Learning Spatial-Semantic Relationship for Facial Attribute Recognition with Limited Labeled Data

Ying Shu[1], Yan Yan[1]*, Si Chen[2], Jing-Hao Xue[3], Chunhua Shen[4], Hanzi Wang[1]

[1] Xiamen University, China [2] Xiamen University of Technology, China
[3] University College London, UK [4] The University of Adelaide, Australia

Abstract

Recent advances in deep learning have demonstrated excellent results for Facial Attribute Recognition (FAR), typically trained with large-scale labeled data. However, in many real-world FAR applications, only limited labeled data are available, leading to remarkable deterioration in performance for most existing deep learning-based FAR methods. To address this problem, here we propose a method termed Spatial-Semantic Patch Learning (SSPL). The training of SSPL involves two stages. First, three auxiliary tasks, consisting of a Patch Rotation Task (PRT), a Patch Segmentation Task (PST), and a Patch Classification Task (PCT), are jointly developed to learn the spatial-semantic relationship from large-scale unlabeled facial data. We thus obtain a powerful pre-trained model. In particular, PRT exploits the spatial information of facial images in a self-supervised learning manner. PST and PCT respectively capture the pixel-level and image-level semantic information of facial images based on a facial parsing model. Second, the spatial-semantic knowledge learned from auxiliary tasks is transferred to the FAR task. By doing so, it enables that only a limited number of labeled data are required to fine-tune the pre-trained model. We achieve superior performance compared with state-of-the-art methods, as substantiated by extensive experiments and studies.

1. Introduction

Facial attribute recognition (FAR), which aims to predict multiple attributes (such as gender, age, and race) of a given facial image, can greatly facilitate a variety of applications, including face verification and identification [19, 29, 32], image generation [9, 36], and image retrieval [4, 33]. However, FAR is challenging due to significant facial appearance variations caused by pose, illumination, occlusion, *etc.*

State-of-the-art deep learning-based FAR methods usually rely heavily on a large number of labeled training data for achieving good classification accuracy. Unfortunately, in many real-world FAR applications, often only a small number of training data are labeled since labeling a massive amount of multi-attribute images can be very time-consuming and costly. As a result, the performance of these deep learning-based FAR methods significantly decreases in real-world applications. Here, we focus on the challenging problem of FAR with limited labeled data.

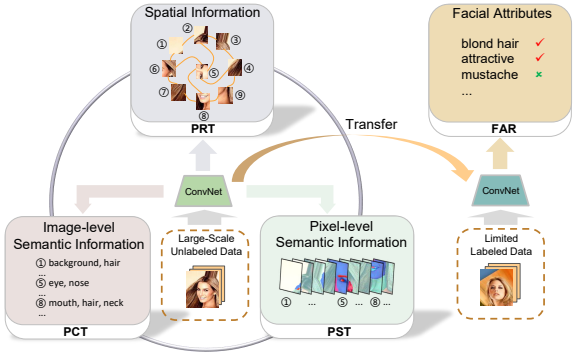


Figure 1 – Illustration of our proposed SSPL method. First, three auxiliary tasks (namely, PRT, PST, and PCT) are jointly learned to model the spatial-semantic relationship of facial images from large-scale unlabeled data, and a pre-trained model is obtained. Then, the pre-trained model is transferred to perform FAR with limited labeled data.

To alleviate the challenge of learning with limited labeled data, considerable efforts [6, 7, 8, 25, 34] have been spent on extracting high-level feature representations from unlabeled data in an unsupervised manner. Among these efforts, self-supervised learning has emerged as a prominent learning paradigm. The training of self-supervised learning involves two tasks: a pretext task and a downstream task [16]. Apart from self-supervised learning methods, some semi-supervised learning methods [1, 13, 18, 23, 26] have also been proposed, where labeled and unlabeled data are simultaneously used for training.

*Corresponding author (email: yanyan@xmu.edu.cn).

The tasks targeted by self-supervised learning and semi-supervised learning methods are usually image classification [15, 30], object detection [11, 12], and semantic segmentation [5, 38]. Different from these tasks, FAR is a multi-attribute classification task, where the spatial-semantic relationship of facial images is critical to classify attributes. For example, to identify the "BigNose" and "PointyNose" attributes, it is natural to locate the nose region and determine whether the nose is big and pointy at a semantic level. Similarly, the "Smiling" and "MouthOpen" attributes are predicted by exploiting the semantic information in the mouth region. Therefore, for FAR, it is pivotal to learn *fine-grained* feature representations, in particular capturing the *spatial-semantic relationship*, from *unlabeled facial data*.

In this work, we propose a novel Spatial-Semantic Patch Learning method (SSPL) to address the problem of effectively learning the spatial-semantic relationship for achieving state-of-the-art FAR with limited labeled data. To this end, as shown in Figure 1, the training of SSPL involves two stages. First, three auxiliary tasks consisting of a Patch Rotation Task (PRT), a Patch Segmentation Task (PST), and a Patch Classification Task (PCT) are jointly proposed and trained to obtain a powerful pre-trained model. Second, the pre-trained model is transferred to perform FAR by fine-tuning on limited labeled data.

