International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01787-w

SPL-Net: Spatial-Semantic Patch Learning Network for Facial Attribute Recognition with Limited Labeled Data

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

Received: 5 December 2021 / Accepted: 12 March 2023

Abstract
Existing deep learning-based facial attribute recognition (FAR) methods rely heavily on large-scale labeled training data. Unfortunately, in many real-world applications, only limited labeled data are available, resulting in the performance deterioration of these methods. To address this issue, we propose a novel spatial-semantic patch learning network (SPL-Net), consisting of a multi-branch shared subnetwork (MSS), three auxiliary task subnetworks (ATS), and an FAR subnetwork, for attribute classification with limited labeled data. Considering the diversity of facial attributes, MSS includes a task-shared branch and four region branches, each of which contains cascaded dual cross attention modules to extract region-specific features. SPL-Net involves a two-stage learning procedure. In the first stage, MSS and ATS are jointly trained to perform three auxiliary tasks (i.e., a patch rotation task (PRT), a patch segmentation task (PST), and a patch classification task (PCT)), which exploit the spatial-semantic relationship on large-scale unlabeled facial data from various perspectives. Specifically, PRT encodes the spatial information of facial images based on self-supervised learning. PST and PCT respectively capture the pixel-level and image-level semantic information of facial images by leveraging a facial parsing model. Thus, a well-pretrained MSS is obtained. In the second stage, based on the pre-trained MSS, an FAR model is easily fine-tuned to predict facial attributes by requiring only a small amount of labeled data. Experimental results on challenging facial attribute datasets (including CelebA, LFWA, and MAAD) show the superiority of SPL-Net over several state-of-the-art methods in the case of limited labeled data.

Keywords Facial attribute recognition · Limited labeled data · Multi-task learning · Multi-label learning · Self-supervised learning · Semantic segmentation

1 Introduction

Facial attributes (such as gender, age, and expression) describe important visual properties of facial images, and provide mid-level representations between low-level features and high-level labels (Cao et al., 2018a). Over the past few years, facial attribute recognition (FAR) has attracted considerable attention from both academia and industry. This is mainly because of its significant importance in various computer vision tasks, including face verification and recognition (He et al., 2018b; Chen et al., 2018; Song et al., 2018; Rao et al., 2019; Zhang et al., 2017b), image editing (Song et

Communicated by Maja Pantic.

✉ Yan Yan
yanyan@xmu.edu.cn

Ying Shu
shuyin9@stu.xmu.edu.cn

Si Chen
chensi@xmut.edu.cn

Jing-Hao Xue
jinghao.xue@ucl.ac.uk

Chunhua Shen
chunhua@me.com

Hanzi Wang
hanzi.wang@xmu.edu.cn

[1] Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China

[2] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China

[3] Department of Statistical Science, University College London, London WC1E 6BT, UK

[4] Zhejiang University, Hangzhou 310058, China

Published online: 11 May 2023

al., 2019; Egger et al., 2018; Huang et al., 2018), and image retrieval (Nguyen et al., 2018; Li et al., 2015).

With the rapid development of deep learning, a large number of FAR methods (Zhang et al., 2014; Liu et al., 2015; Kalayeh et al., 2017; Mahbub et al., 2018; He et al., 2018a; Li et al., 2018; Rudd et al., 2016; Hand & Chellappa, 2017; Cao et al., 2018a) have been proposed and shown promising performance. These methods often rely on abundant labeled data to learn discriminative feature representations for classifying attributes. However, in many real-world applications, only a small amount of labeled training data are provided since labeling massive multi-attribute images is time-consuming and labor-intensive. As a consequence, the performance of these methods may substantially drop in these applications. In this paper, we study the challenging problem of FAR with limited labeled data.

To address the challenge of learning with limited labeled data, many recent efforts (Caron et al., 2018; Noroozi & Favaro, 2016; Gidaris et al., 2018; He et al., 2020; Chen et al., 2020; Sohn et al., 2020; Miyato et al., 2018) have been devoted to extracting feature representations in a self-supervised or semi-supervised learning fashion. Generally, self-supervised learning takes advantage of automatically generated labels for model training, while semi-supervised learning leverages both labeled and unlabeled data to improve the generalization capability of models.

Traditional self-supervised and semi-supervised learning methods usually target at image classification (Wu & Prasad, 2017; Zhai et al., 2019; Misra & Maaten, 2020), object detection (Gao et al., 2019; Tang et al., 2017), and semantic segmentation (Wei et al., 2018; Wang et al., 2020) tasks. Unlike these tasks, FAR is a multi-label learning task, where facial attributes are comprised of global attributes (such as the "Male" attribute) and local attributes (such as the "Smiling" attribute) according to different regions of interest. To predict these attributes, a comprehensive understanding of the spatial-semantic relationship of facial images plays a critical role. For instance, the "Male" and "Attractive" attributes are identified by extracting the semantic information from the whole facial region. Similarly, to predict the "Smiling" and "Mouth-Open" attributes, it is natural to locate the mouth region and determine whether the mouth is smiling and open at a semantic level. Therefore, it is of great significance to learn fine-grained feature representations, in particular capturing the spatial-semantic relationship, for FAR.

Motivated by the above observations, we propose a novel spatial-semantic patch learning network (SPL-Net) method, which effectively exploits the spatial-semantic relationship on large-scale unlabeled facial data, for FAR with limited labeled data. SPL-Net consists of a multi-branch shared subnetwork (MSS), three auxiliary task subnetworks (ATS), and an FAR subnetwork. For MSS, it includes a task-shared branch (denoted T_B) and four region branches (denoted R_B).

