}(G_{WS}, D_S, X_S, X_W) =& \mathrm{E}_{x_s \sim X_S}[log D_T(x_s)] + \\
& \mathrm{E}_{x_w \sim X_W}[log(1 - D_T(G_{WS}(x_w)))].
\end{aligned}
\tag{5}
$$

We then require that mapping a source sample from source to web and back to the source reproduces the original sample, thereby enforcing cycle-consistency and preserving local structural information of the facial images in source domain. In other words, we want $G_{WS}(G_{SW}(x_s)) \approx x_s$ and $G_{SW}(G_{WS}(x_w)) \approx x_w$. This is done by imposing an L_1 penalty on the reconstruction error, which is referred to as the cycle-consistency loss:

$$
\begin{aligned}
L_{cyc}(G_{SW}, G_{WS}, X_S, X_W) =& E_{x_s \sim X_S}[||G_{WS}(G_{SW}(x_s)) - x_s||_1] \\
& + E_{x_w \sim X_W}[||G_{SW}(G_{WS}(x_w)) - x_w||_1].
\end{aligned}
\tag{6}
$$

Futhermore, we add a discriminative loss into the GAN by training a classification model f_W on the generated facial images. In the case of generation, it can be used to penalize the generator loss, which is helpful for preserving the global attributes of the source data, thus ensuring the label consistency. In the case of classification, it attempts to classify the expression, bridging the domain gap between source domain and target domain. To take advantage of the labeled images in source domain, we train f_W by distilling knowledge from f_S. Unlike previous work that distills knowledge through class probabilities, we do so by attention transfer, which has been proven in [30] that attention is a more transferable feature compared with the features from the last layers of the networks. This corresponds to the loss function

$$
\begin{aligned}
& L_W(f_W, G_{SW}(X_S), Y_S) = \\
& \quad \mathrm{E}_{(G_{SW}(x_s), y_s) \sim (X_W{}', Y_S)} - \sum_{k=1}^{K} 1[k = y_i^s] log(\sigma(f_W^{(k)}(G_{SW}(x_s)))) \\
& \quad + \frac{\beta}{2} \sum_{j \in \mathrm{I}} || \frac{Q_S^j}{||Q_S^j||_2} - \frac{Q_W^j}{||Q_W^j||_2} ||_2 .
\end{aligned}
\tag{7}
$$

Here, the first item is a typical softmax cross-entropy loss, which is the same as Eq (3). The second item is the attention transfer loss where $Q_S^j = vec(F(A_S^j))$ and $Q_W^j = vec(F(A_W^j))$ are respectively the j-th attention maps pair of the classifier f_S and f_W in vectorized form, and $F(\cdot)$ is calculated according to Eq (1). $X_W{}'$ are the synthesized facial images from source domain to auxiliary domain. β is the weight of the attention transfer loss.

Taken together, the loss function for the cycle-consistent adversarial image synthesis and FER task is defined as in equation (8) by considering Eqs (4), (5), (6), and (7).

$$
\begin{aligned}
& L_{sum}(f_W, X_S, X_W, Y_S, G_{SW}, G_{WS}, D_S, D_W) \\
& \quad = L_W(f_W, G_{SW}(X_s), Y_s) + L_{GAN}(G_{SW}, D_W, X_W, X_S) \\
& \quad + L_{GAN}(G_{WS}, D_S, X_S, X_W) + L_{cyc}(G_{SW}, G_{WS}, X_S, X_W).
\end{aligned}
\tag{8}
$$

This ultimately corresponds to solving for a classification model f_W according to the optimization problem

$$
f_W^* = \arg \min_{f_W, G_{SW}, G_{WS}} \max_{D_S, D_W} L_{sum} . \tag{9}
$$

Target domain learning: Once we obtain the classification model for the synthesized facial images, we can better address the domain shift problem with limited training data in target domain by transferring attention knowledge from the deeper network f_W to a shallower network f_T, which is the same objective function as Eq (7),

$$
\begin{aligned}
L_T(f_T, X_T, Y_T) =& \mathrm{E}_{(x_t, y_t) \sim (X_T, Y_T)} - \sum_{k=1}^{K} 1[k = y_i^t] log(\sigma(f_T^{(k)}(x_t))) \\
& + \frac{\beta}{2} \sum_{j \in \mathrm{I}} || \frac{Q_W^j}{||Q_W^j||_2} - \frac{Q_T^j}{||Q_T^j||_2} ||_2 .
\end{aligned}
\tag{10}
$$

Ultimately, the optimization function for the target model f_T can be defined as

$$
f_T^* = \arg \min_{f_T} L_T . \tag{11}
$$

Based on the learnt f_T, the FER in the wild with limited training samples is implemented.

