Unsupervised Domain Attention Adaptation Network for Caricature Attribute Recognition

Wen Ji[1†], Kelei He[2,3†*], Jing Huo[1*], Zheng Gu[1], and Yang Gao[1,3]
{hkl,huojing,gaoy}@nju.edu.cn
{jiwen,guzheng}@smail.nju.edu.cn

[1] State Key Laboratory for Novel Software Technology, Nanjing, China
[2] Medical School of Nanjing University, Nanjing, China
[3] National Institute of Healthcare Data Science at Nanjing University, Nanjing, China

Abstract. Caricature attributes provide distinctive facial features to help research in Psychology and Neuroscience. However, unlike the facial photo attribute datasets that have a quantity of annotated images, the annotations of caricature attributes are rare. To facility the research in attribute learning of caricatures, we propose a caricature attribute dataset, namely WebCariA. Moreover, to utilize models that trained by face attributes, we propose a novel unsupervised domain adaptation framework for cross-modality (i.e., photos to caricatures) attribute recognition, with an integrated inter- and intra-domain consistency learning scheme. Specifically, the inter-domain consistency learning scheme consisting an image-to-image translator to first fill the domain gap between photos and caricatures by generating intermediate image samples, and a label consistency learning module to align their semantic information. The intra-domain consistency learning scheme integrates the common feature consistency learning module with a novel attribute-aware attention-consistency learning module for a more efficient alignment. We did an extensive ablation study to show the effectiveness of the proposed method. And the proposed method also outperforms the state-of-the-art methods by a margin. The implementation of the proposed method is available at https://github.com/KeleiHe/DAAN.

Keywords: Unsupervised domain adaptation; Caricature; Attribute recognition; Attention

1 Introduction

Caricatures are facial drawings of human faces with exaggerating facial features. Studying the latent information conveyed by caricatures has been long to the neurologists, psychologists, and also the computer scientists. The recognition of caricatures indicates the mechanism of human thoughts, and the knowledge

[†] These authors contributed equally as co-first authors.

[*] Co-corresponding authors.

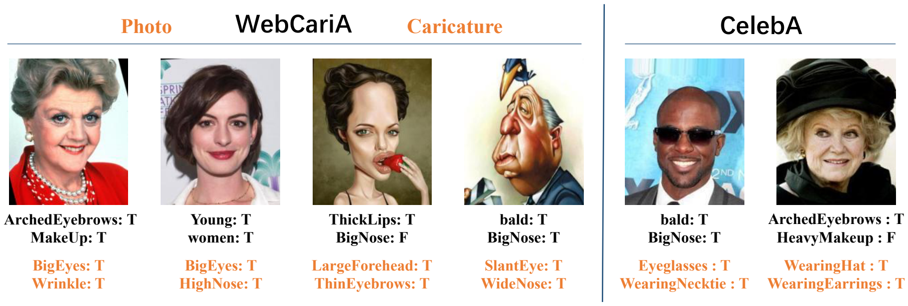

Fig. 1. The comparison of typical cases in WebCariA and CelebA datasets. First two rows denote the same attributes annotated in the WebCariA dataset and the CelebA dataset. The last two rows indicate the distinctive attributes annotated in the two datasets. 'T' indicates 'True', 'F' indicates 'False'.

learned by human during this task. Compared with photos, recognize the identities of caricatures may easier for humans [26, 27]. This indicates the most representative face features are not destroyed even the shape and appearance of faces have been largely changed. By contrast, they are usually harder for the machine learning methods to recognize and comprehend.

Recently, the research area of analyzing the caricatures in machine learning society has been raised, with several datasets are publicly released [1, 18, 14]. Most of the existing researches focus on face recognition [26, 18, 14, 33] and image generation [3, 4, 17] using these datasets. However, as lacked by the annotations of the caricature attributes, a more valuable task of attribute recognition, has rarely been touched. To solve this problem, in this paper, we introduce the WebCariA dataset, by extending a large caricature dataset 'WebCaricture' [14] with the annotation of fifty intrinsic face attributes. We hope it can boost this research area. Specifically, to help understand the intrinsic facial characteristics that are felt by human, the face attributes on WebCariA are purely facial characteristics without the non-face attributes, compared with the existing face attribute in photo datasets (e.g., the attribute of 'Eyeglasses' and 'WearingHat' in CelebA dataset). (See the comparisons of typical examples in Fig. 1) The details of the dataset will be further introduced in Section 3.

The face attribute recognition task has been solved well by deep convolutional networks [21, 7, 16, 31, 19]. It can be concluded as large-scale annotated face attribute datasets are already established, as the facial photos are easy to acquire. By contrast, the number of attribute annotated caricatures is small. Therefore, a natural idea is adapting a method that is trained on the annotated photos to the unannotated caricatures. This raises the problem of unsupervised domain adaptation from photos to caricatures for attribute recognition.