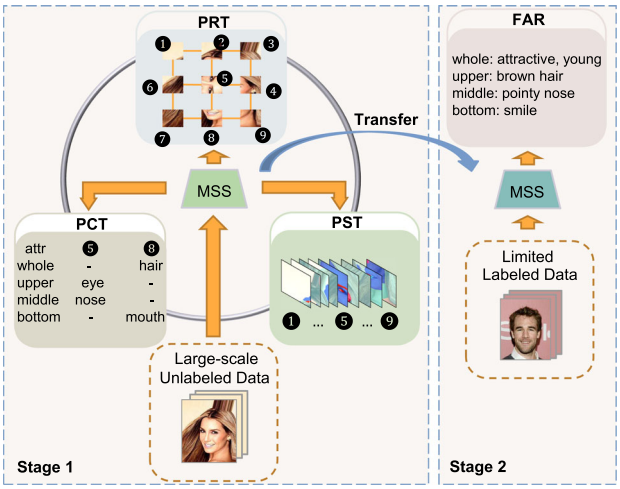

Fig. 1 Illustration of the two-stage learning procedure of our proposed SPL-Net method. In the first stage, MSS and ATS (including PRT, PST, and PCT subnetworks) are jointly trained to perform three auxiliary tasks on large-scale unlabeled data, and a well-pretrained MSS is obtained. In the second stage, the pre-trained MSS is transferred to perform FAR with limited labeled data

T_B extracts shared features from input facial images, while R_B aggregates features from T_B based on cascaded dual cross attention modules. For ATS, it contains a patch rotation task (PRT) subnetwork, a patch segmentation task (PST) subnetwork, and a patch classification task (PCT) subnetwork.

The training of SPL-Net involves a two-stage learning procedure. In the first stage, MSS and ATS are jointly trained to perform three auxiliary tasks on large-scale unlabeled facial data. Therefore, a powerful pre-trained MSS is obtained. Specifically, based on T_B, PRT identifies the rotated patch given several facial patches (one of which is rotated) and PST performs semantic segmentation on a randomly cropped facial patch. Meanwhile, based on R_B, PCT predicts facial components for the same patch in PST. In this way, PRT captures the spatial information of facial images, while PST and PCT 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 can in turn lead to a significant improvement of FAR when only limited labeled data are available. In the second stage, an FAR model (consisting of the pre-trained MSS and the FAR subnetwork) is easily fine-tuned to classify attributes by using labeled data. Figure 1 illustrates the training process of the proposed SPL-Net method.

In summary, the main contributions of our work are as follows:

– We propose a novel SPL-Net to address the problem of FAR with limited labeled data. SPL-Net effectively exploits the spatial-semantic information on unlabeled

facial data to learn a powerful pre-trained model. Therefore, we are able to obtain an accurate attribute prediction model by simply fine-tuning the pre-trained model with limited labeled data.
– We elaborately design three auxiliary tasks to make use of the intrinsic dependencies between patch rotation prediction and patch segmentation/classification. This enables the pre-trained model to extract patch-level fine-grained feature representations.
– Experimental results show that our proposed method consistently outperforms several state-of-the-art methods in the case of limited labeled data, which shows the importance of exploring the spatial-semantic relationship for predicting facial attributes.

This paper is a substantial extension of our previous conference work (Shu et al., 2021). The method in our previous work predicts all the facial attributes based on the same features extracted from the backbone. However, as we mentioned above, identifying global and local attributes generally relies on different facial regions. Therefore, our previous work does not fully exploit the characteristics of different facial attributes. SPL-Net alleviates this limitation from two main aspects. First, we design MSS and the PCT subnetwork with multiple branches to classify facial components. Such a way explicitly accounts for the differences between facial components in the auxiliary task, and thus in turn benefits the FAR model to predict global and local attributes. Second, we innovatively associate different component labels with the corresponding attribute labels to effectively model the instrinsic relationship between facial components and facial attributes. Hence, we can perform PCT in the first stage and FAR in the second stage by using the same network architecture. In this manner, the extended auxiliary task is more suitable for FAR with limited labeled data.

To summarize, we have added the following new significant contributions:

– We design a multi-branch shared subnetwork MSS to encode the region-specific information for different facial regions (which naturally correspond to different attribute groups). In particular, we leverage adversarial training between the whole region branch and the three local region branches. Hence, the whole region branch can capture the global context in facial images, even when randomly cropped facial patches are used as inputs for training those region branches.
– We extend the original PCT subnetwork to the multi-branch structure for classifying facial components. In the PCT subnetwork, we introduce a spatial mutual exclusion loss that explicitly enforces each local branch to focus on its corresponding facial region. This is helpful to classify diverse attributes in the FAR task with limited labeled data.
– By virtue of the above extensions, our new SPL-Net achieves better recognition accuracy than our previous method. Furthermore, we validate the superiority of SPL-Net on the newly released MAAD dataset (Terhörst et al., 2020).

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 introduces the details of our proposed SPL-Net method. Section 4 provides experimental results on three facial attribute datasets. Finally, Sect. 5 presents the conclusion.

2 Related Work

In this section, we review the related work, including facial attribute recognition and learning from unlabeled data, which is closely related to our method.

2.1 Facial Attribute Recognition (FAR)

Currently, deep learning-based methods have become dominant in the field of FAR. They can be roughly categorized into two groups: part-based methods and holistic methods (Zheng et al., 2020).

Part-based methods first locate the regions for different facial attributes, and then predict each attribute in a specific facial region. For example, SPLITFACE (Mahbub et al., 2018) takes several facial segments and a whole facial image as the input and identifies attributes. Kalayeh et al. (2017) leverage a deep semantic segmentation network to improve the prediction of facial attributes. Unlike part-based methods, holistic methods pay more attention to model the relationships among attributes. For instance, Mao et al. (2020) propose to perform FAR based on a deep multi-task and multi-label convolutional neural network (DMM-CNN). Considering the correlations and distinctions between different attributes, several methods perform FAR based on attribute grouping. He et al. (2019) divide facial attributes into six groups and propose an adaptive threshold algorithm to classify attributes. Cao et al. (2018a) resort to the auxiliary information (i.e., attribute grouping and identity information) to customize the network architecture and boost the FAR performance by capturing the local geometric structure.