3.3 Discussion

There are numerous methods about domain adaptation and image generation. In this section, we comment on the differences of the proposed model with three most relevant methods in [20, 30, 67]. (1) In the CycleGAN [67], the cycle-consistency is proposed mainly for image generation, but is agnostic to any particular task. Different from the CycleGAN, our model explicitly incorporates a task-specific classifier to enforce the global attribute-level information. (2) The CyCADA [20] extends the CycleGAN for cross-domain image classification. The authors use the cycle-consistency loss to encourage the cross-domain transformation to preserve local structural information and a semantic loss to enforce semantic consistency. It is mainly used for digit classification and semantic image segmentation, which have sufficient training samples in both source domain and target domain. Different from the CyCADA, we only have limited training samples in target domain. Thus we incorporate attention transfer into our model to distill discriminative knowledge from a FER model trained with sufficient samples. (3) In [30], the authors transfer knowledge from web images for video recognition. Different from this method that directly utilizes a noisy collection of web images for recognition tasks, we use a variation of GAN to automatically generate facial images in the wild with the correct category, which can bridge the domain gap between the source domain and target domain.

4 EXPERIMENTAL RESULTS

In this section, we show experimental results of our method for FER in the wild and facial image synthesis. For the former task, we quantitatively evaluate the expression recognition performance. For the latter one, we show qualitative results of the generated facial images.

4.1 Datasets

To demonstrate the effectiveness of the proposed model, we perform extensive evaluations on a number of popular datasets. For all experiments, we use the 3D facial expression dataset BU-3DFE [49] as source domain, and the static facial expressions in the wild dataset SFEW [8] and EmotioNet [2] as target domain. The Web image dataset (auxiliary domain) is the combination of the Large-scale CelebFaces Attributes (CelebA) dataset [33] and the Labeled Faces in the Wild (LFW) dataset [21]. The images in CelebA and LFW are mainly collected from Internet. The details are as follows.

BU-3DFE: The BU-3DFE dataset has 100 subjects with 3D models and facial images. It contains images depicting seven facial expressions of Anger (AN), Disgust (DI), Fear (FE), Happiness (HA), Sadness (SA), Surprise (SU) and Neutral (NE). With the exception of NE, each of the six prototypic expressions includes four levels of intensity. We render 2D facial images from the 3D models at levels 3 and 4 of the expression, and in 143 discrete poses including 11 pan angles ($0^\circ, \pm5^\circ, \pm15^\circ, \pm25^\circ, \pm35^\circ, \pm45^\circ$), and 13 tilt angles ($0^\circ, \pm5^\circ, \pm10^\circ, \pm15^\circ, \pm20^\circ, \pm25^\circ, \pm30^\circ$), using the 3D range data. Consequently, we have $100 \times 6 \times 143 \times 2 + 100 \times 1 \times 143 \times 1 = 185,900$ facial images in total for our experiments. We randomly divide the 100 subjects into a training set with 80 subjects and a testing one with 20 subjects. As a result, the training set comprises $148,720$ facial images whereas the testing one comprises $37,180$ samples.

SFEW: The SFEW is a facial expression dataset in the wild with 95 subjects. It consists of 700 images (346 images in Set 1, 354 images in Set 2) extracted from movies covering unconstrained facial expressions, varied head poses, changed illumination, large age range, different face resolutions, occlusions, and varied focus. The images are labeled with AN, DI, FE, HA, SA, SU and NE.

EmotioNet: The EmotionNet is a facial expression dataset in the wild. This dataset is labeled by Action Units (AUs) and compound emotions (6 basic emotion categories and 10 compound emotion categories). Since our work focuses on the FER, we obtain $1,141$ facial images with 6 basic expressions of AN, DI, FE, HA, SA, and SU. Although the $1,141$ facial images in this dataset are also labeled with AUs, we only use the basic expression labels of them like the source data. In this way, we can clearly validate the effectiveness of our method on FER in the wild with limited samples. Each time we divide the $1,141$ facial images into a training set with 913 facial images and a testing one with 228 facial images for 5-fold independent cross-validation.