To solve the problem of cross-modality (i.e., face to caricature) attribute recognition on WebCariA dataset, we propose the domain attention adaptation network (DAAN), which has robust cross-domain adaptation ability for face attribute recognition. Specifically, to address the problem of large domain gap between photos and caricatures, DAAN has two main learning schemes to constrain the network, i.e., the inter-domain consistency learning and the intra-domain consistency learning. The inter-domain consistency learning scheme consisting of a pixel-level cross-domain image-to-image translator and the corresponding inter-domain consistency losses. The inter-domain consistency losses force the network to have consistent face attribute predictions on a certain image and its translated one. On the other hand, to align a certain image and a transferred image generated by the image of a different identity and domain property, we further propose the intra-domain consistency learning. It leverages both the feature and the attribute-aware attention map to build up the consistency between the two domains. Herein, we propose the attribute-aware attention-consistency learning by two observations: (1) the conventional feature-based consistency learning that align the distribution of the two domains in a global perspective is not efficient. (2) As a fine-grained classification task, face attributes are often revealed in a small region of the face, using attribute-aware attention will eliminate the noise conveyed by other parts. We did extensive ablation studies to show our proposed components, which can improve the discriminate ability of the network. The experiments also show our proposed method can outperform the state-of-the-art methods by a margin.

In this paper, our contributions are three-fold:

- We introduce the WebCariA dataset with the annotation of fifty intrinsic face attributes on caricatures.
- We propose a novel unsupervised attention adaptation framework for the recognition of attributes on unlabeled caricatures, which outperforms the state-of-the-art methods.
- We propose the attention-consistency learning which transfers the most task-discriminate features to achieve a more efficient adaptation.

2 Related Work

This work proposes an unsupervised domain adaptation method for attribute recognition in unpaired facial photos and caricatures. The related work can be concluded into three aspects: (1) The methods for face attribute recognition; (2) The methods for unsupervised domain adaptation in classification; (3) The generative adversarial learning-based image-to-image translation.

2.1 Face Attribute Recognition Methods

Face attribute recognition is a fine-grained classification task aiming to estimate facial characteristics. Previous works [24, 25, 41, 21, 7] based on convolutional neural networks have achieved satisfactory results. In the literature, early works mostly focus on single attribute learning [10, 36]. After several large face attribute datasets being released, e.g., CelebA and LFWA [21], more works attempt to study on multi-attribute estimation, formulating the inter-attribute relationships in one method. For example, Liu *et al.*[21] and Ding *et al.*[7] proposed to localize face regions for multiple attribute recognition. Typically, multi-task learning (MTL) with multiple classifiers is a common technique to boost the generalization ability of the method by exploiting the inter-attribute relationships. Han *et al.*proposed a deep MTL approach for multi-attribute estimation in heterogeneous faces. Lu *et al.*[23] learn a deep MTL framework that dynamically groups similar tasks together. Zhang *et al.*[38] divide the attributes into different groups according to the location of the attribute. The multi-task learning have been proved to be very effective for facilitating the prediction of face attributes [12, 23, 37, 2, 8, 28].

2.2 Unsupervised Domain Adaptation Methods in Classification

The scenario of unsupervised domain adaptation (UDA) arises when we aiming at constructing a model that learns from an annotated data distribution (i.e., source domain), and need it to generalize well on a different (but related) unannotated data distribution (i.e., target domain). Plenty of UDA methods for classification have been proposed in recent years. [6] We can roughly divide them into two categories: (1) Non-adversarial learning methods: A metric is often used as the objective to directly measure and minimize the discrepancy of high-level features between the source domain and the target domain. For instance, the work in [22] uses the maximum mean discrepancy (MMD) as the metric. Vazquez *et al.*[35] proposed to use a transductive SVM algorithm to solve the UDA problem. The non-adversarial methods have the ability to strongly align two different domains that is efficient for simple data distributions. (2) Adversarial learning methods: The discrepancy between the source and the target domain is usually large. Therefore, the adversarial learning-based methods try to minimize such discrepancy with the learning of the domain distributions. For example, Ganin *et al.*[9] proposes a domain adversarial neural network (DANN) for image classification which contains a loss for label prediction and a loss for domain classification. The maximum classifier discrepancy (MCD) [30] method builds up two classifiers to classify the source samples. Specifically, the two classifiers are forced to have different task-specific decision boundaries for the target sample. To make the method aware of a specific task, the Drop to Adapt (DTA) [20] method is proposed to learn robust and discriminate features by leveraging the adversarial dropout strategy, which supports the cluster assumption. However, these methods still cannot get satisfying performance as they only directly align the two domain distributions. Therefore, several works [29, 13] have adopted image-to-image translation into the UDA framework, with the assumption of the generated intermediate data helps domain alignment. Besides, previous methods align the features learned in two domains, that are redundant and not efficient.

2.3 Image-to-Image Translation

The goal of image-to-image translation is to learn the mapping between input images and output images. GAN has been widely utilized to solve the image-to-image translation problem, as it can learn the latent distribution of the images. For example, Isola *et al*.[15] have proposed the conditional adversarial networks (conditional GAN) for the image-to-image translation problem. Zhu *et al*.[40] used the cycle-consistent adversarial networks (CycleGAN) to tackle the problem without paired training examples. Kim *et al*.proposes a method namely U-GAT-IT [17] which tried to combine the attention mechanism with the generative adversarial networks. It prompts the generator to focus on areas that specifically distinguish between the two domains, and the discriminator to focus on the difference between an original image and transferred image in the target domain.

3 The WebCariA Dataset

To date, several caricature datasets [1, 18, 14] are publicly available for caricature recognition. However, due to lack of caricature attribute annotations, there are little attempts to solve the task of caricature attribute recognition. In order to promote the research of caricature attribute recognition and generation, we construct a face attribute dataset, namely WebCariA, by labeling all images in the WebCaricature [14] dataset. The dataset contains 6024 caricatures and 5974 photos of 252 people. Each image was labeled more than three times with fifty attributes of intrinsic face characters, and the final labels are determined by a voting strategy. The dataset is released at [*].