Specifically, given several facial patches (one of which is rotated), PRT predicts the index of the rotated patch to exploit the spatial information of facial images. Meanwhile, PST performs semantic segmentation to assign a semantic label to each pixel in a randomly selected facial patch and PCT predicts facial component labels of this patch, such that PST and PCT can respectively encode the pixel-level and image-level semantic information of facial images. These three tasks and their joint training effectively capture the spatial-semantic relationship between facial regions, which in turn leads to a significant improvement of FAR when only limited labeled data are available.

Our main contributions are summarized as follows.

- We propose the SSPL method to address the problem of FAR with limited labeled data. SSPL effectively exploits both the spatial and semantic information from unlabeled facial data to obtain a powerful pre-trained model, ensuring that an attribute recognition model can be easily fine-tuned to accurately predict facial attributes by using only limited labeled data.
- We elaborately design three auxiliary tasks (*i.e.*, PRT, PST, and PCT) targeted for FAR. These auxiliary tasks are jointly trained to make use of the intrinsic relationship between patch rotation prediction and patch segmentation/classification. This enables the network to effectively extract semantic-aware fine-grained feature representations.
- Our experiments convincingly show that the proposed method performs favorably against state-of-the-art methods in the case of limited labeled data, demonstrating the potentials of learning the spatial-semantic relationship of facial images for FAR.

2. Related Work

Here we review the closely related deep learning-based work in FAR and learning from unlabeled data.

Facial Attribute Recognition. With the increasing availability of large-scale data, deep learning-based methods have become dominant in the field of FAR. Sharma and Foroosh [27] leverage deep separable convolutions and pointwise convolution to design a lightweight CNN for FAR, which significantly reduces the model parameters and improves the computational efficiency. Mao *et al*. [22] perform FAR based on a Deep Multi-task and Multi-label Convolutional Neural Network (DMM-CNN). He *et al*. [14] propose to use synthesized abstraction images to improve the FAR performance.

The above FAR methods are often trained on large-scale labeled data. However, in many real-world FAR applications, sufficient labels can be difficult to collect. As a result, the performance of these methods greatly degrades. In this work, we address the challenging and less studied problem of FAR with limited labeled data.

Learning from Unlabeled Data. A large number of methods have been proposed to learn features from unlabeled data, which can significantly reduce the high cost of annotating large-scale data.

Self-supervised learning Recently, self-supervised learning methods with deep neural networks have received considerable attention. For example, Caron *et al*. [3] use an image clustering algorithm to generate labels for image classification. In [24], the images are divided into 9 patches and shuffled, and then a pretext task is designed to solve the jigsaw puzzle to identify the correct spatial locations of input patches. Gidaris *et al*. [10] propose to learn to predict the geometric transformation of images.

Semi-supervised learning Current semi-supervised learning methods with deep neural networks roughly contain two categories: 1) consistency regularization-based methods [23, 35]; and 2) proxy label-based methods [28].

The consistency regularization-based methods introduce a regularization term to the objective function to enable the training of unlabeled data. Xie *et al*. [35] develop an Unsupervised Domain Adaptation (UDA) method to make use of realistic noise generated by data augmentation methods. The proxy label-based methods first assign proxy labels to unlabeled data (pseudo-labels), and then train unlabeled and labeled data based on proxy and ground-truth labels. Sohn *et al*. [28] introduce a FixMatch method which com-