The above methods learn the optimized network parameters by training on large-scale labeled data. However, in many real-world applications, a large number of labels can be difficult to collect. As a result, the performance of these methods is greatly influenced when only a few labeled training data are available. Different from these methods, we address the challenging and little-explored problem of FAR with limited

labeled data. In particular, we design SPL-Net with three auxiliary tasks to capture the spatial-semantic relationship on large-scale unlabeled facial data. In this way, a powerful pre-trained model can be obtained and then fine-tuned to accurately classify facial attributes by using only limited labeled data.

2.2 Learning from Unlabeled Data

To alleviate the extensive expense of annotating large-scale data, various methods have been developed by learning from unlabeled data. Among them, self-supervised learning and semi-supervised learning are the two popular paradigms.

Self-Supervised Learning Self-supervised learning methods often learn general features from large-scale unlabeled data without using any human-annotated labels (Jing & Tian, 2021). For example, Caron et al. (2018) employ an image clustering algorithm to generate labels for image classification. Noroozi and Favaro (2016) divide the images into nine patches and shuffle these patches. Then, a pretext task is designed to establish correct spatial positions of input patches by solving the jigsaw puzzle. Gidaris et al. (2018) develop a self-supervised learning method to predict the geometric transformation of images.

Recently, contrastive learning has been widely studied in self-supervised learning. He et al. (2020) develop momentum contrast (MoCo) by constructing dynamic dictionaries for unsupervised visual representation learning. They formulate an instance discrimination task to determine whether a query and a key are encoded views (e.g., different crops) of the same image. Chen et al. (2020) combine several data augmentation methods to transform each sample to generate two correlated views of the same sample, and use convolutional networks to extract image features. Then, a multi-layer perceptron (MLP) is employed to obtain the nonlinear projection of image features, thereby improving the representation quality of features.

Semi-Supervised Learning Current semi-supervised learning methods mainly contain two categories: generative methods and teacher-student methods (Qi & Luo, 2020).

The generative methods learn the real data distribution from training data and then generate new data according to the distribution. Salimans et al. (2016) use a generative adversarial network (GAN) to generate virtual samples, where the unlabeled and generated samples are classified into real classes and a fake class, respectively. They further combine the classification loss and the unsupervised GAN loss to train the model.

For teacher-student methods, a teacher model is first trained to predict the proxy labels of unlabeled data. Then, both labeled and unlabeled data (with the proxy labels) are employed to train a student model. MixMatch (Berthelot et al., 2019) identifies low-entropy labels for data-augmented unlabeled data, and then mixes labeled and unlabeled data based on MixUp (Zhang et al., 2017a). FixMatch (Sohn et al., 2020) introduces a strong augmentation and a weak augmentation to an unlabeled sample, and predicts the labels for the two types of augmentations. Virtual adversarial training (VAT) (Miyato et al., 2018) develops a novel regularization method based on the virtual adversarial loss, which defines the virtual adversarial direction on unlabeled data.

The above methods often learn holistic feature representations in a variety of computer vision tasks, including image classification, object detection, and semantic segmentation. However, they may not be suitable for the FAR task, where each facial attribute is associated with a specific facial region of interest. In this paper, three auxiliary tasks are designed and jointly performed to model the spatial-semantic relationship between facial regions by leveraging patch rotation prediction and patch segmentation/classification. Moreover, MSS is introduced to extract region-specific features which naturally correspond to different attribute groups. In this way, fine-grained feature representations are extracted by our method, which can largely facilitate FAR.

3 Proposed Method

In this section, we first give an overview of the proposed method in Sect. 3.1. Then, we introduce MSS in Sect. 3.2. Next, we describe the details of three auxiliary tasks and the FAR model in Sects. 3.3 and 3.4, respectively. Finally, we summarize the overall training of our method in Sect. 3.5.

3.1 Overview

The network architecture of our proposed SPL-Net method is illustrated in Fig. 2. SPL-Net involves MSS for extracting mid-level features, ATS for performing multi-auxiliary task learning, and an FAR subnetwork for predicting attributes. To address the problem of FAR with limited labeled data, we introduce a two-stage learning procedure. In the first stage, MSS and ATS are jointly trained to perform three auxiliary tasks (i.e., PRT, PST, and PCT) and learn fine-grained feature representations encoding the spatial-semantic information on large-scale unlabeled facial data. Hence, a powerful pre-trained MSS is obtained. In the second stage, an FAR model (consisting of the pre-trained MSS followed by the FAR subnetwork) is easily fine-tuned to classify facial attributes by using only a small amount of labeled facial data.

For PRT, it encodes the spatial information of facial images in a self-supervised learning manner. Specifically, an input facial image is divided into several patches, one of which is randomly chosen and rotated. Then, PRT identifies the rotated patch. For PST and PCT, they respectively exploit the pixel-level and image-level semantic information