Different from the photo dataset CelebA [21] and LFWA [21], we did not mark the attributes that are changeable, such as 'WearingJewelry', 'WearingHats', etc. Instead, we labeled the attributes that can reflect the intrinsic characteristics of the faces such as 'BigNose', 'Bald', etc. (See typical cases compared to CelebA in Fig. 1). Besides, we hope the WebCariA dataset can also promote the research area of caricature generation.

4 Method

In this section, we first provide the pipeline of our unsupervised domain adaptation framework. Secondly, we introduce the integrated-domain generalization learning paradigm which consists of the inter-domain consistency learning and the intra-domain consistency learning. Finally, we describe the multi-task learning setting for more precise face attribute recognition.

4.1 The Unsupervised Domain Adaptation Framework

The overall framework is shown in Fig. 2. The goal of our method is to estimate face attributes on caricatures (i.e., the target domain) with the annotations of the attributes that are only given on photos (i.e., the source domain). Moreover, the images in source and target domain are unpaired, for which forms a typical unsupervised domain adaptation setting. Formally, let us denote the labeled

[*] https://cs.nju.edu.cn/huojing/WebCariA.htm

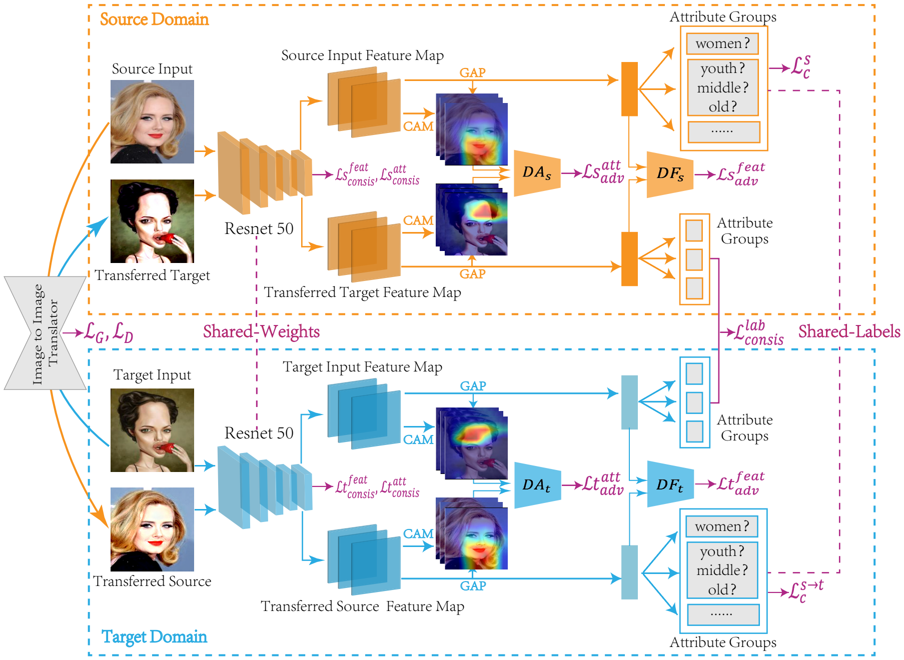

Fig. 2. The overall framework of our proposed unsupervised domain attention adaptation network (DAAN) for unsupervised face attribute recognition on caricatures.

source domain as $S = \{X_s, Y_s^i\}$, $i \in [1, N]$, where $\mathbf{x}_s \in X_s$ and $\mathbf{y}_s \in Y_s$ denote the source image and the corresponding label pair, N denotes the number of face attributes. The unlabeled target domain can be therefore denoted as $T = \{X_t\}$, where $\mathbf{x}_t \in X_t$ denotes the target image.

The domain gap and domain shift are the major problems for the cross-domain learning method. To solve the problems, our method adopted an integrated learning scene for learning both the inter- and intra-domain consistencies. The consistencies are learned by given constraints on the features. The inter-domain consistency learning firstly uses an unsupervised image-to-image translator to generate images from one domain to another by learning the influence of the style discrepancy of the two domains. The semantic information is preserved and the style information is transformed on the transferred images. Specifically, in this work, the image translation is bi-directional, where each image of the source domain will generate one transferred image in the target domain, and vise versa. The bi-directional image-to-image translation module in the framework can be regarded as to help build an intermediate data distribution between the source data and target data.

However, caricatures and photos have a large domain gap, i.e., with large style and shape variations. To constrain the semantic information conveyed by image and its generated one align in features, we introduce the label-consistency learning to minimize the prediction error generated on the two domains.

Obviously, the generated data (from the other domain) and the real data in a certain domain have smaller domain gap, and thus are easier to be aligned together. We further propose an attention adaptation followed by a feature adaptation learning scheme to close the features of the intra-domain images, i.e., the original image and the transferred image (from another domain).

For inference, we use a multi-task learning method to design the face attribute classifiers. By assuming that the attributes have mutually exclusions, we build attribute-groups to involve mutual attribute competitions. The final network is end-to-end, and all the modules we introduced can be trained simultaneously.

