The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

A Deep Cascade Network for Unaligned Face Attribute Classification

Hui Ding,[1] **Hao Zhou,**[2] **Shaohua Kevin Zhou,**[3] **Rama Chellappa**[4]
[1,2,4]University of Maryland, College Park
[3]Siemens Healthineers, New Jersey

Abstract

Humans focus attention on different face regions when recognizing face attributes. Most existing face attribute classification methods use the whole image as input. Moreover, some of these methods rely on fiducial landmarks to provide defined face parts. In this paper, we propose a cascade network that simultaneously learns to localize face regions specific to attributes and performs attribute classification without alignment. First, a weakly-supervised face region localization network is designed to automatically detect regions (or parts) specific to attributes. Then multiple part-based networks and a whole-image-based network are separately constructed and combined together by the region switch layer and attribute relation layer for final attribute classification. A multi-net learning method and hint-based model compression is further proposed to get an effective localization model and a compact classification model, respectively. Our approach achieves significantly better performance than state-of-the-art methods on unaligned CelebA dataset, reducing the classification error by 30.9%.

Introduction

Face attributes describe the characteristics observed from a face image. They were first introduced by Kumar et al. (2009) as mid-level features for face verification (Kumar et al. 2011) and since then have attracted much attention. The last few years have witnessed their successful applications in hashing (Li et al. 2015), face retrieval (Siddiquie, Feris, and Davis 2011), and one-shot face recognition (Jadhav, Namboodiri, and Venkatesh 2016). Recently, researchers have begun to investigate the possibility of synthesizing face images based on face attributes (Radford, Metz, and Chintala 2015; Yan et al. 2016).

Despite their wide applications, face attribute recognition is not an easy task. One reason is that recognizing different face attributes may require attentions to different regions of the face (Moran and Desimone 1985; Posner and Petersen 1990). For example, local attributes like *Mustache* could be recognized by just checking the region containing the mouth. Rest of the face region does not provide useful information and may even hamper this particular attribute recognition. However, recognizing global attributes like *Pale Skin* may require information from the whole face region. Most current studies do not pay special attention to this problem. They either detect facial landmarks and extract hand-crafted features from patches around them (Kumar et al. 2009; Berg and Belhumeur 2013) or train a deep network to classify the attributes by taking a whole face as input (Liu et al. 2015; Wang, Cheng, and Schmidt Feris 2016; Rudd, Günther, and Boult 2016; Hand and Chellappa 2017).

In this paper, we propose a learning-based method that dynamically selects different face regions for unaligned face attribute prediction. It integrates two networks using a cascade: a face region localization (FRL) network followed by an attribute classification network. The localization network detects face areas specific to attributes, especially those that have local spatial support. The classification network selectively leverages information from these face regions to make the final prediction.

For accurate face region detection, our localization network is constructed under a multi-task learning framework. The lower layers which are used to extract low level features are shared by all the tasks while the high-level semantics are learned separately. Moreover, a global average pooling is applied to force the network to learn location-sensitive information (Lin, Chen, and Yan 2014). Although the network is trained in a weakly-supervised manner with attribute labels only, the detected face regions are consistent with what one may expect. As a result, face alignment algorithms which are usually sensitive to occlusion, variations of pose and illumination are not needed.

For each face region (also called a part) detected by our localization network, we train a separate attribute classification network, called a part-based subnet. The localized face parts may not contain enough contextual information for predicting global attributes. Thus, a whole-image-based subnet is also trained. To combine the information from the part-based and whole-image-based subnets, a two-layer fully-connected classifier is built on top of the output attribute scores. The first layer is used to select the relevant subnet for predicting each attribute, while the second layer is designed to model the rich attribute relations. The integrated system is called the parts and whole (PaW) network.

Since the face region localization network is supervised by attribute labels, it is appealing to adapt its weights to initialize the subnets in PaW. However, features from the local-

ization network, which are mainly designed for localization purpose, are generally not very discriminative for attribute classification. To this end, a multi-net learning method is proposed. It utilizes a network with enhanced attribute classification capability to train the localization network to find a more discriminative solution.

A naive implementation of the PaW network is problematic since the number of total parameters increases linearly with the number of attributes, and the subnet adapted from the FRL network is not very compact. To jointly train the PaW network end-to-end, a hint-based model compression technique is further proposed. This not only leads to a compact model with only $11M$ parameters, but also reduces the training time significantly.

We applied the proposed method to CelebA dataset (Liu et al. 2015). With no use of alignment information, our method achieves an accuracy of **91.23**%, reducing the classification error by a significant margin of **30.9**% compared with state-of-the-art (Liu et al. 2015). Moreover, our designed model could select the most relevant face region for predicting each face attribute.

To summarize, the contributions of this paper are listed below:

- A weakly-supervised localization network is designed to accurately locate attribute regions.
- A hybrid classification network is proposed to dynamically choose the pertinent face regions for predicting different attributes.
- A hint-based model compression technique is explored to obtain a compact model.
- Performance of unaligned face attribute classification is significantly improved by the proposed method.