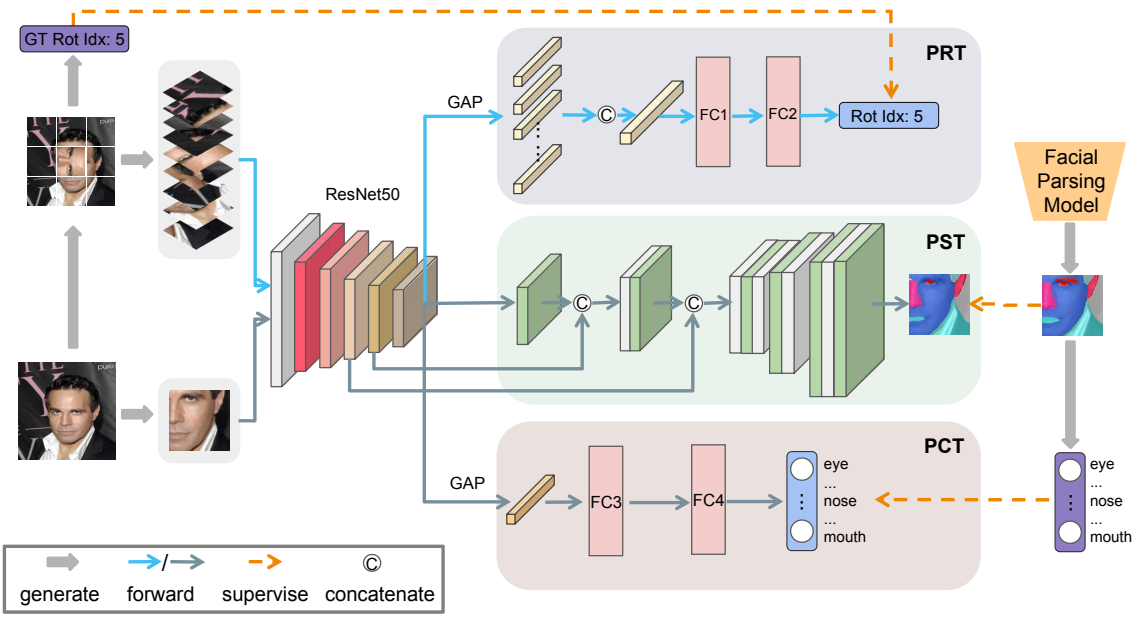


Figure 2 – Overview of three auxiliary tasks in SSPL, including a Patch Rotation Task (PRT), a Patch Segmentation Task (PST), and a Patch Classification Task (PCT). PRT exploits the spatial information of facial images in a self-supervised learning manner. PST and PCT respectively capture the pixel-level and image-level semantic information of facial images by using a facial parsing model.

bines consistency regularization and proxy-labeling to perform semi-supervised learning.

The above methods usually learn holistic feature representations. They may not be suitable for FAR, where various facial attributes correspond to different facial regions. In this paper, we model the spatial-semantic relationship between facial regions by leveraging patch rotation prediction and patch segmentation/classification. Thus, fine-grained feature representations can be extracted, which are important for FAR.

3. Our Method

In this section, we first give an overview of the proposed method and then discuss the details of key components.

3.1. Overview

SSPL includes three auxiliary tasks and a target FAR task. The training of SSPL involves two stages. First, three auxiliary tasks (consisting of a Patch Rotation Task (PRT), a Patch Segmentation Task (PST), and a Patch Classification Task (PCT)) are jointly trained to learn semantic-aware fine-grained feature representations based on large-scale unlabeled facial data, and thus a powerful pre-trained model (based on ResNet-50) is obtained. Second, the target FAR task fine-tunes the pre-trained model with limited labeled data and then predicts facial attributes.

An overview of the three auxiliary tasks in SSPL is shown in Figure 2. More specifically, PRT encodes the spatial information of facial images based on self-supervised learning. In particular, the input facial image is divided into $m \times m$ patches and one of them is randomly chosen and rotated. PRT is trained to predict the index of the rotated patch, where the index label corresponds to a number in $\{1, 2, ..., m^2\}$. PST and PCT respectively exploit the pixel-level and image-level semantic information of facial images based on a facial parsing model. On one hand, PST performs semantic segmentation on a randomly selected facial patch and assigns a semantic label to each pixel in the patch. We leverage an externally-trained facial parsing model (BiSeNet [37]) to generate proxy semantic labels. On the other hand, PCT predicts facial component labels of the same patch as that in PST, where the proxy component labels are obtained by aggregating semantic labels from BiSeNet.

3.2. Patch Rotation Task (PRT)

In this subsection, we develop a pretext task PRT to fully exploit the spatial relationship between different patches. The network architecture of PRT consists of a ResNet-50 backbone, a Global Average Pooling (GAP) layer, and two Fully-Connected (FC) layers.

Given an input facial image $\mathbf{I}$, it is first divided into $m \times m$ different patches $\{\mathbf{p}_1, \cdots, \mathbf{p}_{m^2}\}$. Then, one patch $\mathbf{p}_r$ is randomly selected and rotated. PRT takes these patches as the input and predicts the index of the rotated patch.

To be specific, each patch is sequentially fed into the ResNet-50 backbone to extract the patch feature map $\mathbf{F}_i \in \mathbb{R}^{c\times w\times h}$, where c, w, and h represent the channel, width, and height of the feature map, respectively. Then, all the patch feature maps are fed into a GAP layer and concatenated to obtain a whole feature $\mathbf{f}_p$. Next, $\mathbf{f}_p$ is flattened and fed into two FC layers and a softmax layer to give the probabilities of the index predictions $\boldsymbol{t} = [t_1, \cdots, t_{m^2}]$, where $t_i \in [0, 1]$. Note that we concatenate the patch features extracted from the backbone rather than stacking the original patches, to avoid a network simply identifying correlations between low-level texture statistics. Therefore, the network is able to learn high-level primitives and structures to predict the correct index.

In order to prevent the network from taking shortcuts (*e.g.*, edge continuity, pixel intensity distribution, and chromatic aberration) when predicting the index of a rotation patch, similarly to [24], we perform color jitter for each patch and then normalize each patch independently. As a result, the network can effectively capture the spatial information between a patch and its surrounding patches.

The loss of PRT uses the cross-entropy:

$$\mathcal{L}_R = \sum_{i=1}^{m^2} -q_i \log(t_i), \tag{1}$$

where $q_i = 1$ if $i = r$ and $q_i = 0$ otherwise.

It is worth noting that Gidaris *et al*. [10] propose a self-supervised learning method, which predicts the rotated angle of an input image. However, such a method cannot fully exploit the geometric structure of facial images for FAR, since different facial attributes are often associated with different facial regions. The proposed PRT divides the facial image into several patches and learns the spatial relationship between them. Therefore, the proposed pretext task may work better for the FAR task.