4.2 Inter-Domain Consistency Learning

Unsupervised image-to-image translation. In the cross-domain attribute recognition task, the distribution between the source and target domains are large. Photos are shots of the real-world that often contain rich illumination, texture and noise. And caricatures are facial drawings are often concise, with clear outlines and exaggerated facial features. In order to eliminate the domain gap between the photo and caricatures, we first utilize an unsupervised image-to-image translator to alleviate the style discrepancy between these two domains. The translator takes a source image $\mathbf{x}_s$ and a target image $\mathbf{x}_t$ as inputs, and translate them from one domain to the other. The obtained corresponding translated images are denoted as $\mathbf{x}_{s\to t}$ and $\mathbf{x}_{t\to s}$. Here we use the generator and discriminator proposed in [17] as the image-to-image translator $\mathbf{G}_{s\to t}$ and $\mathbf{G}_{t\to s}$. Therefore, the loss are denoted as $\mathcal{L}_G$, $\mathcal{L}_D$, respectively. In this way, the style difference between the source domain and the target domain is eliminated, making their domain distribution more consistent with each other. Then, the features of the four images are extracted by a feature extractor. The feature extractor *Resnet* takes $\mathbf{x}_s$ and $\mathbf{x}_{t\to s}$ as inputs, then the feature vectors $\mathbf{F}_s$ and $\mathbf{F}_{t\to s}$ is obtained after the global average pooling (GAP). The features $\mathbf{F}_t$ and $\mathbf{F}_{s\to t}$ in the target domain are acquired accordingly.

Label-consistency learning. By assuming the semantic information is not destroyed during image translation, we give $\mathbf{x}_s$ and $\mathbf{x}_{s\to t}$ the same labels to calculate the classification loss. Here, we use the cross-entropy loss to calculate the attribute estimating errors with the input of $\mathbf{x}_s$ and $\mathbf{x}_{s\to t}$,

$$\mathcal{L}_c^s = -\mathbb{E}_{\mathbf{x}_s,\mathbf{y}_s\sim S}\left[\mathbf{y}_s^\top \log C\left(\mathit{Resnet}\left(\mathbf{x}_s\right)\right)\right] \tag{1}$$

$$\mathcal{L}_c^{s\to t} = -\mathbb{E}_{\mathbf{x}_{s\to t},\mathbf{y}_s\sim T'}\left[\mathbf{y}_s^\top \log C\left(\mathit{Resnet}\left(\mathbf{x}_{s\to t}\right)\right)\right] \tag{2}$$

where C is the classifier, S' and T' are the distributions of the transferred image features. This also helps to make full use of annotations given in the source domain for semantic attribute consistency.

On the other hand, the labels of $\mathbf{x}_t$ and $\mathbf{x}_{t\to s}$ should be predicted as consistent as possible, therefore, we design the label-consistency loss for $\mathbf{x}_t$ and $\mathbf{x}_{t\to s}$ to control semantic attribute consistency,

$$\mathcal{L}_{consis}^{lab} = -\mathbb{E}_{\mathbf{x}_t \sim T, \mathbf{x}_{t\to s} \sim S'} \left[\| C\left(Resnet\left(\mathbf{x}_t\right)\right) - C\left(Resnet\left(\mathbf{x}_{t\to s}\right)\right) \|_2 \right] \quad (3)$$

$$\mathcal{L}_{inter} = \lambda_G \mathcal{L}_G + \lambda_D \mathcal{L}_D + \lambda_l \mathcal{L}_{consis}^{lab} \quad (4)$$

The proposed method achieves style consistency and semantic consistency between the source domain and the target domain by adopting the above-mentioned techniques. However, attribute recognition is a fine-grained classification task, these consistencies are built on a global perspective. They are not sensitive to the tiny differences between the two domains in the cross-domain face attribute recognition task. And the network still lack of the constraints between the transferred image and original image in one domain, e.g., $\mathbf{x_{s\to t}}$ and $\mathbf{x_t}$. We propose the intra-domain consistency learning to solve the problem.

4.3 Intra-Domain Consistency Learning

To make the source (/target) domain image and the transferred target (/source) image have stronger consistency constraints, we use the generative adversarial strategy to make the distribution of features align between the source domain and target domain. We introduce two discriminators D_{Fs} and D_{Ft} for the output features of the two domains, to construct the feature-level domain adaptation.

Under the adversarial learning setting, the loss $\mathcal{L}_{consis}^{feat}$ of the features can be written as,

$$\mathcal{L}_{consis}^{feat} = \mathbb{E}\left[\sum \mathbf{Y}_s \log\left(D_{Fs}\left(\mathbf{F}_{t\to s}\right)\right)\right] + \mathbb{E}\left[\sum \mathbf{Y}_{s\to t} \log\left(D_{Ft}\left(\mathbf{F}_t\right)\right)\right] \quad (5)$$

where $\mathbf{Y}$ is the originated domain label of the samples, and have the same shape to $\mathbf{F}$. If the samples come from source domain the elements of Y are 1 else 0. As the two pair of cross-domain features are jointly optimized, the final consistency loss in Eq. 5 aggregate the errors raised by the two features. And in the discriminate process, the intra-domain feature consistency for source domain try to achieve feature-level alignment between $\mathbf{F}_s$ and $\mathbf{F}_{t\to s}$, which can be defined as follows,

$$\mathcal{L}s_{adv}^{feat} = \mathbb{E}\left[\log\left(D_{Fs}\left(\mathbf{F}_s\right)\right)\right] + \mathbb{E}\left[\log\left(1 - D_{Fs}\left(\mathbf{F}_{t\to s}\right)\right)\right] \quad (6)$$