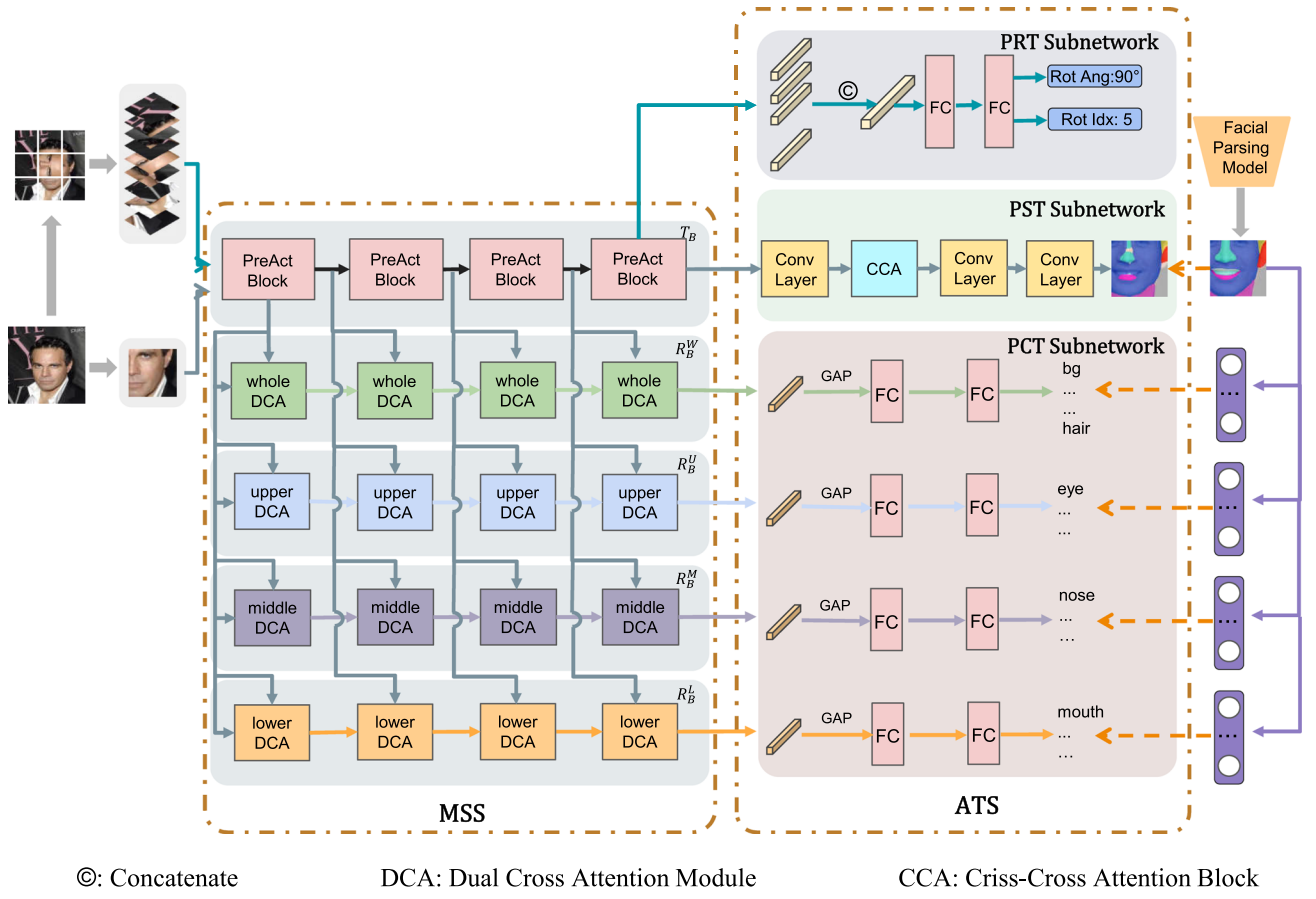

Fig. 2 The network architecture of the proposed SPL-Net method. SPL-Net involves MSS, ATS (consisting of PRT, PST, and PCT subnetworks), and an FAR subnetwork (which adopts the network structure, consisting of GAP layers and FC layers, same as the PCT subnetwork). MSS contains T_B and R_B (including R_B^W, R_B^U, R_B^M, and R_B^L). T_B is based on PreAct ResNet-18 while each region branch of R_B is comprised of cascaded DCA modules

of facial images. To achieve this, PST performs semantic segmentation on a randomly cropped facial patch and assigns a semantic label to each pixel in this patch, while PCT predicts facial components for the same input patch in PST.

Note that the ground-truth semantic labels and facial component labels are usually not provided in facial attribute datasets. In this paper, we take advantage of an externally trained facial parsing model (BiSeNetV2 (Yu et al., 2021)) to generate proxy semantic labels and proxy facial component labels (obtained by aggregating predicted semantic labels from BiSeNet) for PST and PCT, respectively. Therefore, during the training of auxiliary tasks in the first stage, all the labels are automatically generated to reduce the burden of labeling large-scale facial data.

3.2 Multi-branch Shared Subnetwork (MSS)

MSS includes a task-shared branch (denoted T_B) and four region branches (denoted R_B), as shown in Fig. 2. In this paper, T_B, which is based on PreAct ResNet-18 (He et al., 2016b) (consisting of four PreAct blocks), extracts features for both the PRT and PST subnetworks. R_B, which contains a whole region branch (denoted R_B^W), an upper region branch (denoted R_B^U), a middle region branch (denoted R_B^M), and a lower region branch (denoted R_B^L), extracts four different region-specific features for the PCT subnetwork. These four branches share the same network architecture, and each of them is composed of cascaded dual cross attention (DCA) modules.

The detailed network architecture of the DCA module is given in Fig. 3.

For the channel path, given two input features $\mathbf{f}_1$ and $\mathbf{f}_2$, they are first concatenated along the channel dimension to obtain a concatenated feature $\mathbf{f}_{con}$, i.e., $\mathbf{f}_{con} = \text{concat}(\mathbf{f}_1, \mathbf{f}_2)$, where concat$(\cdot)$ represents the channel-wise concatenation operation. Then, $\mathbf{f}_{con}$ is fed into a channel attention (CA) block (Hu et al., 2018) to calculate the channel attention mask $\mathbf{m}_{CA}$, i.e., $\mathbf{m}_{CA} = \text{CA}(\mathbf{f}_{con})$, where CA$(\cdot)$ represents the CA block. Next, the output feature $\mathbf{f}_{CA}$ of the channel path is derived by adding $\mathbf{f}_{con}$ to the product between $\mathbf{m}_{CA}$

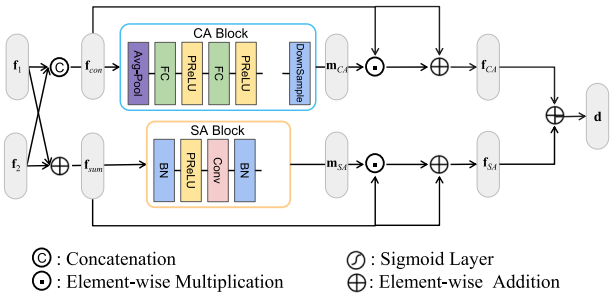

Fig. 3 The network architecture of the dual cross attention (DCA) module

and $\mathbf{f}_{con}$, which can be expressed as

$$\mathbf{f}_{CA} = \mathbf{f}_{con} \oplus (\mathbf{m}_{CA} \odot \mathbf{f}_{con}), \tag{1}$$

where '$\odot$' and '$\oplus$' denote the element-wise multiplication and element-wise addition operations, respectively.