3.3. Patch Segmentation Task (PST)

In this subsection, we propose PST to predict semantic labels of pixels in a patch. Considering that PST and PRT share the same backbone, instead of using the whole facial image, we employ a randomly cropped facial patch as the input of PST. Such an approach can avoid shortcuts for PRT, since shortcuts exploit the relevant information (such as low-level statistics in facial images) helpful for solving the pretext task (PRT), but not for the target task.

Specifically, a facial patch $\mathbf{p}_s$ is randomly cropped from the original facial image $\mathbf{I}$ and taken as the input of PST. The patch is fed into the ResNet-50 backbone to extract different levels of features. To take full advantage of multi-level information, the output feature map from the 4th residual block of ResNet-50 is fed into an upsampling layer and then concatenated with the output feature map from the 3rd residual block of ResNet-50. Then, the concatenated feature map is fed into a convolutional layer, followed by batch normalization, ReLU, and an upsampling layer. Next, the output feature map from the 2nd residual block of ResNet-50 and that from the previous layer are concatenated, and fed into another convolutional layer, followed by batch normalization and ReLU. Finally, another three pairs of upsampling layers and convolutional layers with batch normalization and ReLU are used to classify each pixel of the final feature map into different semantic classes.

Given J semantic classes and the class prediction probabilities for the k-th pixel as $\boldsymbol{h}_s = [h_{k1}, \cdots, h_{kJ}]$, we can formulate the loss of the k-th pixel in $\mathbf{p}_s$ as

$$\mathcal{L}_{pixel} = \sum_{j=1}^{J} -q_{kj} \log(h_{kj}), \tag{2}$$

where $q_{kj} = 1$ if j is the ground-truth label of the k-th pixel and $q_{kj} = 0$ otherwise.

Generally, the semantic labels of facial images are not available in the face attribute datasets. In this paper, we use an externally trained facial parsing model (*i.e.*, BiSeNet [37]) to directly predict the semantic labels of the input patch $\mathbf{p}_s$ as the proxy semantic labels for PST. Note that the original predicted output of BiSeNet contains 19 categories, and we merge them into 8 categories (*i.e.*, background, skin, eye, ear, nose, mouth, neck, and hair) according to their spatial positions. To alleviate the overfitting of incorrect proxy semantic labels, we further make use of the label smoothing strategy [31], which is defined as

$$q'_{kj} = (1 - \epsilon)q_{kj} + \frac{\epsilon}{J}, \tag{3}$$

where q'_{kj} is the smoothed ground-truth label and ϵ is a smoothing parameter empirically set to 0.1 as in [31].

With Eq. (2) and Eq. (3), the loss of PST is defined as

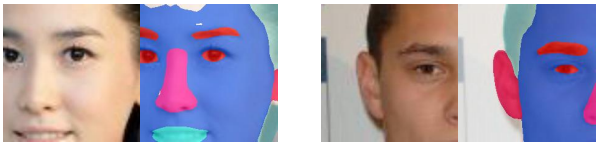
$$\mathcal{L}_S = \frac{1}{K} \sum_{k=1}^{K} \left(\sum_{j=1}^{J} -q'_{kj} \log(h_{kj}) \right), \tag{4}$$

where K is the total number of pixels in $\mathbf{p}_s$.

3.4. Patch Classification Task (PCT)

PST exploits the pixel-level semantic information of facial images. Nevertheless, the target FAR task is an image-level classification task. Therefore, we further develop PCT to predict the facial components of a given input patch, which can explicitly capture the image-level semantic information of facial images.

In this paper, PCT adopts the same input as PST instead of the whole facial image. Note that, if the whole facial images are used as input, PCT will be trained with similar proxy component labels (*i.e.*, most facial components



(a) (b)

Figure 3 – Examples of two input facial patches and the corresponding semantic masks from CelebA. The facial components "ear" and "nose" exist in (a) and (b), respectively. But they are not the dominant facial components.

exist), which may not be helpful to train PCT due to the significantly imbalanced distribution of proxy component labels.

The architecture of PCT is similar to that of PRT, except that the output channels of the last two FC layers are different due to the different numbers of predicted classes. Specifically, the facial patch $\mathbf{p}_s$ is first fed into the ResNet-50 backbone to obtain a feature map $\mathbf{F}_s \in \mathbb{R}^{c' \times w' \times h'}$, where c', w', and h' represent the channel, width, and height of the feature map, respectively. Then, $\mathbf{F}_s$ is used as the input of a GAP layer to obtain a feature map $\mathbf{f}_s$. Finally, $\mathbf{f}_s$ is fed into two FC layers to predict facial component labels.

In this paper, the facial components are the same as the semantic classes used in PST. However, PST is a pixel-level classification task (*i.e.*, predicting the semantic labels of pixels), while PCT is an image-level classification task (*i.e.*, predicting the existence of facial components in a patch). The proxy component labels of the input patch are generated by aggregating the pixel-level semantic labels given by BiSeNet. In particular, the proxy component labels of the input patch are denoted as a vector, that is, $\mathbf{y}_s = [y_0, \cdots, y_J]$. Here, $y_i = 1$ denotes the existence of a facial component, and $y_i = 0$ otherwise.