Related Works

Face Attribute Recognition Early works (Kumar et al. 2009; Berg and Belhumeur 2013) on face attribute recognition used manually defined face parts to extract features and then train a linear SVM classifier. This strategy though is well suited for near-frontal faces, is heavily dependent on the accuracy of landmark detection. Recently, with the emergence of large-scale data and deep neural networks, holistic methods (Liu et al. 2015; Wang, Cheng, and Schmidt Feris 2016; Huang et al. 2016) have produced better performance than the part-based method. Liu et al.(2015) noticed that a deep model pre-trained for face recognition implicitly learns attributes. Huang et al. (2016) employed a quintuplet loss to combat the imbalanced data distribution problem. These methods typically use the whole face image to train a deep network, ignoring the fact that different facial attributes have different attentional facial regions. This problem has been recently noticed in (Ehrlich et al. 2016; Murrugarra-Llerena and Kovashka 2017). Murrugarra-Llerena and Kovashka (2017) created human gaze maps for each attribute such that only features within the saliency maps are used for attribute recognition. Our method differs from the aforementioned approaches in the sense that *the face parts are localized automatically without relying on detected landmarks or human gaze data.* Moreover, our classification network can dynamically select the relevant face regions for predicting different attributes.

Weakly Supervised Object Localization Despite training with only image-level labels, recent works (Oquab et al. 2015; Zhou et al. 2016; Cinbis, Verbeek, and Schmid 2017) showed that deep Convolutional Neural Networks (CNN) have remarkable object localization ability. Zhou et al. (2016) proposed a class activation mapping method to localize the objects with class labels only. The design of our face region localization network is motivated by this work. However, to fully utilize the correlations among different face attributes, the localization network is designed in a multi-task learning framework.

Model Compression To obtain a compact model, several methods including network distillation (Bucilu, Caruana, and Niculescu-Mizil 2006), parameter pruning (LeCun et al. 1989) have been proposed. Recently, knowledge distillation (Hinton, Vinyals, and Dean 2015) has been shown to be very effective to teach a small student model. However, it can not be directly applied to our problem: the teacher net uses soft labels which contain rich ambiguous information to supervise the student net, while for attribute classification, the output has only one logit for each attribute. Thus, a new loss function based on hints is proposed to replace soft label supervision.

Proposed Method

The proposed method contains two networks: a localization network and an attribute classification network. An overview of the framework is shown in Figure 1. First, we adopt the multi-net learning method to train a face region localization (FRL) network. Then one attentional region is detected for each attribute by the FRL network, which is fed into the PaW network for attribute prediction. To train the PaW end-to-end, a hint-based method is further applied to compress the model. The details of the proposed approach are discussed below.

Face Region Localization (FRL) Network

One challenge in designing a face region localization algorithm is that we do not have the labeled regions available. Murrugarra-Llerena and Kovashka (2017) used human gaze to label the related region for each attribute, however, this is both time consuming and expensive. Inspired by the success in weakly supervised object localization (Zhou et al. 2016), we apply a global average pooling (GAP) network for the localization task, and train it in a weakly-supervised way where only face attribute labels are needed. In this network structure, a GAP layer is used to pool features from the last convolutional layer, and a fully-connected layer is followed to predict the attribute score. A localization heatmap, H_j, for the j-th attribute, is obtained by applying the class activation mapping method. $H_j = \sum_{i=1}^{N} w_{j,i} F_i, i = 1, ..., N$, where F_i is the output feature maps from the last convolutional layer and $w_{j,i}$ is the i-th weight of the fully connected layer for predicting the j-th attribute. N is chosen to be 32 in our experiments.