For the spatial path, $\mathbf{f}_1$ and $\mathbf{f}_2$ are first added together to obtain a feature $\mathbf{f}_{sum}$, i.e., $\mathbf{f}_{sum} = \mathbf{f}_1 \oplus \mathbf{f}_2$. Instead of concatenating the features, the element-wise addition is advantageous to preserve spatial details of two features. Then, $\mathbf{f}_{sum}$ is fed into a spatial attention (SA) block (Woo et al., 2018) to calculate the spatial attention mask $\mathbf{m}_{SA}$, i.e., $\mathbf{m}_{SA} = \mathrm{SA}(\mathbf{f}_{sum})$, where $\mathrm{SA}(\cdot)$ represents the SA block. Next, the output feature $\mathbf{f}_{SA}$ of the spatial path is derived by adding $\mathbf{f}_{sum}$ to the product between $\mathbf{m}_{SA}$ and $\mathbf{f}_{sum}$, which can be formulated as

$$\mathbf{f}_{SA} = \mathbf{f}_{sum} \oplus (\mathbf{m}_{SA} \odot \mathbf{f}_{sum}). \tag{2}$$

Finally, the output feature $\mathbf{d}$ of the DCA module is obtained by combining $\mathbf{f}_{CA}$ with $\mathbf{f}_{SA}$ as

$$\begin{aligned} \mathbf{d} &= \mathrm{DCA}(\mathbf{f}_1, \mathbf{f}_2) \\ &= \mathbf{f}_{CA} \oplus \mathbf{f}_{SA}, \end{aligned} \tag{3}$$

where $\mathrm{DCA}(\cdot, \cdot)$ denotes the DCA module.

Similar to existing attention mechanisms (Zhang et al., 2018; Fu et al., 2019; Chen et al., 2017; Zhao et al., 2018), the DCA module involves a channel path and a spatial path. Concretely, it combines the CA block in SENet (Hu et al., 2018) and the SA block in CBAM (Woo et al., 2018). In fact, the DCA module can be comprised of any CA and SA blocks. Note that existing attention mechanisms take in a single feature as the input and generate an enhanced feature representation. Nevertheless, unlike these mechanisms, the DCA module accepts two features (i.e., one is from the PreAct Block in T_B and the other is from the previous DCA module in R_B^k) as the input. The concatenation operation and the element-wise addition operation are individually used to combine the two input features before the CA and SA blocks. By aggregating the features along the channel and spatial dimensions, we can effectively exploit the shared information from T_B and the region-specific information from R_B.

As shown in Fig. 2, the DCA module at the first layer of R_B^k ($k \in \{W, U, M, L\}$) takes the feature $\mathbf{o}_1$ from the first PreAct block in T_B and its copy as the input. For the DCA module at the n-th ($n \in \{2, 3, 4\}$) layer of R_B^k, the feature $\mathbf{o}_n$ from the n-th PreAct block in T_B and the attention feature $\mathbf{d}_{n-1}^k$ from the previous DCA module in R_B^k are taken as the input. Therefore, the output feature $\mathbf{d}_n^k$ of the n-th DCA module in R_B^k can be described as

$$\mathbf{d}_n^k = \begin{cases} \mathrm{DCA}(\mathbf{o}_n, \mathbf{o}_n), & n = 1, \\ \mathrm{DCA}(\mathbf{o}_n, \mathbf{d}_{n-1}^k), & n \geq 2. \end{cases} \tag{4}$$

On the one hand, if the whole facial image is used as the input of MSS, PCT is trained with similar facial component labels while PRT leverages shortcuts to identify the rotated patch (detailed explanations will be described in Sect. 3.3). In this way, both PCT and PRT fail to perform well on auxiliary tasks. Therefore, the whole region branch in MSS adopts a randomly cropped facial patch as the input in the first stage. On the other hand, the whole region branch is designed to capture the global context information of the whole facial image. To address this, we take advantage of adversarial training between the whole region branch and the three local region branches to enforce the whole region branch to aggregate the information from local branches.

Specifically, a feature fusion block consisting of a convolutional layer and a batch normalization layer is used to aggregate three region-specific features $\mathbf{d}_4^U$, $\mathbf{d}_4^M$, and $\mathbf{d}_4^L$ from the 4-th DCA modules of three local region branches, which can be expressed as

$$\mathbf{d}_4^{agg} = \mathrm{g}(\mathrm{concat}(\mathbf{d}_4^{\mathrm{U}}, \mathbf{d}_4^{\mathrm{M}}, \mathbf{d}_4^{\mathrm{L}})), \tag{5}$$

where $\mathbf{d}_4^{agg}$ is the aggregated feature and $\mathrm{g}(\cdot)$ denotes the convolutional operation followed by batch normalization.

Then, the distributions of $\mathbf{d}_4^{agg}$ and $\mathbf{d}_4^W$ extracted by R_B^W are constrained to be as close as possible. In this way, the feature extracted from the whole region branch can easily capture the global semantic context with the help of three region-specific features from different local region branches. To achieve this, a discriminator D (consisting of four fully-connected (FC) layers) is introduced to play a mini-max game between R_B^W and R_B^U, R_B^M, R_B^L. That is, R_B^W tries to minimize the divergence between $\mathbf{d}_4^{agg}$ and $\mathbf{d}_4^W$, while D aims to distinguish $\mathbf{d}_4^{agg}$ from $\mathbf{d}_4^W$. Mathematically, adversarial training can be formulated as

$$\min_D \max_{R_B^W} \mathcal{L}_{MSS}^{adv}(R_B^W, D), \tag{6}$$

where the adversarial loss $\mathcal{L}_{MSS}^{adv}$ is defined as

$$\mathcal{L}_{MSS}^{adv} = -\mathbb{E}[\log(D(\mathbf{d}_4^{agg}))] - \mathbb{E}[\log(1 - D(\mathbf{d}_4^{W}))]. \quad (7)$$

The whole region branch is optimized to extract features similar to the aggregated features from three region-specific features. Meanwhile, notice that the whole region branch is optimized with the joint loss (Eq. (16)) containing the classification loss corresponding to the whole facial components in the first stage. It is also fine-tuned with the classification loss (Eq. (17)) corresponding to the global attributes and the whole facial images as inputs in the second stage. Therefore, by back-propagating the gradients of the loss, the whole region branch can hold the global view of facial images to some extent.