It is worth mentioning that the proxy component label is obtained by aggregating semantic labels, and thus it is tolerant of small semantic label errors. Usually, a few facial components exist in the input patch $\mathbf{p}_s$ and some of them only involve a relatively small number of pixels in the patch, as illustrated in Figure 3. Therefore, we only choose the top n dominant facial components in each input patch and label them as 1. For the rest of facial components, we label them as 0.

The loss of PCT adopts the binary cross-entropy:

$$\mathcal{L}_C = \sum_{j=1}^{J} \left(y_j \log(x_j) + (1 - y_j) \log(1 - x_j) \right), \qquad (5)$$

where x_j is the output prediction probability of the j-th facial component.

3.5. Joint Loss Function

Finally, the joint loss function of SSPL can be formulated as

$$\mathcal{L}_{SSPL} = \mathcal{L}_R + \lambda_1 \mathcal{L}_S + \lambda_2 \mathcal{L}_C, \qquad (6)$$

where λ_1 and λ_2 are the regularization parameters to balance different losses.

Using Eq. (6), the three auxiliary tasks can be jointly trained in an end-to-end manner. Note that all labels used for training the three auxiliary tasks are automatically generated to alleviate the burden of labeling large-scale facial data.

4. Experiments

In this section, we first briefly introduce two public facial attribute datasets. Then, we describe the implementation details. Next, we perform ablation studies to show the importance of each auxiliary task in SSPL and the influence of key parameters on the final performance. Finally, we compare SSPL with several state-of-the-art methods.

4.1. Datasets

CelebA [21] is a popular large-scale facial attribute dataset used to evaluate the FAR performance. It consists of 202,599 facial images with 40 attribute annotations per image. CelebA is divided into 3 parts, including 162,770 images for training, 19,867 images for validation, and 19,962 images for test.

LFWA [21] is another widely-used facial attribute dataset. It contains 13,143 facial images with the same attribute annotations as the CelebA dataset. In LFWA, 6,263 images, 2,800 images, and 4,080 images are used for training, validation, and test, respectively.

Here, we use the default training set (without using labels) provided by CelebA or LFWA to train three auxiliary tasks. Moreover, we randomly choose a proportion of the training set (with labels), and use the default validation and test sets of CelebA or LFWA in the FAR task. All experiments are performed 10 times and the average recognition accuracy is reported.

4.2. Implementation Details

In PRT, the number of patches per side m is set to 3. Thus, there are in total $3 \times 3 = 9$ patches. The input image $\mathbf{I}$ is first resized to 255×255, and then 9 patches with the size of 85×85 are cropped from the input image. Finally, a patch with the size of 64×64 is randomly cropped from each 85×85 patch and resized to 224×224. Thus, we prevent the model from using low-level texture statistics, which are not beneficial for the downstream task [24]. Given a generated index, the corresponding patch is rotated by 90 degrees. In PST and PRT, a patch with the size of 75×75 is randomly cropped from the original image, and then resized to the

Table 1 – Ablation studies: The recognition accuracy (%) obtained by six variants of SSPL with the different proportions of labeled training data on the CelebA and LFWA datasets.

	CelebA					LFWA				
Proportion	0.2%	0.5%	1%	2%	100%	5%	10%	20%	50%	100%
# of training samples	325	843	1,627	3,255	162,770	313	6,26	1,252	3,131	6,263
Baseline	82.57	85.33	87.14	88.24	91.73	73.90	75.50	78.78	83.76	84.11
PRT	85.36	86.97	87.82	88.93	91.72	77.31	79.96	82.55	84.87	85.72
PST+PCT	85.39	86.07	87.30	88.27	91.72	76.82	78.88	82.04	84.01	85.65
PRT+PCT	85.55	86.97	88.06	89.01	91.75	77.93	80.52	82.64	84.92	86.10
PRT+PST	85.96	87.58	88.31	89.05	91.76	78.08	81.17	82.90	84.89	86.15
SSPL	**86.67**	**88.05**	**88.84**	**89.58**	**91.77**	**78.68**	**81.65**	**83.45**	**85.43**	**86.53**

size of 224×224. In PRT, the number of dominant facial components n is set to 4. We use ResNet-50 (without pre-training) as the backbone.

For the FAR task, we simply replace the last two FC layers of ResNet-50 backbone (trained in three auxiliary tasks) with one FC layer (with 40 output nodes) and fine-tune the whole network. We use PyTorch to implement SSPL and all the experiments are performed on a GTX 2080 GPU. For the three auxiliary tasks, the batch size is set to 40 and the model is trained for 80 epochs. The values of λ_1 and λ_2 in Eq. (6) are empirically set to 0.05 and 0.50, respectively. For the target task, the batch size is set to 64, and the model is trained for 60 epochs. The Adam optimizer [17] is adopted with the initial learning rate of 1×10^{-4}, $\beta_1 = 0.500, \beta_2 = 0.999$ and weight decay of 5×10^{-4}. The warm-up strategy is used to update the learning rate during training, where the value of the learning rate is linearly increased from 1×10^{-3} to 3.5×10^{-3} in the first 15 epochs, and then remains at 1.5×10^{-5} until the end of training. The Adam optimizer and the warm-up strategy are used in both pre-training and fine-tuning.