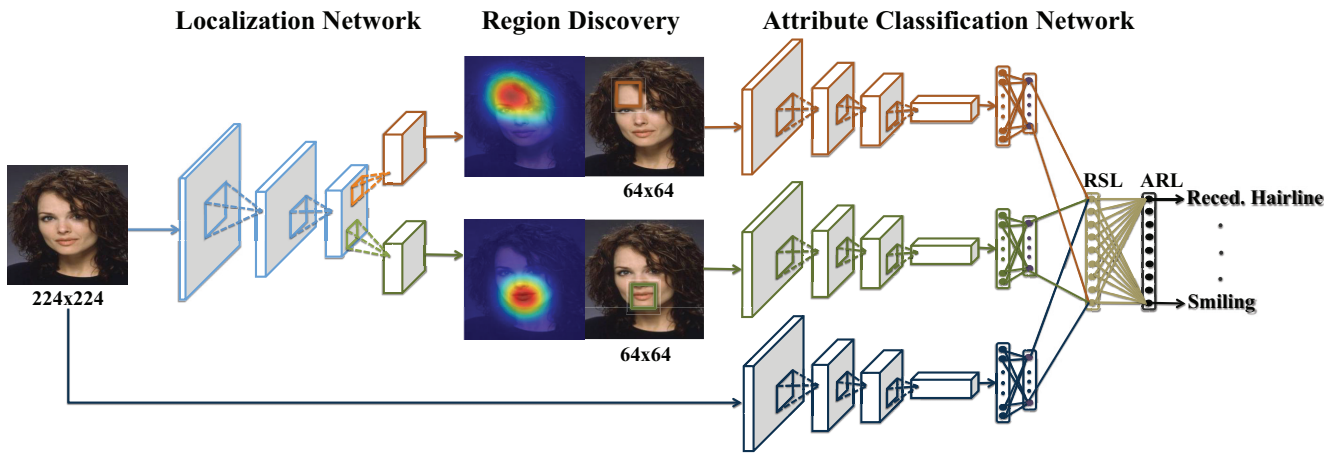

Figure 1: Overview of our face attribute recognition framework. It consists of a facial region localization (FRL) network and a Parts and Whole (PaW) classification network. The localization network detects a discriminative part for each attribute. Then the detected face regions and the whole face image are fed into the PaW classification network. The region switch layer (RSL) selects the relevant subnet for predicting the attribute, while the attribute relation layer (ARL) models the attribute relationships.

We design the FRL network using multi-task learning (Caruana 1998) strategy, where each attribute can be seen as one separate task. It has five VGGNet (Simonyan and Zisserman 2015) convolutional modules shared by all the attributes, and a domain adapted convolutional layer which has M different branches for each attribute, where $M = 40$ is the number of face attributes. The weights of the network are initialized from the VGG-Face CNN (Parkhi, Vedaldi, and Zisserman 2015) which is trained on a large-scale face recognition dataset.

Multi-Net Learning Since the supervision of the FRL network comes from the attribute tags, it is appealing to transfer its weights to the subnets in PaW for faster convergence and better performance. However, training the FRL net in a plain way leads to less discriminative features due to GAP regularization (Zhou et al. 2016). This is also verified in our experiments. To this end, a multi-net learning (MNL) method is proposed to boost the classification performance of the GAP feature, which yield improved final attribute classification.

The network architecture for MNL is shown in Figure 2. Except for the FRL network (blue and red boxes), another two fully-connected layers (gray box) are also attached to the output of the fifth convolutional module. We call it a classification branch because of its improved performance on the classification task compared with the localization branch. The idea is to simultaneously train the two different types of networks with the same attributes loss. Meanwhile the first several convolutional layers are constructed to be shared between them. The gradients from both classification and localization branches are backpropagated to the shared layers. This extra supervision from the classification branch regularizes the training process to search for a more discriminative solution. Interestingly, we find this simple learning strategy is beneficial for both branches in terms of classification performance. *After the multi-net training is completed, the classification branch is removed, and only the localization branch is kept for extracting attribute-specific heatmaps.*

To localize the face region, we upsample the location heatmap to the original image size 224×224, and find the position that corresponds to the maximum value. Then, a 64×64 patch centered around this position is cropped from the original image as the detected face region. We empirically found this patch size to be sufficient for most face parts. This process is repeated for each attribute and M face regions are obtained.

Attribute Classification Network

As shown in Figure 1, the proposed attribute classification network PaW contains M part-based subnets and one whole-image-based subnet. After getting the predicted attributes scores from each subnet, a two-layer fully-connected classifier is adopted to combine them.

Parts and Whole (PaW) Classification Network Suppose x_0 represents the whole face image, $x_1, ..., x_M$ represent face region related to each face attribute. $g_i, i \in 0, .., M$ represent the $(M + 1)$ subnets. Each x_i is first fed into its corresponding subnet g_i to predict the M attribute scores $\{s_{i,j}\}$, where $s_{i,j}$ represents the predicted score of the j-th attribute by the i-th subnet. The reason why we train each part-based subnet to predict M attributes instead of the one related to the input region is based on the observation that some attributes can usually be predicted by other ones (Torfason et al. 2016). The predicted scores $s_{i,j}$ will be fed into a region switch layer (RSL) which is designed as $r_j = \sum_{i=0}^{M} W_{ij} s_{ij}, j = 1, ..., M, W \in R^{(M+1)\times M}$ whose element in the i-th row and j-th column is W_{ij}. RSL adopts a group fully-connected structure, where the j-th output is

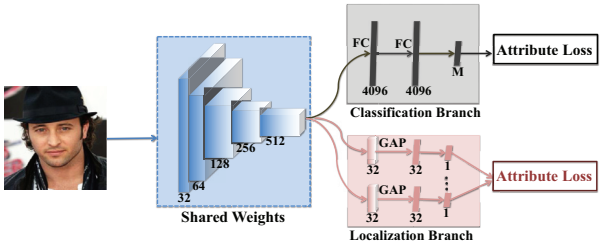

Figure 2: Multi-Net Learning.