Web image Dataset: The Web image dataset consists of the CelebA [33] and LFW [21]. The CelebA contains $202,599$ face images of ten thousand identities from the Internet, with approximately 20 images per person on average. The LFW contains $13,233$ face images collected from the Internet with large intra-personal variations in poses and backgrounds. It contains $5,749$ people, only 85 have more than 15 images, and $4,069$ people have only one image.

4.2 Implementation Details

The network is constructed as shown in Figure 3. We first use the lib face detection algorithm with 68 landmarks [50] to crop out the faces, and resize them as 256×256. For the failed images, we manually crop the faces from them. As the number of training samples in the target domain is extremely scare, for each detected facial image in the training set, we select a certain number of images with similar low-level characteristics from the generated facial images. Specifically, we put all the generated images and target images into AlexNet pre-trained on ImageNet [26], and get features from the first convolutional layer. Next, the features are projected by TSNE [27], and t nearest samples are selected for each facial image in the training set by using k-Nearest Neighbor (KNN). Here, t is 8 and 3 for SFEW and EmotioNet, respectively. Although we select some training samples from the generated facial images, the number of the training data in the target domain is still much smaller than that in the source domain. To stabilize the training process, we design the network for the GAN based on the techniques in the CycleGAN [67]. Specifically, this network contains two stride-2 convolutions, 9 residual blocks, and two fractionally-stride convolutions with stride $\frac{1}{2}$. For the discriminator network, we use a 70×70 PatchGAN [28], which is adopted to classify whether the 70×70 overlapping image patches are real or fake. We adopt a 50-layer residual network for the classification models f_S and f_W according to the results as discussed in Section 4.3. The model is implemented by TensorFlow [1] and trained with ADAM optimizer [25], which is used with a learning rate of 0.0002 and momentum 0.5. All weights are initialized from a zero-centered normal distribution with a standard deviation of 0.02.

4.3 Results of FER on Source Domain

In this part, we investigate several different basic classification models on source domain, i.e., Convolutional Neural Network (AlexNet [26]), VGGNet-19 network [42], ResNet-38 [17], ResNet-50 [17], and ResNet-101 [17]. All networks are trained and validated using the same training and test subset of BU-3DFE dataset. The detailed results over each expression obtained from different methods are shown in Table 2. The average FER accuracy is reported in the last column of the table, which reveals that the ResNet-50 network works best. Comparison with the ResNet-50, the performance of ResNet-101 shows a small drop. Thus, in our experiment, the basic classification model is fixed as ResNet-50.

Table 2: Comparison results of the FER on source domain with different classification models.

Method / Emotion	AN	DI	FE	HA	NE	SA	SU	Ave.
AlexNet [26]	**75.59**	84.20	55.09	87.89	79.66	55.06	88.87	75.18
VGGNet-19 [42]	68.77	88.38	58.14	93.10	79.07	66.03	88.75	77.46
ResNet-38 [17]	74.05	**90.97**	60.29	93.57	71.10	66.94	90.36	78.15
ResNet-50 [17]	71.41	88.48	61.65	**94.01**	**80.74**	75.28	95.32	**80.98**
ResNet-101 [17]	72.43	87.78	**65.37**	88.98	77.82	**75.49**	**96.77**	80.66

4.4 Evaluation on Domain Shift

In any adaptation experiment, it is crucial to understand the nature of the discrepancy between the different sources of data, thus we first provide intuitive understanding of the domain shift between source domain and target domain, which is shown in Figure 4. We visualize the distribution of the features in the first convolutional layer of AlexNet pre-trained on ImageNet [26], which are then projected by TSNE, and we choose the first three dimensions having the biggest contributions. In Figure 4, the orange points denote the features in source set, and the yellow and blue ones denote the features in SFEW and EmotioNet, respectively. Clearly, there are large domain gaps between the source and target domains.