Note that both PS-MCNN (Cao et al., 2018a) and our MSS adopt the multi-branch structure to extract features for different facial regions. However, these two methods are significantly different. PS-MCNN aggregates features from different branches by a simple concatenation layer. In contrast, our MSS aggregates features by employing an attention module (i.e., DCA), which emphasizes the important information and suppresses the irrelevant information in the features along the channel and spatial dimensions. By leveraging cascaded DCA modules, each region branch learns informative features more effectively. Besides, compared with ResNet-50 (He et al., 2016a) used in our previous work (Shu et al., 2021), our MSS exploits the spatial characteristics of facial images by extracting region-specific features since each facial attribute corresponds to a specific facial region. Hence, a well-pretrained MSS can be obtained in the first stage and facilitate the training of the FAR model in the second stage, as verified in our experiments in Sec. 4.3.

3.3 Auxiliary Tasks

In this subsection, we give the details of three auxiliary tasks.

3.3.1 Patch Rotation Task (PRT)

We design PRT to model the spatial relationship between facial patches. As illustrated in Fig. 2, the network architecture of PRT contains T_B and a PRT subnetwork (composed of a global average pooling (GAP) layer and two FC layers).

Given an input facial image $\mathbf{I}$ from unlabeled facial data, it is first evenly divided into $m \times m$ different patches, denoted by $\{\mathbf{p}_1, \cdots, \mathbf{p}_{m^2}\}$. Then, one patch $\mathbf{p}_r$ is randomly chosen and rotated by degree d that is randomly selected from 90, 180, and 270 degrees. PRT takes these patches as the input and aims to identify the rotated patch and the corresponding rotation angle.

Note that the random selection of a patch is guided by the semantic mask (see Sect. 3.3.2 for more details) generated by BiSeNetV2 (Yu et al., 2021). That is, when a selected patch contains only the background, we will discard it and choose another patch randomly until it involves the facial component.

To be specific, these $m \times m$ patches are first concatenated along the channel dimension, and fed into a preprocessing block (consisting of a 1×1 convolutional layer followed by a batch normalization layer and a PReLU layer) to reduce the number of feature channels and improve the training efficiency. Then, the output from the preprocessing block is passed through several PreAct blocks to extract the patch feature $\mathbf{p}_{PRT} \in \mathbb{R}^{c \times w \times h}$, where c, w, and h represent the channel, width, and height of the feature, respectively. Next, the patch feature is fed into a GAP layer to obtain a feature $\mathbf{f}_{PRT}$. After that, $\mathbf{f}_{PRT}$ is flattened and fed into two FC layers and two softmax layers to predict the probabilities of m^2 patches being rotated and the probabilities of three degrees being chosen, i.e., $\boldsymbol{t}^p = [t_1^p, \cdots, t_{m^2}^p] \in \mathbb{R}^{1 \times m^2}$ with $t_i^p \in [0, 1]$, and $\boldsymbol{t}^r = [t_1^r, \cdots, t_3^r] \in \mathbb{R}^{1 \times 3}$ with $t_i^r \in [0, 1]$. The index of the largest element in $\boldsymbol{t}^p$ corresponds to that of the predicted rotated patch, and the index of the largest element in $\boldsymbol{t}^r$ indicates the predicted rotation angle.

Similar to Noroozi and Favaro (2016), we apply color jitter to each patch and then normalize each patch independently. In this way, we avoid the model simply taking shortcuts between low-level texture statistics (e.g., edge continuity, pixel intensity distribution, and chromatic aberration) when identifying the rotated patch. Therefore, the network is capable of extracting high-level primitives and structures, thus effectively modeling the spatial relationship between a patch and its neighboring patches.

The loss of PRT employs the standard cross-entropy loss, which is formulated as

$$\mathcal{L}_{PRT} = -\left(\sum_{i=1}^{m^2} \mathbb{1}_{[i=r]} \log(t_i^p) + \sum_{i=1}^{3} \mathbb{1}_{[i=d]} \log(t_i^r)\right), \quad (8)$$

where $\log(\cdot)$ denotes the logarithm function; $\mathbb{1}_{[i=r]}$ outputs 1 when $i = r$ and 0 otherwise; $\mathbb{1}_{[i=d]}$ outputs 1 when $i = d$ and 0 otherwise.

It is worth pointing out that Gidaris et al. (2018) develop a self-supervised learning method to predict the rotation angle of an input image. However, this method is originally designed for image classification, object detection, and semantic segmentation, and thus it does not fully take into account the intrinsic geometric structure of images. For FAR, different facial attributes are often associated with different facial regions. Hence, by exploiting the spatial contextual information between patches, our design of PRT is more appropriate for the FAR task.

3.3.2 Patch Segmentation Task (PST)

We develop PST to perform semantic segmentation, which predicts the semantic label of each pixel in a patch. Conventional semantic segmentation methods often consider the whole image as the input. However, such a manner may cause PRT to leverage shortcuts (such as low-level statistics in facial images) to identify the rotated patch since PST and PRT share the same T_B. Therefore, we use a randomly cropped facial patch as the input of PST.

As shown in Fig. 2, the network architecture of PST consists of T_B and a PST subnetwork (composed of a convolutional layer and a Criss-cross attention (CCA) block (Huang et al., 2019) followed by two convolutional layers. Different from the encoder-decoder structure used in our previous work (Shu et al., 2021), the PST subnetwork aggregates full-patch dependencies in horizontal and vertical directions by the CCA block. This way accurately captures the contextual information from all patch pixels and benefits the performance improvement of semantic segmentation.