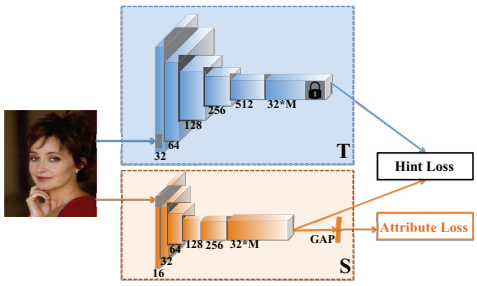

Figure 3: Hint-based Model Compression.

only connected with the j-th attribute scores predicted by all subnets. Especially, it could balance the scores from the part-based and whole-image-based subnets by putting more weight to the one that is more important. An attribute relation layer (ARL), which is a fully-connected layer, then takes these $r_j, j \in 1, ..., M$ as input to predict the final score for each face attribute. ARL here is used to further model the high correlations among the face attributes. The PaW network is trained end-to-end with the sigmoid cross entropy loss: $L_{attr} = \sum_{j=1}^{M} y_j \log o_j + (1 - y_j) \log(1 - o_j)$, where y_j's are the attributes labels, and o_j's are the outputs from the ARL layer.

Hint-based Model Compression Training the PaW network in a naive way is both memory demanding and time consuming, since the total number of network parameters increases substantially as the number of attributes becomes large, and the subnet architecture adapted from the FRL network is not very compact. To obtain a compact subnet model, we further propose a model compression technique. Motivated by (Abu-Mostafa 1992; Ding, Zhou, and Chellappa 2017), we design a hint loss to make the student net (SNet) reconstruct the feature maps from the teacher net (TNet). It can be expressed as:

$$L_{hint}(w) = ||T_k(I) - S_l(I, w)||_2, \quad (1)$$

where k (l) is the chosen layer of the teacher (student) net to transfer (add) supervision, w are the weights of the student net to be learned, and I is the input whole face image. The network architecture is shown in Figure 3. Besides the hint loss, the student network is also supervised by the attributes loss. Thus, the total loss function can be written as $L_S = \lambda_1 L_{hint} + \lambda_2 L_{attr}$. The FRL network trained by MNL is adopted as the teacher network to teach the whole-image-based subnet (or the student net). Since it is fully-convolutional and deeper layer generally captures high-level semantics (Zeiler and Fergus 2014; Escorcia, Carlos Niebles, and Ghanem 2015), we set the supervision layer k to be the teacher network's last convolutional layer. During training, the weights of the teacher network are frozen, and only the student network is learned. The whole training is carried out in two stages: first setting $\lambda_1 = 1, \lambda_2 = 0$, and training S with only the hint loss. In this way, the knowledge of the teacher network could help the student network find a good initialization. Then we set $\lambda_1 = 0, \lambda_2 = 1$ and train S with attribute loss only. After the whole-image-based subnet is learned, its weights are used to initialize all the part-based subnets in PaW.

Training Methodology

The whole training process is carried out as follows:

1. First, MNL is adopted to train the FRL network with superior classification performance;
2. Then hint-based compression method is applied to train a compact whole-image-based subnet g_0 using the learned FRL network as the teacher net.
3. Initialize each part-based subnet $\{g_i\}_{i=1}^{M}$ using the weights from g_0 and then train each subnet g_i independently using the corresponding attentional face region;
4. By fixing all the part-based subnets and the whole-image-based subnet, the RSL and ARL are learned;
5. Finally, the PaW network is fine-tuned by back-propagating errors from ARL to all the lower layers of the part-based subnets and the whole-image-based subnet.

All the subnets and the two layer fully-connected model are trained under the supervision of attribute labels. The third and forth steps initialize the classification model to be close to a good local minimum, which is important for the successful training of PaW.

Experiments

Dataset

We use the CelebA dataset (Liu et al. 2015) in our experiments, since it has been widely used for face attributes classification. It consists of 202,599 face images collected from the Internet and annotated with 40 binary attributes. As suggested in (Liu et al. 2015), 162,770 of these images are used for training, 19,867 and 19,962 are reserved for validation and testing respectively. Both unaligned and aligned sets are provided and we applied our method on the unaligned one (**uCelebA**). To conduct experiments on uCelebA, we use the publicly available face detector (Zhang et al. 2016) to detect faces. For 560 images which have no face detected, we use the provided landmarks to get the groundtruth bounding box (we empirically expand the minimum bounding box containing all landmarks twice to cover the neck and hair region). For 15,181 images with multiple faces detected, we select the bounding box that has maximum overlap with the

groundtruth bounding box. This is the only preprocessing step applied to the unaligned images.

Implementation details

We applied MNL to train the FRL network. The learning rate is fixed to be 0.0001, and the network is trained for 10 epochs with batch size of 128. The FRL network is then compressed with a learning rate of $1e^{-7}$ for the hint loss training and 0.0001 for the attribute loss training. The part-based subnets are trained for 15 epochs with the weights initialized from the whole-image-based subnet. After that, the RSL and ARL are trained with a learning rate of 0.1 with all subnets fixed. Finally, a learning rate of 0.001 is applied to train the PaW network in an end-to-end manner. Stochastic gradient descent (SGD) is used to train all the networks. The momentum and weight decay are set at 0.9 and 0.0005 for all the experiments respectively. Horizontal flipping is applied for data augmentation. We use Caffe (Jia et al. 2014) to implement our networks.