In order to quantify the domain shift between the source domain dataset BU-3DFE and the target domain datasets SFEW and EmotioNet, we train a source model (source-to-source) and find that the accuracy is relatively high when evaluating within the source domain (80.98%) as shown in Table 3. However, the performance

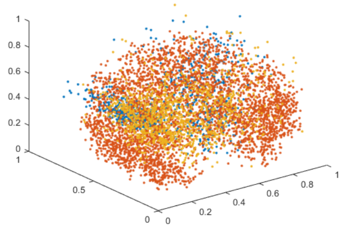


Figure 4: Feature distribution of the source (orange) and target (yellow, blue) domain.

catastrophically drops when evaluating within the target domain (SFEW and EmotioNet), which indicates the importance of the FER in the wild by using domain adaptation model.

Table 3: Results about domain shift for the FER.

Train	Test	Expressions							Ave.
		AN	DI	FE	HA	SA	SU	NE	
BU-3DFE	BU-3DFE	71.41	88.48	61.65	94.01	80.74	75.28	95.32	80.98
BU-3DFE	SFEW	17.21	23.52	18.93	55.62	21.32	9.83	10.30	22.40
BU-3DFE	EmotioNet	28.57	16.67	10.00	65.79	37.04	23.08	-	30.19

4.5 Model Analysis

To help analyze our CycleAT model and show the benefit of each module, we design several baseline methods as follows.

- **Source Domain to Target Domain (S2T):** This baseline uses the facial images in source domain (BU-3DFE) to pre-train the classification model ResNet-50, and then tests it on target domain, i.e., SFEW and EmotioNet. We use the output score from the last layer of the ResNet-50 to classify the given facial images in the wild, where voting is applied.
- **S2T_fine-tune:** This baseline uses the facial images in target domain to fine-tune the pre-trained classification model. Then we use the fine-tuned classifier to classify the facial images in target domain, and get the average FER accuracy over all expressions.
- **S2T_attention:** This baseline adds attention transfer to the training process of the target domain. Specifically, we first pre-train the classification model on BU-3DFE, and then transfer attention from it to the target classifier trained on SFEW and EmotioNet. By comparing it with S2T_fine-tune, we can evaluate the effect of attention transfer module.
- **S2T_attention&generator (S2T_att.&gen.):** This baseline adds facial image generator to the S2T_attention, and the attention and generator are trained separately. Specifically, we first train the image generator, and get the labeled facial images in the wild. Then, the generated facial images are used to train the classification model, and distill attention knowledge from the classifier trained on source domain. By comparing it with the S2T_attention, we can evaluate the effect of the proposed auxiliary data guided cycle-consistant generative model. Besides, we can also validate the effect of the global structural consistency in our method by comparing with the proposed CycleAT.

In Table 4, we show the differences between the above baseline methods and the propose CyCA-AT. The detailed comparison results are illustrated as follows.

Table 4: Differences among the evaluated models.

Method / Modules	fine-tune	attention transfer	image gen.	global consis.
S2T	-	-	-	-
S2T_fine-tune	√	-	-	-
S2T_attention	√	√	-	-
S2T_att.&gen.	√	√	√	-
CycleAT	√	√	√	√

Table 5: Comparison results on the SFEW dataset.

Method / Emotion	AN	DI	FE	HA	NE	SA	SU	Ave.
S2T	17.21	23.52	18.93	55.62	21.32	9.83	10.30	22.40
S2T_fine-tune	19.67	24.54	20.48	53.78	17.55	16.62	17.45	24.30
S2T_attention	27.43	25.65	19.87	**60.43**	11.32	15.72	**24.30**	26.39
S2T_att.&gen.	31.68	24.45	21.78	52.45	**24.87**	29.96	22.43	29.66
CycleAT	**34.91**	**25.95**	**23.61**	55.85	23.43	**32.00**	19.52	**30.75**

Table 6: Comparison results on the EmotioNet.