Specifically, a $c \times c$ patch $\mathbf{p}_s$ is randomly cropped from the original facial image $\mathbf{I}$ and used as the input of PST. The patch is fed into T_B to extract a feature, which is then passed through the PST subnetwork to classify each pixel into different semantic classes. Suppose that we have J semantic classes and the class prediction probabilities for the d-th pixel are denoted $\boldsymbol{h} = [h_{d1}, \cdots, h_{dJ}]$, we can formulate the loss of the d-th pixel in $\mathbf{p}_s$ as

$$\mathcal{L}_{pixel} = -\sum_{j=1}^{J} q_{dj} \log(h_{dj}), \tag{9}$$

where q_{dj} denotes the label distribution; $q_{dj} = 1$ if j is the ground-truth label of the d-th pixel and $q_{dj} = 0$ otherwise.

Generally, the semantic labels of facial images are not available in facial attribute datasets. Therefore, we make use of an externally trained facial parsing model (i.e., BiSeNetV2 (Yu et al., 2021)) to predict the semantic labels for all pixels of the input patch $\mathbf{p}_s$. These predicted labels are used as the proxy semantic labels for PST. In this paper, BiSeNetV2 is pre-trained on ImageNet and fine-tuned with only limited labeled data (we employ the same number of training data in CelebA-HQ (Karras et al., 2017) as that of limited labeled data used in the second stage, instead of using the whole CelebA-HQ).

BiSeNetV2 may give incorrect proxy semantic labels when applied to facial attribute datasets due to domain discrepancy and limited training data. To alleviate the overfitting caused by incorrect labels, we further leverage the label smoothing strategy (Szegedy et al., 2016), which is formulated as

$$q'_{dj} = (1 - \epsilon) q_{dj} + \frac{\epsilon}{J}, \tag{10}$$

where q'_{dj} is the modified label distribution and ϵ is a smoothing parameter empirically set to 0.1 as in Szegedy et al. (2016).

With Eqs. (9) and (10), the loss of PST is defined as

$$\mathcal{L}_{PST} = \frac{1}{D} \sum_{d=1}^{D} \left(-\sum_{j=1}^{J} q'_{dj} \log(h_{dj}) \right), \tag{11}$$

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

3.3.3 Patch Classification Task (PCT)

PST encodes the pixel-level semantic information of facial images by performing semantic segmentation. Nonetheless, the FAR task is an image-level multi-attribute classification task, where each facial attribute often corresponds to the semantic context of a whole/local facial region. Hence, we further develop PCT to predict facial components of a given input. In this way, the image-level semantic information of facial images can be explicitly captured.

PCT adopts the same input (i.e., a randomly cropped facial patch) as PST. Note that, if the whole facial image is taken as the input, most facial components exist and thus PCT is trained with similar facial component labels. Such a manner is detrimental to the PCT training since the distribution of facial component labels is highly imbalanced.

As shown in Fig. 2, the network architecture of PCT is composed of T_B, R_B, and a PCT subnetwork (consisting of four parallel GAP layers and four parallel FC layers). As we mentioned previously, each branch of R_B aggregates features from T_B based on cascaded DCA modules according to a specific region of interest. Therefore, R_B can extract both the global and local information of a given input facial image. More specifically, given a facial patch $\mathbf{p}_s$, it is first fed into T_B and R_B to extract four region-specific features. Then, these features are fed into the PCT subnetwork to predict facial components.

In this paper, the facial components predicted in PCT are the same as the semantic classes used in PST. However, PST and PCT are two different tasks. PST is a pixel-level classification task (i.e., assigning a label to each pixel in a patch) while PCT is an image-level classification task (i.e., predicting the existence of facial components in a patch).

Due to the lack of ground-truth facial component labels in facial attribute datasets, we also employ BiSeNet to assign the proxy facial component labels of an input patch. Each proxy label is generated by aggregating pixel-level semantic labels predicted by BiSeNet, and thus it is tolerant of small label errors. Thus, the proxy component labels of the input patch are denoted as a vector, that is, $\mathbf{y}_s = [y_0, \cdots, y_J]$.

Fig. 4 Examples of two input facial patches and their corresponding semantic masks from CelebA. The "ear" and "right eye" exist in **a** patch 1 and **b** patch 2, respectively. But they are not the dominant facial components

Here, $y_i = 1$ denotes the existence of a facial component, and 0 otherwise. In particular, we divide J facial component labels into four groups (a whole group W, an upper group U, a middle group M, and a lower group L) according to their spatial locations, and each group has J_k ($k \in \{W, U, M, L\}$) facial component labels. The detailed group configuration is listed in Sect. 4.2. Accordingly, the PCT subnetwork involves a whole branch and three local branches, where each branch predicts facial components in a group.

Usually, a few facial components exist in $\mathbf{p}_s$, where some of them only involve a relatively small number of pixels. Some examples are illustrated in Fig. 4. Hence, we only choose the top v dominant facial components in the patch and label them as 1. For the rest of facial components, we label them as 0.

The classification loss of PCT adopts the binary cross-entropy loss, which is defined as

$$\mathcal{L}_{PCT}^{cls} = -\sum_{k}\sum_{j=1}^{J_k}\left(z_j^k \log(p_j^k) + (1 - z_j^k)\log(1 - p_j^k)\right), \quad k \in \{W, U, M, L\}, \tag{12}$$

where p_j^k is the output prediction probability of the j-th facial component in group k; z_j^k denotes the proxy facial component label of the j-th facial component; $z_j^k = 1$ indicates the existence of a facial component, and 0 otherwise.