Ablative Analysis

Face Region Localization In this section, we evaluate the FRL network qualitatively. Figure 4 shows the location heatmaps corresponding to several attributes. We observe that the localized parts are quite semantically meaningful, even though some face images have large pose variations or under occlusion. For example, the eye area produces the highest response for the *Arched Eyebrow* attribute even though the woman wears sunglasses. While for the attribute of *Wavy Hair*, the network localizes the head region although the man wears a hat. We also examine it quantitatively in the **Classification Results** section to show that accurate region localization is essential for good classification results.

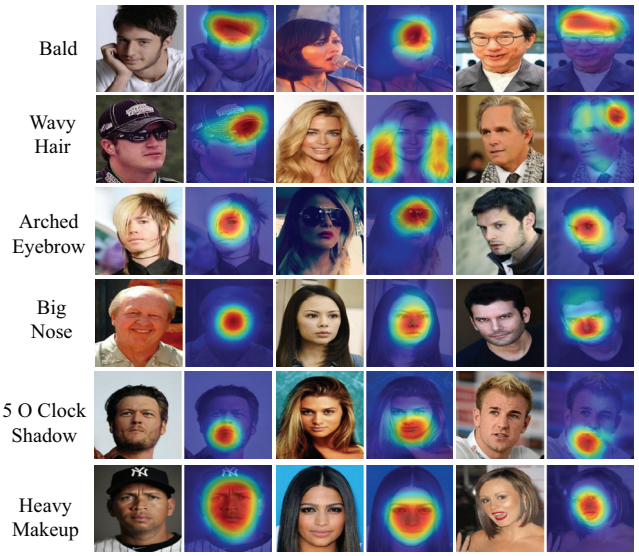

Figure 4: Location heatmaps from the face region localization network. Face regions that correlate with facial attributes are discovered.

Table 1: Average classification accuracy on uCelebA dataset.

Methods	Classif. Branch	Loc. Branch
Without MNL	-	91.01
MNL	91.05	**91.07**

Table 2: Fine-grained classification accuracy on CUB-200 dataset.

Methods	Classif. Branch	Loc. Branch
Without MNL on full image	-	67.40
MNL on full image	72.10	**71.66**
Without MNL on crop	-	71.90
MNL on crop	75.76	**76.03**

Multi-Net Learning In this section, we study the ability of MNL for obtaining a localizable and discriminative deep representation. Table 1 summarizes the attribute classification results from classification and localization branches. We find that MNL consistently improves the classification performance of the localization branch, achieving an accuracy of 91.07% vs. 91.01% with/without MNL.

To further test the proposed MNL, we applied it on the popular CUB-200-2011 dataset (Wah et al. 2011) for fine-grained object recognition. The dataset contains 11,788 images, with 5,994 images for training and 5,794 for testing. The network architecture is the same as the one used in uCelebA, except that the last layer is replaced with 200 output nodes (the number of classes). The weights are initialized from VGGNet (Simonyan and Zisserman 2015). Table 2 summarizes the results. We find that the localization branch performs worse than the classification branch, with almost 4% performance gap. After applying MNL, the accuracy of the localization branch is improved from 67.40% to 71.66% when using the full image. We also adopt the same localization technique as (Zhou et al. 2016) to identify the bounding box of the birds in both the training and testing sets. With the cropped bird images as training data, the performance of the localization branch is further improved from 71.90% to 76.03%. This further demonstrates that MNL is able to improve the discriminativeness of the GAP-based localization network.

Hint-based Model Compression In this section, we analyze the effectiveness of our model compression technique. To show the flexibility and robustness of our method, we experiment with three student nets (SNet1, SNet2 and SNet3) with different sizes. Table 3 shows the network architectures and their classification results. We use $s \times s \times n(t)$ to denote kernel size $s \times s$ with n output feature maps, where t is the number of repeated convolution modules. We observe that the proposed method is able to compress a deep network to a relatively shallow network, with little performance drop. For SNet3, which achieves an accuracy of 90.60%, the depth is shortened from 14 to 5, and the number of parameters is reduced from 19M to 0.27M.