Method / Emotion	AN	DI	FE	HA	SA	SU	Ave.
S2T	28.57	16.67	10.00	65.79	37.04	23.08	30.19
S2T_fine-tune	30.00	23.33	14.00	63.16	48.15	20.00	33.10
S2T_attention	52.86	35.00	14.00	75.66	56.30	18.46	42.05
S2T_att.&gen.	57.14	**40.00**	**18.00**	78.95	59.26	23.08	46.07
CycleAT	**58.57**	38.33	**18.00**	**79.34**	**61.48**	**24.62**	**46.72**

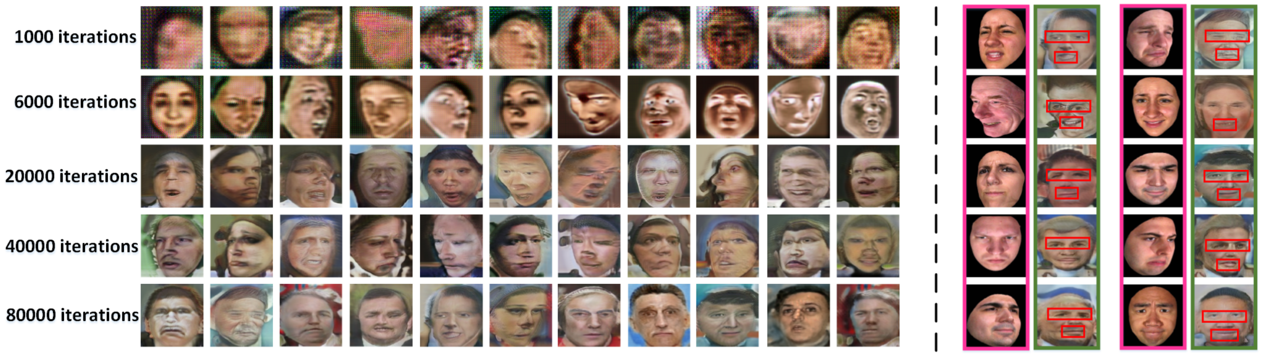
Comparison results on the SFEW: Table 5 shows the detailed comparison results over each facial expression between our method and four aforementioned baseline methods. Among the seven expressions, happiness is easier to be recognized. This is most likely because of the fact that the muscle deformations are relatively large compared with others, which also coincides with the findings of source domain as shown in Table 2. The average recognition accuracy shown in Table 5 indicates that our method achieves better results. Overall, it outperforms all the methods with a 1.11% to 8.35% improvement on the FER accuracy. Based on the results, it is clear that S2T_attention and S2T_att.&gen can separately improve the performance by 1.9% and 5.36% when compared with S2T_fine-tune. We attribute this to the web guided attention transfer and image generator strategy adopted in our method.

Comparison results on the EmotioNet: In Table 6, we show the comparison results of different methods on the EmotioNet dataset. As can be seen, the average recognition accuracy shown in the last column of this table indicates that the proposed model drastically improves performance, and the degree of improvement varies between 0.75% and 16.53%. Especially, when we add attention transfer strategy we can obtain a 8.95% promotion. The cycle-consistent adversarial image generator adopted in baseline S2T_att.&gen. further improves the FER accuracy to 46.07%. Our method achieves an average recognition accuracy of 46.72%. As shown in Table 6, we can also learn that among the six expressions, the happiness is easier to be recognized in all the methods. On the one hand it is because the muscle deformations of happiness are relatively large compared with others. On the other hand it is thanks to the imbalanced training data. There are 609 facial images with happiness expression among all the training samples in this dataset.

4.6 Comparison with State of the Arts

In this section, we compare the proposed model with state-of-the-art methods on the two target domains.

Comparison results on the SFEW: We compare our method with the current state-of-the-art results reported in [10] including MvDA, GMLDA, GMLPP, DS-GPLVM, and the baseline designed by the dataset creators [8] on the SFEW. The detailed results over each expression obtained from different methods are shown in Table 7.



(a) synthesized facial images in different iterations

(b) comparison with source domain

Figure 5: (a) The facial image generation process over different iterations. (b) The input images (pink) in source domain and their generated images (green) by using web facial images in auxiliary domain.

Table 7: Comparison with state of the arts on SFEW.

Method / Emotion	AN	DI	FE	HA	NE	SA	SU	Ave.
Baseline	23.00	13.00	13.90	29.00	23.00	17.00	13.50	18.90
MvDA	23.21	17.65	27.27	40.35	**27.00**	10.10	13.19	22.70
GMLDA	23.21	17.65	**29.29**	21.93	25.00	11.11	10.99	19.99
GMLPP	16.07	21.18	27.27	39.47	20.00	19.19	16.48	22.80
DS-GPLVM	25.89	**28.24**	17.17	42.98	14.00	**33.33**	10.99	24.70
CycleAT	**34.91**	25.95	23.61	**55.85**	23.43	32.00	**19.52**	**30.75**

Table 8: Comparison with three state-of-the-art methods on EmotioNet.