For the features in the target domain, the alignment can be written as,

$$\mathcal{L}t_{adv}^{feat} = \mathbb{E}\left[\log\left(D_{Ft}\left(\mathbf{F}_{s\to t}\right)\right)\right] + \mathbb{E}\left[\log\left(1 - D_{Ft}\left(\mathbf{F}_t\right)\right)\right] \quad (7)$$

The features convey the whole information of an image, that is rough and not efficient. It is proved that discriminative features often make more contribution to the classification task [34]. And a slight difference between the source and target domain can seriously affect the classification results in face attributes classification, because it is a fine-grained classification problem. Moreover, the distortion and exaggerations of caricature are often beyond realism which makes the gap of photos and caricatures very large. Therefore, we propose the attribute-attention alignment between the original images and translated images (from the other domain). The attention maps $\mathbf{A}_s$, $\mathbf{A}_{t\to s}$, $\mathbf{A}_t$ and $\mathbf{A}_{s\to t}$ of all inputs are calculated by class activation map (CAM) [39]. Take the source original image as example, $\mathbf{A}_s^i = \sum_{j=1}^n w_j^i \mathbf{F'}_{s,j}$, where $\mathbf{F'}_{s,j}$ is the jth channel of output feature maps $\mathbf{F}'_s$ which is a metric before GAP and w_j^i is the jth weight of the fully connected layer for predicting the ith category. Similar to the feature domain adaptation, we propose two attention-based discriminators, i.e., D_{As} which is performed on $\mathbf{A}_s$ and $\mathbf{A}_{t\to s}$, and D_{At} which is performed on $\mathbf{A}_t$ and $\mathbf{A}_{s\to t}$ to construct the attribute-attention domain adaptation.

The attribute attention-based adaptation can locate the decisive image region for classification. And with the help of the discriminators, the network will pay more attention to the inconsistent regions of the two domains to eliminate their tiny differences. Therefore, attention is suitable for domain alignment in face attributes classification.

The attention-consistency loss can be defined as follow:

$$\mathcal{L}_{consis}^{att} = \mathbb{E}\left[\sum \mathbf{Y'}_s \log\left(D_{As}\left(\mathbf{A}_{t\to s}\right)\right)\right] + \mathbb{E}\left[\sum \mathbf{Y'}_{s\to t} \log\left(D_{At}\left(\mathbf{A}_t\right)\right)\right], \quad (8)$$

where $\mathbf{Y'}$ is the label of the domain where the sample is coming from, its shape is same to $\mathbf{A}$. If the samples are drawn from the source domain, the elements of $\mathbf{Y'}$ is 1 else 0. And the attention-based intra-domain consistencies $\mathbf{A}_s$ and $\mathbf{A}_{t\to s}$ for the source domain are defined as,

$$\mathcal{L}s_{adv}^{att} = \mathbb{E}\left[\log\left(D_{As}\left(\mathbf{A}_s\right)\right)\right] + \mathbb{E}\left[\log\left(1 - D_{As}\left(\mathbf{A}_{t\to s}\right)\right)\right] \quad (9)$$

Similarly, D_{At} aligns the distributions between $\mathbf{A}_t$ and $\mathbf{A}_{s\to t}$,

$$\mathcal{L}t_{adv}^{att} = \mathbb{E}\left[\log\left(D_{At}\left(\mathbf{A}_{s\to t}\right)\right)\right] + \mathbb{E}\left[\log\left(1 - D_{At}\left(\mathbf{A}_t\right)\right)\right] \quad (10)$$

Overall, the final intra-domain consistency loss can be written as,

$$\begin{aligned} \mathcal{L}_{intra} &= \lambda_f(\mathcal{L}s_{consis}^{feat} + \mathcal{L}t_{consis}^{feat}) + \mathcal{L}s_{adv}^{feat} + \mathcal{L}t_{adv}^{feat} & (11)\\ &+\lambda_a(\mathcal{L}s_{consis}^{att} + \mathcal{L}t_{consis}^{att}) + \mathcal{L}s_{adv}^{att} + \mathcal{L}t_{adv}^{att} & (12) \end{aligned}$$

4.4 Multi-Task Attribute Recognition

For the task of caricature attribute recognition, the attributes are not independent. It is not suitable to treat them as a pure multi-class classification problem, as done by the conventional attribute recognition methods. Thus, the work in [38] partitioned all the attributes into different groups based on the global and local spatial regions.

The idea is natural but not very reasonable to the learning theory. By contrast, we observe that in most face attribute datasets, there are mutually exclusive relationships among the face attributes. For example, the attributes of young, middle-aged and old are related to the group of age. Obviously, the mutually exclusive relationships among the attributes provides stronger and more stable constraints to build up the classifiers. Herein, for the classifiers, we use one fully-connected layer to map the features to a certain number of classes for each attribute group. The cross-entropy loss is used to calculate the error. The results of each classifier is concatenated to obtain the final prediction result.

Finally, the objective function involving all the previous mentioned losses can be written as,

$$\mathcal{L} = \mathcal{L}_c^s + \mathcal{L}_c^{s\to t} + \mathcal{L}_{inter} + \mathcal{L}_{intra} \tag{13}$$

4.5 Implementation Details

We implement the proposed method using the open-source framework *PyTorch*. We use ResNet-50 [11] without the layer after the global average pooling layer as the feature extractor, and share the extractor through all inputs in this work. The discriminators are randomly initialized with a structure similar to [32], consisting five convolutional layers with size of {3 × 3,1,1} for {kernel, stride, padding}. We use the Stochastic Gradient Descent (SGD) algorithm to optimize the network, with a batch size of 40. The learning rate for the feature extractor and the classifiers is 0.05, with momentum of 0.9 and weight decay of 5×10^{-4}. The learning rate is decayed by the polynomial strategy [5] with the power of 0.75. For the discriminators, we use Adam optimizer with the learning rate of 1×10^{-4}. For the adversarial losses, we use the learning rates of 0.02, 0.1 and 0.1 for λ_l, λ_f and λ_a, respectively.