4.3. Ablation studies

We conduct ablation studies to evaluate the influence of different auxiliary tasks (*i.e.*, PRT, PST, and PCT) and critical parameters (including the number of patches and the number of dominant facial components) of the proposed method on the final performance.

Influence of different auxiliary tasks. We evaluate six variants of the proposed method, including: (1) the method (denoted as "PRT") that only adopts PRT as the auxiliary task; (2) the method (denoted as "PST+PCT") that uses PST and PCT as the auxiliary tasks; (3) the method (denoted as "PRT+PCT") that uses PRT and PCT as the auxiliary tasks; (4) the method (denoted as "PRT+PST") that uses PRT and PST as the auxiliary tasks; (5) the proposed SSPL method, which effectively combines PRT, PST, and PCT in an integrated network; and (6) the baseline method that is based on the ResNet-50 backbone. All the variants are trained end-to-end from scratch. The results obtained by six variants with the different proportions of labeled training data are given in Table 1.

From Table 1, SSPL achieves higher accuracy than the baseline method on both CelebA and LFWA. When a smaller proportion of labeled training data is used, the improvements obtained by SSPL are more evident (*e.g.*, SSPL outperforms the baseline by 4.10% on CelebA and by 4.78% on LFWA when 0.2% and 5% of labeled training data are respectively used).

Note that when 100% of labeled training data is used, SSPL and the baseline method achieve similar results in CelebA. This is because sufficient training data are used to obtain the optimized network parameters for the baseline method. Compared with PST+PCT, SSPL also improves the accuracy, since it additionally adopts PRT. This shows the effectiveness of the pretext task PRT, which exploits the spatial information of facial images based on self-supervised learning to improve the FAR performance in the case of limited labeled data.

PST takes advantage of semantic segmentation to extract fine-grained semantic information from images. As shown in Table 1, PRT+PST achieves higher accuracy (*e.g.*, 0.60% improvements on CelebA and 0.77% improvements on LFWA when 0.2% and 5% of labeled training data are respectively used) than PRT. Meanwhile, compared with PRT+PCT, SSPL achieves higher accuracy (*e.g.*, 1.12% improvements on 0.2% of CelebA and 0.75% improvements on 5% of LFWA). This can be ascribed to the fact that joint training of PRT and PST effectively exploits the relationship between pixel-level semantic segmentation and patch rotation prediction, which improves the performance.

PCT learns the semantic relationship between different patches by identifying the facial components of a randomly chosen patch. Compared with PRT+PST, SSPL also improves the performance on CelebA and LFWA. Therefore, the image-level semantic information plays an important role to improve the performance of FAR with limited labeled data. By combing PRT, PST, and PCT, SSPL achieves the best performance among all the variants. Compared with the baseline, SSPL improves the accuracy from

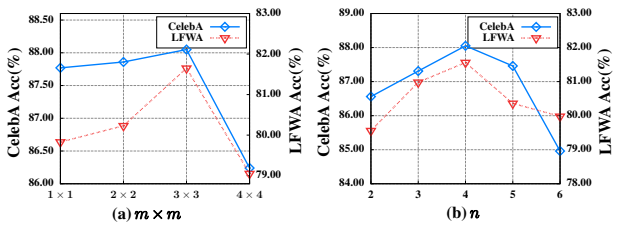


Figure 4 – Ablation studies: Influence of (a) the number of patches and (b) the number of dominant facial components on the final performance on CelebA and LFWA when 0.5% and 10% of the training labeled data of CelebA and LFWA are used, respectively.

82.57% to 86.67% on CelebA and from 73.90% to 78.68% on LFWA, when 0.2% and 5% of labeled training data are respectively used. This shows the importance of modeling the spatial-semantic relationship between facial regions, which can be beneficial for the FAR task.

Influence of the number of patches $m \times m$. We evaluate the performance of SSPL with the different numbers of patches $m \times m$ in PRT, including 1×1, 2×2, 3×3, and 4×4. The results are shown in Figure 4 (a). We can see that when the number of patches $m \times m$ is set to 3×3, our method achieves the best performance. On one hand, when the number of patches is larger, the semantically consistent facial image is over-segmented into many small patches. On the other hand, when the number of patches is smaller, the large patch contains many facial components. Too large or too small values of the number of patches adversely affect the extraction of features encoding the spatial information.

Influence of the number of dominant facial components n. We evaluate the influence of the number of dominant facial components in PCT on the final performance. The experimental results are given in Figure 4 (b). Our proposed method obtains the best results when the value of n is set to 4. When the values of n are too large, some facial components that involve only a few pixels are chosen as dominant facial components and when n is too small, some dominant facial components are ignored. Both cases will lead to performance degradation.