Method / Emotion	AN	DI	FE	HA	SA	SU	Ave.
Multi-SVM	35.71	18.33	10.00	69.08	34.81	21.54	31.58
AlexNet	38.57	20.00	14.00	75.13	45.93	20.00	35.61
VGG-16	37.14	25.00	10.00	77.76	49.63	**24.62**	37.36
CycleAT	**58.57**	**38.33**	**18.00**	**79.34**	**61.48**	**24.62**	**46.72**

The mean FER accuracy is reported in the last column. The results clearly show that our method outperforms all existing methods with a 6.05% to 11.85% improvement in terms of the FER accuracy. Note that all other models cannot achieve good performance in the surprise expression. However, the proposed model can significantly improve the performance attained by the cooperation of the generated images and attention transfer, which can distill attention knowledge from the FER model trained with sufficient samples.

Comparison results on the EmotioNet: We cannot find existing methods that conduct experiments on this dataset under the same conditions with us. Thus we compare our model with three state-of-the-art methods on the EmotioNet including multi-SVM [18], AlexNet [26], and VGG-16 [42]. All the methods make use of the same training and testing samples. As the number of training samples in the EmotioNet is scare, the AlexNet and VGG-16 used here are pre-trained on the ImageNet. The average recognition accuracies across all the expressions of each method are reported in Table 8. Clearly, our method can achieve the highest recognition accuracy of 46.72%. This may attribute to the generated facial images, which is able to not only reduce the domain shift between source domain and target domain, but also promote the target classifier by providing attention knowledge.

4.7 Qualitative Results

We visualize the image generation process and qualitative results of facial images synthesis from source domain (BU-3DFE) to auxiliary domain (Web images) in Figure 5. In Figure 5(a), each row shows some random generated samples under different iterations. Based on the figure, it is clear that the backgrounds in the web images can be gradually incorporated into the generated facial images. After several iterations, the generated images become more and more natural, which are as similar as the facial images in the target domain. In Figure 5(b), we randomly select several input facial images from the source domain, which are denoted with the pink rectangle. The corresponding synthesized facial images are shown in the green rectangle. By comparing the generated facial images with the ground truth, it is clear that the expression attributes (texture appearance around the eyes, mouth, and nose) have been preserved by our model, which ensures the label consistency between the original and generated facial images. More importantly, compared with the exaggerated facial expressions in the laboratory controlled environment, the generated facial images have more natural expressions, which are closer to the real scenarios.

5 CONCLUSIONS

We have presented a cycle-consistent adversarial attention transfer method that unifies cycle-consistent adversarial models with attention transfer strategy used in classification model. The CycleAT is suitable for large domain shift problems, especially for target domain with limited training samples. Besides, we experimentally validated our model on two benchmark facial expression datasets in the wild. Comparison results with the state-of-the-art methods show the superior performance of the proposed model. The proposed model has great potential to serve as general framework for domain adaptation that only limited training samples are available in target domain. Thus, in the future, we would extend current work to be a general framework, and apply it into other applications.

6 ACKNOWLEDGMENT

This work is supported by National Nature Science Foundation of China under Grant U1705262, 61672267, 61672268, 61572498, 61532009, 61728210, 61721004, 61751211, and 61472379, the Open Project Program of the National Laboratory of Pattern Recognition under Grant 201700022, the PKU-NTU Joint Research Institute (JRI) sponsored by a donation from the Ng Teng Fong Charitable Foundation, the Undergraduate Student Scientific Research Innovation Projects in Jiangsu Province under Grant KYCX17_1811, the Key Research Program of Frontier Sciences, CAS, Grant NO. QYZDJ-SSW-JSC039, and the Beijing Natural Science Foundation 4172062. The National Key Research and Development Program of China under Grant NO. 2017YFB1002804.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Carlos Fabian Benitez-Quiroz, Ramprakash Srinivasan, Aleix M Martinez, et al. 2016. EmotioNet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Computer Vision and Pattern Recognition (CVPR)*. 5562–5570.