To further compare our approach with existing methods, we also train our models on the *aligned* CelebA dataset. The

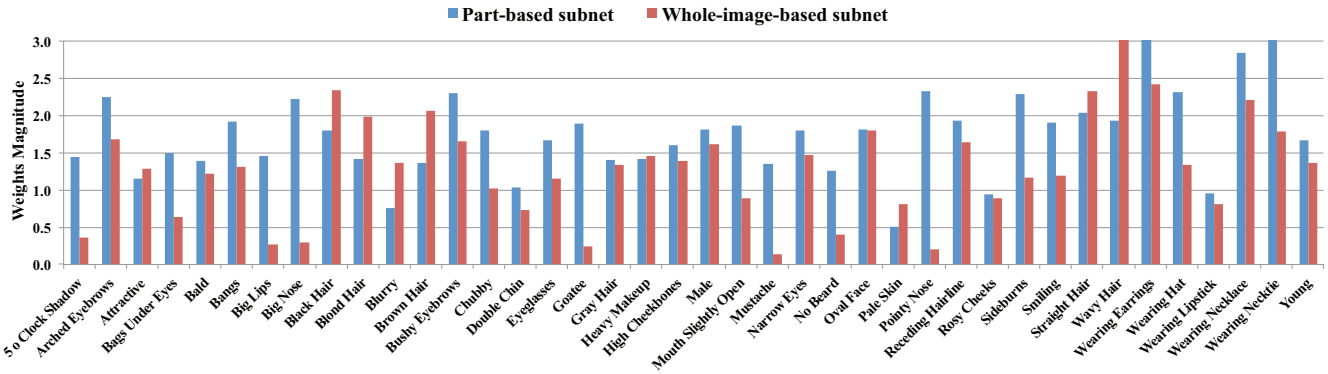

Figure 5: Visualization of the region switch layer weights. For each attribute, the blue and the red bar represent the weight values of RSL that corresponds to the part-based subnet and whole-image-based subnet respectively. It shows that the weights of the part-based subnets are higher for the local attributes. For global attributes, the whole-image-based subnet is assigned larger weight.

Table 3: Comparison of average accuracy and compactness between different compressed models on uCelebA dataset.

Layer	TNet	SNet1	SNet2	SNet3
Conv1	3x3x32(2)	3x3x32	3x3x32	3x3x16
Pool1	2x2x32	2x2x32	2x2x32	2x2x16
Conv2	3x3x64(2)	3x3x64	3x3x64	3x3x32
Pool2	2x2x64	2x2x64	2x2x64	2x2x32
Conv3	3x3x128(3)	3x3x128	3x3x128	3x3x64
Pool3	2x2x128	2x2x128	2x2x128	2x2x64
Conv4	3x3x256(3)	3x3x256	3x3x256	3x3x128
Pool4	2x2x256	2x2x256	2x2x256	2x2x128
Conv5	3x3x512(3)	3x3x512	3x3x512	1x1x1280
Conv6	3x3x1280	3x3x1280	1x1x1280	n/a
Classifier	GAP	GAP	GAP	GAP
	FC40	FC40	FC40	FC40
Accuracy	91.07	91.02	90.89	90.60
Param.	19M	6M	2M	0.27M

Table 4: Comparison of average accuracy and compactness on the aligned CelebA dataset.

Method	Accuracy	Param.
SOMP (Lu et al. 2017)-thin-32	89.96	0.22M
SOMP (Lu et al. 2017)-branch-32	90.74	1.49M
Low Rank (Denton et al. 2014)	90.88	4.52M
SNet3	**90.89**	**0.27M**

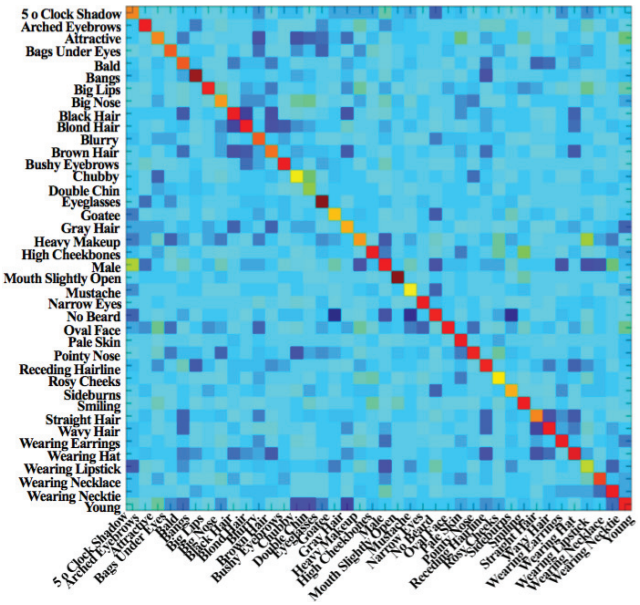

Figure 6: Attribute relation weights learned on uCelebA dataset. Red and yellow colors indicate high values while blue and green colors denote low values.

results are summarized in Table 4. We find that our SNet3 model achieves similar or better accuracy compared to these state-of-the-art methods, while being much more compact and thus faster.

PaW Classification Network In this section, we evaluate the classification performance of the proposed PaW network. Before showing the results, we first explore whether RSL assigns appropriate weights to different subnets for attribute prediction and whether ARL learns meaningful attributes correlations.