5 Experimental Results

Because annotated photos are often easy to acquire, in this paper, the method is built to estimate caricature attributes by the annotated photos, and using the photos in WebCariA dataset as the source domain. We use the metrics of accuracy (Acc.) and F1 score to measure the performance through the experiments. Please note that the attribute recognition task is very task-imbalanced for each attribute, the F1 score is more representative to show the real performance of the methods.

5.1 Ablation Study

We did extensive experiments for DAAN, including (1) the evaluation of the effectiveness of the MTL method, (2) the performance comparison of different network configurations of DAAN. The parameter analysis of DAAN is reported in *Supplementary Material*.

Firstly, we construct the 'Source Only' method, which is only trained on the source domain and evaluate on the target domain; and the 'Target Only' method, which is trained on the target domain and also evaluates on the target domain, forms a fully-supervised fashion. These two methods are usually used as the baselines to indicate the lower and upper bound of the performance of domain adaptation. For the proposed DAAN, we construct different configurations of DAAN, with or without the consistency learning of label, feature and attribute-aware attention.

Table 1. The effectiveness of the proposed multi-task learning strategy. 'MultiTask' denotes the network is learned with the attribute groups.

Method	Avg. Acc	Avg. F1
Source Only	0.8054	0.6770
Source Only(MultiTask)	0.8050	0.6922
DTA [20]	0.8100	0.6941
DTA(MultiTask) [20]	0.8076	0.7000
Target Only	0.8474	0.7358
Target Only(MultiTask)	0.8526	0.7601

The effectiveness of multi-task learning. Table 1 shows the performance in average accuracy and average F1 score through all attributions on the WebCariA dataset. The table suggests that using the multi-task learning strategy consistently improves the performance of the network. For the two baseline networks 'Source Only' and 'Target Only' which are purely unsupervised and supervised methods, group attributes with multi-task learning improve the average F1 score by 1.52% and 2.43%, respectively. To evaluate the generalization of the proposed MTL, we implement the state-of-the-art unsupervised domain adaptation method in [20] with a single classifier, denote as DTA, and with the proposed MTL, denote as DTA(MultiTask). The experiments show that, for the well designed unsupervised domain adaptation method, our proposed MTL still improves the F1 score by 0.59%.

Evaluation of different adaptation constraints. To evaluate the effectiveness of different adaptation constraints proposed by the method, we first construct DAAN with only one consistency constraint of label, feature and attribute-aware attention as 'DAAN-L','DAAN-F' and 'DAAN-A', respectively. Then, we compose two and three constraints into DAAN. The evaluations in average accuracy and F1 score are reported in 2. We make three conclusions according to the table: (1) All the constraints help to regularize the features under the UDA setting. (2) Among one factor constrained DAANs, the attention-consistency constraint performs better than the other two constraints, indicates the attention

Table 2. Performance comparison with different configurations of DAAN. (Bests are in Bold)

Method		label	feat.	att.	Avg. Acc	Avg. F1
DAAN	**-L**	✓			0.8122	0.7038
	-F		✓		0.8169	0.7094
	-A			✓	0.8147	0.7107
	-LF	✓	✓		0.8219	0.7181
	-LA	✓		✓	0.8212	0.7192
	-LFA	✓	✓	✓	**0.8239**	**0.7215**

is more efficient. (3) The combination of three constraints can further improve the model, and achieve the overall best performance. This reveals multiple constraints are useful to the UDA setting, which has more freedom in parameter space compared with fully-supervised learning.

5.2 Comparison with the State-of-the-art Methods.

We compare DAAN with several recent state-of-the-art methods for image classification. The compared methods include DAN [22], DANN [9], MCD [30] and DTA [20]. The compared methods are reimplemented on the WebCariA dataset without our proposed MTL by mutually exclusive attribute groups. The performance in average accuracy and F1 score is reported in Table 3, and the attribute-wise performance in the F1 score is illustrated in Fig. 3. The table suggests that the four state-of-the-art methods can both improve the performance of the network when adapting to another domain, and DTA performs best among the four methods in the second set of rows. So we further implement it with our MTL setting (denoted as 'DTA(MultiTask)'). As shown by the table, compared DAAN-LF with DTA, our proposed method gets a 1.19% improvement in average accuracy and a 2.4% improvement in average F1 score. Compared with DAAN-LFA with DTA and DTA(MultiTask), the improvements are 1.39%, 1.63% in average accuracy and 2.74%, 2.15% in average F1 score, respectively. This shows the effectiveness of DAAN compared with the current state-of-the-art methods. Furthermore, as shown in Fig. 3, DAAN-LFA performs consistently best in most attributes, when compared with Source Only, DAN and DTA method. It is worth noting that DAAN performs especially well in attributes that are hard to estimate (i.e., with F1 score under 0.6), e.g., SquareFace, HookNose, etc. This demonstrates the robustness of the proposed DAAN. We also analyze the generalization ability for DAAN on small benchmark datasets, which is reported in *Supplementary Material*.