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 7.

[4] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Neural Information Processing Systems (NIPS)*. 343–351.

[5] Wen Sheng Chu, Fernando De La Torre, and Jeffrey Cohn. 2017. Selective transfer machine for personalized facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 39, 3 (2017), 529–545.

[6] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. 2016. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 38, 8 (2016), 1548–1568.

[7] Guoxian Dai, Jin Xie, and Yi Fang. 2017. Metric-based generative adversarial network. In *ACM Multimedia (ACM MM)*. ACM, 672–680.

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

[9] Paul Ekman and Wallace V. Friesen. 1976. Pictures of facial affect. In *Palo Alto,CA,USA: Consulting Psychologists Press*.

[10] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. 2015. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE Transactions on Image Processing (TIP)* 24, 1 (2015), 189–204.

[11] Junyu Gao, Tianzhu Zhang, Xiaoshan Yang, and Changsheng Xu. 2018. P2t: Part-to-target tracking via deep regression learning. *IEEE Transactions on Image Processing (TIP)* 27, 6 (2018), 3074–3086.

[12] Weifeng Ge and Yizhou Yu. 2017. Borrowing treasures from the wealthy: deep transfer learning through selective joint fine-tuning. In *Computer Vision and Pattern Recognition (CVPR)*, Vol. 6.

[13] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. 2017. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 1358–1367.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*. 580–587.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Neural Information Processing Systems (NIPS)*. 2672–2680.

[16] Jinwei Gu, Xiaodong Yang, Shalini De Mello, and Jan Kautz. 2017. Dynamic facial analysis: from bayesian filtering to recurrent neural networks. In *Computer Vision and Pattern Recognition (CVPR)*.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[18] Nikolas Hesse, Tobias Gehrig, Hua Gao, and Hazım Kemal Ekenel. 2012. Multi-view facial expression recognition using local appearance features. In *International Conference on Pattern Recognition (ICPR)*. 3533–3536.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[20] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. 2017. CyCADA: cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213* (2017).

[21] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. 1, 2 (2007).

[22] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. Joint fine-tuning in deep neural networks for facial expression recognition. In *IEEE International Conference on Computer Vision (ICCV)*. 2983–2991.

[23] Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino. 2017. Generative attribute controller with conditional filtered generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*. 6089–6098.

[24] Pooya Khorrami, Thomas Paine, and Thomas Huang. 2015. Do deep neural networks learn facial action units when doing expression recognition. In *IEEE International Conference on Computer Vision (ICCV)*.

[25] Diederik Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NIPS)*. 1097–1105.

[27] Van Der Maaten L. 2014. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research* 15, 1 (2014), 3221–3245.

[28] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2016. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802* (2016).

[29] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. 2017. Perceptual generative adversarial networks for small object detection. In *Computer Vision and Pattern Recognition (CVPR)*.

[30] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2017. Attention transfer from web images for video recognition. In *ACM Multimedia (ACM MM)*. ACM, 1–9.

[31] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. 2014. Facial expression recognition via a boosted deep belief network. In *Computer Vision and Pattern Recognition (CVPR)*. 1805–1812.

[32] Si Liu, Yao Sun, Defa Zhu, Renda Bao, Wei Wang, Xiangbo Shu, and Shuicheng Yan. 2017. Face aging with contextual generative adversarial nets. In *ACM Multimedia (ACM MM)*. 82–90.

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*. 3730–3738.

[34] Yadan Lv, Zhiyong Feng, and Chao Xu. 2014. Facial expression recognition via deep learning. In *International Conference on Smart Computing*. 303–308.

[35] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *Neural Information Processing Systems (NIPS)* (2014), 2672–2680.

[36] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. 2012. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision (ECCV)*. Springer, 808–822.

[37] Ognjen Rudovic, Maja Pantic, and Ioannis Patras. 2013. Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35, 6 (2013), 1357–1369.

[38] Ognjen Rudovic, Ioannis Patras, and Maja Pantic. 2010. Regression-based multi-view facial expression recognition. In *International Conference on Pattern Recognition (ICPR)*. IEEE, 4121–4124.