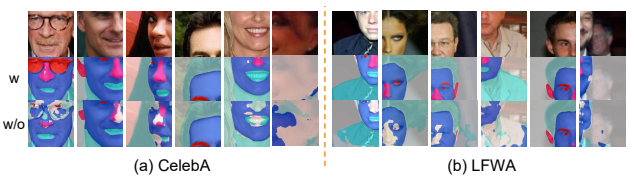


Figure 5 – Semantic masks generated by SSPL with (denoted as "w") and without the label smoothing strategy (denoted as "w/o") on (a) CelebA and (b) LFWA.

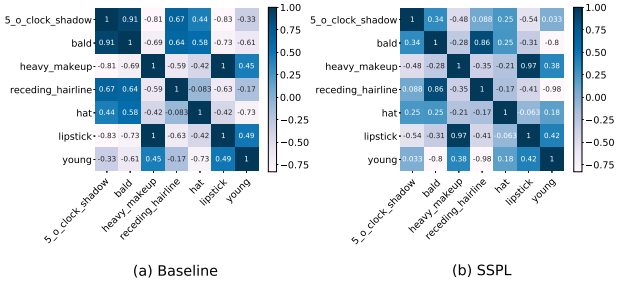


Figure 6 – The correlation maps of seven randomly selected facial attributes obtained by (a) the baseline and (b) SSPL on CelebA.

4.4. Visualization

In this section, we visualize several examples of semantic masks generated by SSPL with and without the label smoothing strategy, as shown in Figure 5. Moreover, we plot the correlation maps of several randomly selected facial attributes obtained by the baseline and SSPL, as illustrated in Figure 6. To be specific, we randomly select seven facial attributes, and use the predicted outputs of the trained model to calculate the correlation map. Here, we employ 0.2% of labeled training data in CelebA.

From Figure 5, the semantic masks generated by SSPL contain less noise than the SSPL without the label smoothing strategy. This demonstrates the effectiveness of the label smoothing strategy. SSPL can generate accurate semantic masks by jointly training three auxiliary tasks, which can be beneficial to capture the pixel-level semantic information of the input facial images. Note that there are some false-detected masks. For example, in the last columns of CelebA and LFWA in Figure 5, most pixels in the facial patch are classified as the background, due to blurring. However, as these blurry patches do not greatly contribute to the learning of semantic information, the false-detected masks will have no much influence on the final performance.

From Figure 6, compared with baseline, SSPL shows better correlation responses between facial attributes. For example, the "5_o_clock_shadow" attribute is negatively correlated with the "bald" attribute (the correlation value is 0.34 obtained by SSPL and that is 0.91 by baseline), while the "receding_hairline" and "bald" attributes are strongly related to each other (the correlation value is 0.86 obtained by SSPL and that is 0.64 by baseline).

4.5. Comparison with State-of-the-Art Methods

In this section, we compare the proposed SSPL method with ten state-of-the-art methods, including five supervised FAR methods [22, 27, 20, 2, 14], three self-supervised learning methods [3, 24, 10], and two semi-supervised learning methods [28, 23], on the CelebA and LFWA

Table 2 – Recognition accuracy (%) obtained by our proposed SSPL method and ten state-of-the-art methods with the different proportions of labeled training data on the CelebA and LFWA datasets.

	CelebA					LFWA				
Proportion	0.2%	0.5%	1%	2%	100%	5%	10%	20%	50%	100%
# of training samples	325	843	1,627	3,255	162,770	313	626	1,252	3,131	6,263
DMM [22]	-	-	-	-	91.70	-	-	-	-	86.56
SlimCNN [27]	79.90	80.20	80.96	82.32	91.24	70.90	71.49	72.12	73.45	76.02
AFFAIR [20]	-	-	-	-	91.45	-	-	-	-	86.13
PS-MCNN [2]	-	-	-	-	**92.98**	-	-	-	-	**87.36**
He *et al*. [14]	-	-	-	-	91.81	-	-	-	-	85.20
DeepCluster [3]	83.21	86.13	87.46	88.86	91.68	74.21	77.42	80.77	84.27	85.90
JigsawPuzzle [24]	82.88	84.71	86.25	87.77	91.57	73.90	77.01	79.56	83.29	84.86
Rot [10]	83.25	86.51	87.67	88.82	91.69	74.40	76.67	81.52	84.90	85.72
FixMatch [28]	80.22	84.19	85.77	86.14	89.78	71.42	72.78	75.10	80.87	83.84
VAT [23]	81.44	84.02	86.30	87.28	91.44	72.19	74.42	76.26	80.55	84.68
SSPL (Ours)	**86.67**	**88.05**	**88.84**	**89.58**	91.77	**78.68**	**81.65**	**83.45**	**85.43**	86.53

datasets, respectively. For five supervised FAR methods, we only use labeled training data in the FAR task to train the models. For self-supervised learning methods, we adopt all the unlabeled training data in the pretext task to obtain the initial network parameters, and then use the different proportions of training data in the downstream FAR task for fine-tuning. For semi-supervised learning methods, we train the models using both unlabeled and labeled training data. The experimental results are given in Table 2.