Face Region Selection We visualize the weights of RSL in Figure 5. Although each subnet predicts M attribute scores simultaneously, only the weights of the corresponding part-based subnet against the whole-image-based subnet are shown here. The weight magnitude indicates the importance of the subnet for predicting the attribute. Interestingly, we find that the part-based subnet related to the local attribute, *e.g. 5 o Clock Shadow* and *Bushy Eyebrows*, is always assigned the largest weight among the $M+1$ subnets. We also observe that for global attributes, *e.g. Attractive*,

Table 5: Performance comparison with state of the art methods on 40 binary facial attributes. The best results are shown in bold.

		5 o Clock Shadow	Arched Eyebrows	Attractive	Bags Under Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Hair	Heavy Makeup	High Cheekbones	Male
uCelebA	LNets+ANet (Liu et al. 2015)	91.00	79.00	81.00	79.00	98.00	95.00	68.00	78.00	88.00	95.00	84.00	80.00	90.00	91.00	92.00	99.00	95.00	97.00	90.00	87.00	98.00
	Part-only	93.90	81.86	81.88	84.07	98.72	95.71	70.63	83.48	87.97	95.16	95.83	87.53	91.73	95.05	95.92	99.46	97.19	97.93	90.26	86.20	96.65
	Whole-only	93.95	81.43	82.06	84.11	98.57	95.45	70.66	82.91	89.08	95.52	96.01	88.63	92.32	95.12	95.98	99.40	96.90	98.07	90.67	86.57	97.10
	PaW	**94.64**	**83.01**	**82.86**	**84.58**	**98.93**	**95.93**	**71.46**	**83.63**	**89.84**	**95.85**	**96.11**	**88.50**	**92.62**	**95.46**	**96.26**	**99.59**	**97.38**	**98.21**	**91.53**	**87.44**	**98.39**

		Mouth Slightly Open	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy Nose	Receding Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wearing Earrings	Wearing Hat	Wearing Lipstick	Wearing Necklace	Wearing Necktie	Young		**Average**
uCelebA	LNets+ANet (Liu et al. 2015)	92.00	95.00	81.00	95.00	66.00	91.00	72.00	89.00	90.00	96.00	92.00	73.00	80.00	82.00	99.00	93.00	71.00	93.00	87.00		87.30
	Part-only	93.55	96.63	86.96	95.71	73.03	96.86	76.40	92.87	94.77	97.63	91.98	82.53	81.29	89.07	98.75	92.96	87.13	96.69	86.51		90.46
	Whole-only	93.24	96.59	87.19	95.40	74.48	96.85	76.06	92.95	94.83	97.50	91.61	82.18	82.63	89.13	98.50	93.58	87.14	96.77	87.14		90.60
	PaW	**94.05**	**96.90**	**87.56**	**96.22**	**75.03**	**97.08**	**77.35**	**93.44**	**95.07**	**97.64**	**92.73**	**83.52**	**84.07**	**89.93**	**99.02**	**94.24**	**87.70**	**96.85**	**88.59**		**91.23**

Blurry, *Heavy Makeup*, and *Pale Skin*, the whole-image-based subnet achieves the highest weight. Intuitively those global attributes should obtain more information from the whole-image-based subnet. This validates the region selection ability of the RSL.

Face Attribute Correlation The learned ARL weights are visualized in Figure 6. We find that attribute pairs that are mutually exclusive such as (*Attractive, Blurry*), (*Black Hair, Blond Hair*) and (*No Beard, Goatee*) are assigned lowest weights. Rarely co-occurring attribute pairs like (*Male, Heavy Makeup*) are also assigned low weights. Pairs of attributes such as (*Chubby, Double Chin*), (*Heavy Makeup, Wearing Lipstick*) and (*Smiling, High Cheekbones*) that commonly co-occur are given relatively higher weights. Moreover, the weights are asymmetric, for example, a person who wears lipstick is very unlikely to have a beard, but not the other way round. This is also reflected in the learned weights. This shows that ARL captures the attribute relationships.

Classification Results We show that our model achieves state-of-the-art results on uCelebA dataset. In the following experiments, each subnet adopts the architecture of SNet3 in Table 3.

We compare PaW with two baselines:

1. Part-only: each part net is trained on the detected face region to predict all face attributes. Then the attribute score from the most related part-based subnet is adopted for testing.

2. Whole-only: this method does not have part nets. It is trained with the whole face image only and is used to directly predict all attributes.

Table 5 summarizes the classification performances. We observe that the PaW net performs consistently better than either the Part-only or Whole-only method alone, achieving an accuracy of 91.23% vs. 90.60% for Part-only and 90.46% for Whole-only on uCelebA. This shows that RSL learns to selectively combine information from part-based and whole-image-based subnets. For unaligned face attribute classification on uCelebA dataset, we achieve the highest recognition rates across the board on all attributes and decrease the average recognition error from 12.70% to 8.77%, a reduction of 30.9%. Our method on the aligned CelebA also achieves an accuracy of 91.33% vs. 90.94% compared with the state-of-the-art (Rudd, Günther, and Boult 2016). This validates the effectiveness of the proposed attribute classification network. Also, the small performance gap on uCelebA and the aligned CelebA means that we practically eliminate the alignment step, and hence no special annotations are needed. Although the PaW network contains multiple part-based and whole-image-based subnets, the total number of parameters is only 11 M.