[39] Enver Sangineto, Gloria Zen, Elisa Ricci, and Nicu Sebe. 2014. We are not all equal: personalizing models for facial expression analysis with transductive parameter transfer. In *ACM Multimedia (ACM MM)*. ACM, 357–366.

[40] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. 2017. Scribbler: controlling deep image synthesis with sketch and color. In *Computer Vision and Pattern Recognition (CVPR)*.

[41] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2017. Learning bases of activity for facial expression recognition. *IEEE Transactions on Image Processing (TIP)* 26, 4 (2017), 1965–1978.

[42] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

[43] Kihyuk Sohn, Sifei Liu, Guangyu Zhong, Xiang Yu, Ming-Hsuan Yang, and Manmohan Chandraker. 2017. Unsupervised domain adaptation for face recognition in unlabeled videos. (2017).

[44] Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson. 2008. Generating facial expressions with deep belief nets. In *Affective Computing*.

[45] Usman Tariq, Jianchao Yang, and Thomas S Huang. 2014. Supervised super-vector encoding for facial expression recognition. *Pattern Recognition (PR)* 46 (2014), 89–95.

[46] Ying li Tian, Takeo Kanade, and Jeffrey F Cohn. 2002. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 229–234.

[47] Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, Vol. 4. 7.

[48] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-Modal retrieval. In *ACM Multimedia (ACM MM)*. ACM, 154–162.

[49] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. 2006. A 3D facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 211–216.

[50] Shiqi Yu, Jia Wu, Shengyin Wu, and Dong Xu. 2016. Lib face detection. https://github.com/ShiqiYu/libfacedetection/. (2016).

[51] Sergey Zagoruyko and Nikos Komodakis. 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations (ICLR)*.

[52] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 31, 1

(2009), 39–58.

[53] Feifei Zhang, Qirong Mao, Xiangjun Shen, Yongzhao Zhan, and Ming Dong. 2018. Spatially coherent feature learning for pose-invariant facial expression recognition. *ACM Transactions on Multimedia Computing, Communications, and Application (ACM TOMM)* 14, 1 (2018), 27.

[54] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. 2018. Joint pose and expression modeling for facial expression recognition. In *Computer Vision and Pattern Recognition (CVPR)*. 3359–3368.

[55] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. 2017. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing (TIP)* 26, 9 (2017), 4193–4203.

[56] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. 2017. Multi-task Correlation Particle Filter for Robust Object Tracking. In *Computer Vision and Pattern Recognition (CVPR)*. 1–9.

[57] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. 2018. Learning Multi-task Correlation Particle Filters for Visual Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* PP, 99 (2018), 1–1.

[58] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. 2018. Robust Structural Sparse Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* PP, 99 (2018), 1–1.

[59] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, Jingwei Yan, and Keyu Yan. 2016. A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Transactions on Multimedia (TMM)* 18 (2016), 2528–2536.

[60] Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age progression/regression by conditional adversarial autoencoder. In *Computer Vision and Pattern Recognition (CVPR)*, Vol. 2.

[61] Rui Zhao, Quan Gan, Shangfei Wang, and Qiang Ji. 2016. Facial expression intensity estimation using ordinal information. In *Computer Vision and Pattern Recognition (CVPR)*. 3466–3474.

[62] Yiru Zhao, Bing Deng, Jianqiang Huang, Hongtao Lu, and Xian-Sheng Hua. 2017. Stylized adversarial autoEncoder for image generation. In *ACM Multimedia (ACM MM)*. 244–251.

[63] Wenming Zheng. 2014. Multi-view facial expression recognition based on group sparse reduced-rank regression. *IEEE Transactions on Affective Computing (TAC)* 5, 1 (2014), 71–85.

[64] Yuhui Zheng, Byeungwoo Jeon, Le Sun, Jianwei Zhang, and Hui Zhang. 2017. Student's t-hidden Markov model for unsupervised learning using localized feature selection. *IEEE Transactions on Circuits and Systems for Video Technology* (2017).

[65] Yuhui Zheng, Le Sun, Shunfeng Wang, Jianwei Zhang, and Jifeng Ning. 2018. Spatially Regularized Structural Support Vector Machine for Robust Visual Tracking. *IEEE Transactions on Neural Networks and Learning System* (2018).

[66] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. 2012. Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2562–2569.

[67] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*.