We can see that our SSPL method achieves similar or better performance than state-of-the-art FAR methods (such as DMM, Slim-CNN, PS-MCNN, and AFFAIR) on both datasets when 100% labeled data are used to train the models. These FAR methods can extract discriminative features from large-scale data. Note that DMM adopts a dynamic weighting scheme and an adaptive thresholding strategy to train the model. PS-MCNN designs a complicated network architecture consisting of four task specific networks (TSNets) and a shared network (SNet) to learn features for each group of attributes and different groups of attributes, respectively. In contrast, SSPL only uses the simple ResNet-50 model. This demonstrates the effectiveness of the pre-trained model in the auxiliary tasks. Moreover, the proposed method significantly outperforms Slim-CNN by a large margin when only a small proportion of training data (such as 0.2%, 0.5%, 1%, or 2%) are used. This is because we jointly train the auxiliary tasks to exploit the spatial-semantic information of facial images, so that effective semantic-aware features are extracted in the FAR task.

The SSPL method obtains much better performance than the competing self-supervised methods under the small proportions of labeled training data. In particular, when less training data are used, the performance improvements obtained by our method are more evident. This indicates the excellent capability of our method to extract effective features with limited labeled data, due to the superiority of exploiting both spatial and semantic information from unlabeled facial data based on three auxiliary tasks.

Compared with semi-supervised learning methods, our SSPL method achieves substantially higher accuracy with limited labeled data. VAT exploits unlabeled data by minimizing the distance between an image and a transformed version of the image, while FixMatch simultaneously employs consistency regularization and proxy-labeling strategies. However, these methods are based on holistic features. Therefore, they cannot effectively model the spatial relationship between different facial patches, which is critical for FAR. In contrast, SSPL learns the spatial-semantic correlation of facial images and extracts fine-grained features, leading to performance improvements.

5. Conclusion

In this paper, we have presented a new Spatial-Semantic Patch Learning (SSPL) method to effectively perform FAR in the case when only limited labeled data are available. The SSPL method involves three auxiliary tasks and a target FAR task. The auxiliary tasks, including PRT, PST, and PCT are developed to fully exploit the spatial-semantic information of facial images from unlabeled facial data to build a powerful pre-trained model. The target FAR task fine-tunes the pre-trained model by using limited labeled data. Extensive experiments on the CelebA and LFWA datasets show the superiority of our proposed method against state-of-the-art methods to address FAR with only limited labeled data.

Acknowledgements

This work was in part supported by the National Natural Science Foundation of China under Grants 62071404 and 61872307, by the Natural Science Foundation of Fujian Province under Grant 2020J01001, and by the Youth Innovation Foundation of Xiamen City under Grant 3502Z20206046.

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. Advances in Neural Inf. Process. Syst.*, 2019.

[2] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proc. Eur. Conf. Comp. Vis.*, 2018.

[4] Bor-Chun Chen, Yan-Ying Chen, Yin-Hsi Kuo, and Winston H. Hsu. Scalable face image retrieval using attribute-enhanced sparse codewords. *IEEE Trans. Multimedia*, 2013.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[6] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[7] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015.

[8] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Proc. Advances in Neural Inf. Process. Syst.*, 2014.

[9] Wen Gao, Shiguang Shan, Xiujuan Chai, and Xiaowei Fu. Virtual face image generation for illumination and pose insensitive face recognition. In *Proc. IEEE Int. Conf. Multimedia Expo.*, 2003.

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Proc. Int. Conf. Learn. Representations*, 2018.

[11] Ross Girshick. Fast r-cnn. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014.

[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Proc. Advances in Neural Inf. Process. Syst.*, 2005.

[14] Keke He, Yanwei Fu, Wuhao Zhang, Chengjie Wang, Yu-Gang Jiang, Feiyue Huang, and Xiangyang Xue. Harnessing synthesized abstraction images to improve facial attribute recognition. In *Proc. Int. Joint Conf. Artificial Intell.*, 2018.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[16] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. Learn. Representations*, 2014.

[18] Durk P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Proc. Advances in Neural Inf. Process. Syst.*, 2014.

[19] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2009.

[20] Jianshu Li, Fang Zhao, Jiashi Feng, Sujoy Roy, Shuicheng Yan, and Terence Sim. Landmark free face attribute prediction. *IEEE Trans. Image Process.*, 2018.

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015.

[22] Longbiao Mao, Yan Yan, Jing-Hao Xue, and Hanzi Wang. Deep multi-task multi-label CNN for effective facial attribute classification. *IEEE Trans. Affective Comput.*, 2020.

[23] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[24] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proc. Eur. Conf. Comp. Vis.*, 2016.

[25] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[26] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015.

[27] Ankit Sharma and Hassan Foroosh. Slim-CNN: A light-weight CNN for face attribute prediction. *arXiv preprint arXiv:1907.02157*, 2019.

[28] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[29] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Proc. Advances in Neural Inf. Process. Syst.*, 2014.

[30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015.

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016.

[32] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014.

[33] Zhong Wu, Qifa Ke, Jian Sun, and Heung-Yeung Shum. Scalable face image retrieval with identity-based quantization and multireference reranking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011.

[34] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018.

[35] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Proc. Advances in Neural Inf. Process. Syst.*, 2020.

[36] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *Proc. Eur. Conf. Comp. Vis.*, 2016.

[37] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proc. Eur. Conf. Comp. Vis.*, 2018.

[38] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.