5.3 Visualization of the Attribute-wise Attention Maps

The CAM reveals the attention of the network when it makes decisions on a certain class. We visualize the CAMs for the proposed method compared with the

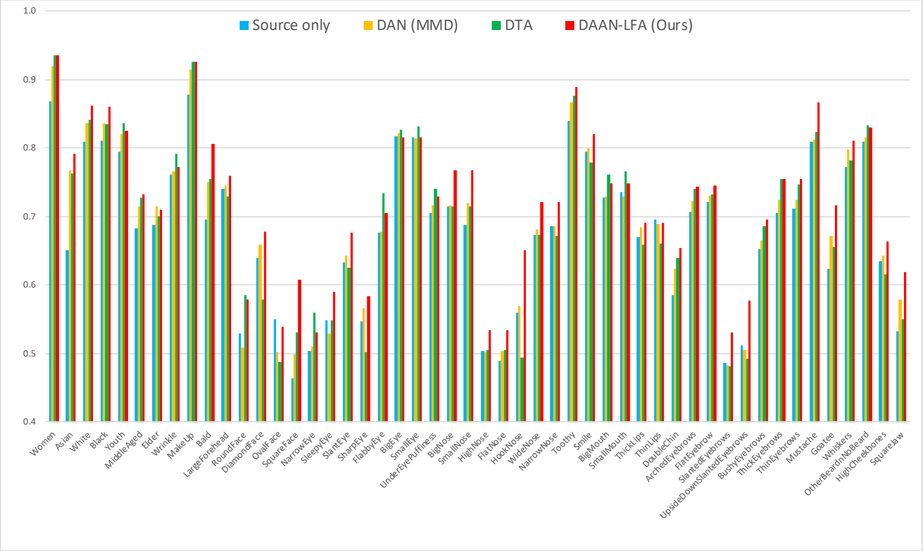

Fig. 3. The performance comparison in F1 score with Source Only, DAN (MMD), DTA and the proposed DAAN-LFA.

Table 3. Comparison with the state-of-the-art methods for performance in average Acc. and F1 score. As MCD method proposed two classifiers in the network, we report both of them separately.

Method	Avg. Acc	Avg. F1
Source Only	0.8054	0.6770
Source Only(MultiTask)	0.8050	0.6922
DAN [22]	0.8098	0.6921
DANN [9]	0.7783	0.6604
MCD [30]	0.7967 (0.7993)	0.6920 (0.6914)
DTA [20]	0.8100	0.6941
DTA(MultiTask) [20]	0.8076	0.7000
DAAN-LF(Ours)	0.8219	0.7181
DAAN-LA(Ours)	0.8212	0.7192
DAAN-LFA(Ours)	**0.8239**	**0.7215**
Target Only	0.8474	0.7358
Target Only(MultiTask)	0.8526	0.7601

state-of-the-art methods on the test set in Fig. 4. As the face attributes are often spatial-related, more precise attention regions for the specific attribute indicates higher predict performance. It is suggested by the figure that DAAN-LFA generates consistently best region activation through the cases. (More visualization results for DAAN is shown in *Supplementary Material*.)

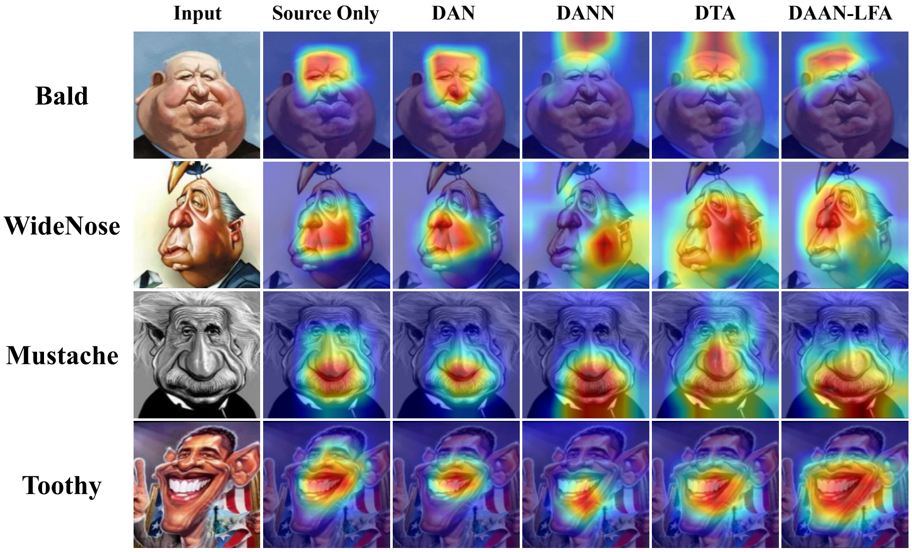

Fig. 4. The visualization of CAM for the proposed compared with the state-of-the-art methods on typical cases. Obviously, the proposed DAAN-LFA consistently generates high quality CAMs of accurate positions and active region sizes.

6 Conclusions

In this paper, we extend the 'WebCaricature' with fifty intrinsic face attributes to prepare the WebCariA dataset, for boosting the research in caricature attribute recognition. Then, to solve the problem of lacking annotated caricature attributes, we propose an unsupervised domain adaptation framework, i.e., DAAN, to estimate the caricature attributes by training the network with only given the face attribute labels. DAAN integrated with both inter- and intra-domain consistency learning paradigms. Specifically, in DAAN, an image-to-image translator and a label-consistency learning constrain are contained to fill the gap between photos and caricatures. To align the domain between the transferred images and original images in one certain domain, we propose the attribute-aware attention-consistency learning constrain, which is more efficient than the feature consistency constrain. The experiments show our framework can outperform the state-of-the-art methods by a reasonable margin.