To test the importance of the FRL network, we further employ a baseline that divides each image into 4×4 non-overlapping blocks to simulate crude part detectors. Then part-based subnets and whole-image-based subnet are trained the same way as before. It achieves an average accuracy of 90.95% on uCelebA. However, we found that the weights corresponding to the whole-image-based net in the RSL are always higher than those of the part-based subnets for predicting *all* the attributes. This is because coarse region localization makes the part-based subnets unreliable, thus all the predictions are essentially made by the whole-image-based subnet only. This validates the effectiveness of the proposed FRL network.

Conclusions

In this paper, we propose to learn attentional face regions to improve attribute classification performance under unaligned condition. To this end, a weakly-supervised face region localization network is first designed. Then the information from those detected regions are selectively combined by the hybrid classification network. Visualization shows our method not only discovers semantic meaningful attributes regions, but also captures rich correlations among attributes. Moreover, our results outperform state-of-the-art by a significant margin on the unaligned CelebA dataset.

Acknowledgments

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should

not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

Abu-Mostafa, Y. S. 1992. A method for learning from hints. In *NIPS*.

Berg, T., and Belhumeur, P. 2013. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*.

Bucilu, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *ACM SIGKDD*.

Caruana, R. 1998. Multitask learning. In *Learning to learn*.

Cinbis, R. G.; Verbeek, J.; and Schmid, C. 2017. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE TPAMI*.

Denton, E. L.; Zaremba, W.; Bruna, J.; LeCun, Y.; and Fergus, R. 2014. Exploiting linear structure within convolutional networks for efficient evaluation. In *NIPS*.

Ding, H.; Zhou, S. K.; and Chellappa, R. 2017. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. *FG*.

Ehrlich, M.; Shields, T. J.; Almaev, T.; and Amer, M. R. 2016. Facial attributes classification using multi-task representation learning. In *CVPR Workshops*.

Escorcia, V.; Carlos Niebles, J.; and Ghanem, B. 2015. On the relationship between visual attributes and convolutional networks. In *CVPR*.

Hand, E. M., and Chellappa, R. 2017. Attributes for improved attributes: A multi-task network for attribute classification. In *AAAI*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, C.; Li, Y.; Change Loy, C.; and Tang, X. 2016. Learning deep representation for imbalanced classification. In *CVPR*.

Jadhav, A.; Namboodiri, V. P.; and Venkatesh, K. 2016. Deep attributes for one-shot face recognition. In *ECCV Workshops*.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM international conference on Multimedia*.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and simile classifiers for face verification. In *ICCV*.

Kumar, N.; Berg, A.; Belhumeur, P. N.; and Nayar, S. 2011. Describable visual attributes for face verification and image search. *IEEE TPAMI*.

LeCun, Y.; Denker, J. S.; Solla, S. A.; Howard, R. E.; and Jackel, L. D. 1989. Optimal brain damage. In *NIPS*.

Li, Y.; Wang, R.; Liu, H.; Jiang, H.; Shan, S.; and Chen, X. 2015. Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In *ICCV*.

Lin, M.; Chen, Q.; and Yan, S. 2014. Network in network. *ICLR*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*.

Lu, Y.; Kumar, A.; Zhai, S.; Cheng, Y.; Javidi, T.; and Feris, R. 2017. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. *CVPR*.

Moran, J., and Desimone, R. 1985. Selective attention gates visual processing in the extrastriate cortex. *Front. Cogn. Neurosci*.

Murrugarra-Llerena, N., and Kovashka, A. 2017. Learning attributes from human gaze. In *WACV*.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*.

Parkhi, O. M.; Vedaldi, A.; and Zisserman, A. 2015. Deep face recognition. In *BMVC*.

Posner, M. I., and Petersen, S. E. 1990. The attention system of the human brain. *Annual review of neuroscience*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rudd, E. M.; Günther, M.; and Boult, T. E. 2016. Moon: A mixed objective optimization network for the recognition of facial attributes. In *ECCV*.

Siddiquie, B.; Feris, R. S.; and Davis, L. S. 2011. Image ranking and retrieval based on multi-attribute queries. In *CVPR*.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR*.

Torfason, R.; Agustsson, E.; Rothe, R.; and Timofte, R. 2016. From face images and attributes to attributes. In *ACCV*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, J.; Cheng, Y.; and Schmidt Feris, R. 2016. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *CVPR*.

Yan, X.; Yang, J.; Sohn, K.; and Lee, H. 2016. Attribute2image: Conditional image generation from visual attributes. In *ECCV*.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *ECCV*.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE SPL*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*.