Acknowledgement

This work is supported in part by National Science Foundation of China under Grant No. 61806092, and in part by Jiangsu Natural Science Foundation under Grant No. BK20180326.

References

1. Abaci, B., Akgul, T.: Matching caricatures to photographs. Signal, Image and Video Processing **9**(1), 295–303 (2015)
2. Abdulnabi, A.H., Wang, G., Lu, J., Jia, K.: Multi-task cnn model for attribute prediction. IEEE Transactions on Multimedia **17**(11), 1949–1959 (2015)
3. Brennan, S.E.: Caricature generator: The dynamic exaggeration of faces by computer. Leonardo **40**(4), 392–400 (2007)
4. Cao, K., Liao, J., Yuan, L.: Carigans: Unpaired photo-to-caricature translation. ACM Transactions on Graphics **37**(6), 244 (2018)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
6. Csurka, G.: Domain adaptation for visual applications: A comprehensive survey. arXiv preprint arXiv:1702.05374 (2017)
7. Ding, H., Zhou, H., Zhou, S.K., Chellappa, R.: A deep cascade network for unaligned face attribute classification. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
8. Ehrlich, M., Shields, T.J., Almaev, T., Amer, M.R.: Facial attributes classification using multi-task representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 47–55 (2016)
9. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189 (2015)
10. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. IEEE transactions on pattern analysis and machine intelligence **35**(10), 2401–2412 (2013)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. He, K., Wang, Z., Fu, Y., Feng, R., Jiang, Y.G., Xue, X.: Adaptively weighted multi-task deep network for person attribute classification. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1636–1644. ACM (2017)
13. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: Proceedings of the 35th International Conference on Machine Learning (2018)
14. Huo, J., Li, W., Shi, Y., Gao, Y., Yin, H.: Webcaricature: a benchmark for caricature recognition. arXiv preprint arXiv:1703.03230 (2017)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
16. Jacob, L., philippe Vert, J., Bach, F.R.: Clustered multi-task learning: A convex formulation. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 21, pp. 745–752. Curran Associates, Inc. (2009), http://papers.nips.cc/paper/3499-clustered-multi-task-learning-a-convex-formulation.pdf
17. Kim, J., Kim, M., Kang, H., Lee, K.H.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: International Conference on Learning Representations (2019)

18. Klare, B.F., Bucak, S.S., Jain, A.K., Akgul, T.: Towards automated caricature recognition. In: 2012 5th IAPR International Conference on Biometrics (ICB). pp. 139–146. IEEE (2012)
19. Kumar, A., Daume III, H.: Learning task grouping and overlap in multi-task learning. ICML (2012)
20. Lee, S., Kim, D., Kim, N., Jeong, S.G.: Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 91–100 (2019)
21. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
22. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International conference on machine learning. pp. 97–105 (2015)
23. Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., Feris, R.: Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5334–5343 (2017)
24. Luo, P., Wang, X., Tang, X.: Hierarchical face parsing via deep learning. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2480–2487. IEEE (2012)
25. Luo, P., Wang, X., Tang, X.: A deep sum-product architecture for robust facial attributes analysis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2864–2871 (2013)
26. Mauro, R., Kubovy, M.: Caricature and face recognition. Memory & Cognition **20**(4), 433–440 (1992)
27. Perkins, D.: A definition of caricature and caricature and recognition. Studies in Visual Communication **2**(1), 1–24 (1975)
28. Rudd, E.M., Günther, M., Boult, T.E.: Moon: A mixed objective optimization network for the recognition of facial attributes. In: European Conference on Computer Vision. pp. 19–35. Springer (2016)
29. Russo, P., Carlucci, F.M., Tommasi, T., Caputo, B.: From source to target and back: Symmetric bi-directional adaptive gan. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
30. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3723–3732 (2018)
31. Smith, V., Chiang, C.K., Sanjabi, M., Talwalkar, A.S.: Federated multi-task learning. In: Advances in Neural Information Processing Systems. pp. 4424–4434 (2017)
32. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472–7481 (2018)
33. Valentine, T., Lewis, M.B., Hills, P.J.: Face-space: A unifying concept in face recognition research. The Quarterly Journal of Experimental Psychology **69**(10), 1996–2019 (2016)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)

35. Vázquez, D., López, A.M., Ponsa, D.: Unsupervised domain adaptation of virtual and real worlds for pedestrian detection. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). pp. 3492–3495. IEEE (2012)
36. Wang, X., Guo, R., Kambhamettu, C.: Deeply-learned feature for age estimation. In: 2015 IEEE Winter Conference on Applications of Computer Vision. pp. 534–541. IEEE (2015)
37. Wang, Z., He, K., Fu, Y., Feng, R., Jiang, Y.G., Xue, X.: Multi-task deep neural network for joint face recognition and facial attribute prediction. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. pp. 365–374. ACM (2017)
38. Zhang, Y., Shen, W., Sun, L., Li, Q.: Position-squeeze and excitation module for facial attribute analysis. BMVC (2018)
39. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
40. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
41. Zhu, Z., Luo, P., Wang, X., Tang, X.: Multi-view perceptron: a deep model for learning face identity and view representations. In: Advances in Neural Information Processing Systems. pp. 217–225 (2014)