To explicitly enforce each local branch of the PCT subnetwork to focus on its corresponding facial region, we propose a spatial mutual exclusion (SME) loss. Specifically, a normalization operation is first applied to the outputs (denoted by $\mathbf{l}^U$, $\mathbf{l}^M$, and $\mathbf{l}^L$) of local branches of the PCT subnetwork, and thus the normalized features $\bar{\mathbf{l}}^U$, $\bar{\mathbf{l}}^M$, and $\bar{\mathbf{l}}^L$ are

$$\bar{\mathbf{l}}^k = \text{sigmoid}(\mathbf{l}^k - \mathbf{m}), k \in \{U, M, L\}, \tag{13}$$

where $\mathbf{m} = (\mathbf{l}^U \oplus \mathbf{l}^M \oplus \mathbf{l}^L)/3$ represents the average feature, and sigmoid$(\cdot)$ is the Sigmoid function which maps the value of an element in $\mathbf{l}^k$ larger than $\mathbf{m}$ closer to 1 and that smaller than $\mathbf{m}$ closer to 0.

Then, the SME loss is defined as

$$\mathcal{L}_{PCT}^{sme} = \bar{\mathbf{l}}^U \odot \bar{\mathbf{l}}^M \odot \bar{\mathbf{l}}^L. \tag{14}$$

By minimizing the SME loss, three local branches are concerned with different facial regions.

With Eqs. (7), (12), and (14), the loss of PCT is given as

$$\mathcal{L}_{PCT} = \mathcal{L}_{MSS}^{adv} + \mathcal{L}_{PCT}^{cls} + \mathcal{L}_{PCT}^{sme}. \tag{15}$$

3.3.4 Joint Loss

Based on the above formulation, the joint loss of SPL-Net can be derived as

$$\mathcal{L}_{joint} = \mathcal{L}_{PRT} + \lambda_1 \mathcal{L}_{PST} + \lambda_2 \mathcal{L}_{PCT}, \tag{16}$$

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

3.4 FAR Model

After the joint training of three auxiliary tasks in the first stage, a comprehensively pre-trained MSS is learned. Then, an FAR model, containing the pre-trained MSS and an FAR subnetwork (consisting of four parallel GAP layers and four parallel FC layers), is fine-tuned to predict facial attributes in the second stage.

Given an input facial image $\mathbf{I}$ with C attribute labels, it is first fed into T_B to extract features. Then, four region branches extract region-specific features from T_B. Finally, these features are fed into the FAR subnetwork to predict facial attributes. According to the different spatial locations of facial attributes, all the attribute labels are divided into four groups $\{W, U, M, L\}$, where each group has C_k ($k \in \{W, U, M, L\}$) attribute labels. The detailed group configuration is given in Sect. 4. Therefore, each branch of the FAR subnetwork classifies facial attributes in a group.

The loss of FAR adopts the binary cross-entropy loss, which is defined as

$$\mathcal{L}_{FAR} = -\sum_{k}\sum_{i=1}^{C_k}\left(y_i^k \log(x_i^k) + (1 - y_i^k)\log(1 - x_i^k)\right), \quad k \in \{W, U, M, L\}, \tag{17}$$

where x_i^k represents the output prediction probability of the i-th facial attribute in a branch of the FAR subnetwork; y_i^k represents the ground-truth label of the i-th facial attribute; $y_i^k = 1$ indicates the existence of a facial attribute, and 0 otherwise.

Algorithm 1 The two-stage learning procedure of SPL-Net.

Require: Unlabeled facial data $\mathcal{U}$; labeled facial data $\mathcal{L}$; the training epochs of each stage, K_1, K_2; the number of steps to update the discriminator, K_d; the number of image patches, $m \times m$.
Ensure: A trained FAR model.
// Stage 1: Performing multi-auxiliary task learning.
for each $k_1 = 1$ to K_1 **do**
  for each mini-batch $\mathcal{U}_b$ in $\mathcal{U}$ **do**
    for $i = 1$ to $|\mathcal{U}_b|$ **do**
      Randomly crop a patch $\mathbf{p}_s$ from $\mathbf{I} \in \mathcal{U}_b$;
      Divide $\mathbf{I}$ into $m \times m$ patches $\mathbf{P} = \{\mathbf{p}_1, \cdots, \mathbf{p}_{m^2}\}$;
      Randomly select a patch $\mathbf{p}_r$ from $\mathbf{P}$ and rotate it by one randomly chosen degree from 90, 180, and 270 degrees;
    end for
    for $k_d = 1$ to K_d **do**
      Calculate the adversarial loss $\mathcal{L}_{MSS}^{adv}$ by Eq. (7);
      Fix MSS and three auxiliary task subnetworks, and update the D;
    end for
    Calculate the joint loss $\mathcal{L}_{joint}$ by Eq. (16);
    Fix the D, and update MSS and three auxiliary task subnetworks;
  end for
end for
// Stage 2: Fine-tuning the FAR model.
for each $k_2 = 1$ to K_2 **do**
  for each mini-batch in $\mathcal{L}$ **do**
    Calculate the FAR loss $\mathcal{L}_{FAR}$ by Eq. (17);
    Update MSS and the FAR subnetwork simultaneously;
  end for
end for

3.5 Overall Training

The overall training process of SPL-Net is summarized in Algorithm 1. Generally, it involves a two-stage learning procedure. In the first stage, three auxiliary tasks are jointly performed to capture the spatial-semantic relationship on large-scale unlabeled facial data in a multi-task learning fashion. Thus, a pre-trained MSS is learned. In the second stage, an FAR model is fine-tuned with limited labeled facial data.

4 Experiments

In this section, we perform extensive experiments to show the superiority of our proposed SPL-Net method. First, we briefly introduce three public facial attribute datasets. Then, we give the implementation details. Next, we perform ablation studies to validate the effectiveness of each auxiliary task in SPL-Net, and discuss the influence of several key parameters of SPL-Net on the final performance. Finally, we compare SPL-Net with several state-of-the-art methods and analyze the computational complexity of SPL-Net.

4.1 Datasets

CelebA (Liu et al., 2015) is a popular large-scale facial attribute dataset, which is widely used to evaluate the FAR performance. It contains 202,599 facial images with 40 attribute annotations per image. The facial images are collected with large pose variations, illumination changes, and background clutter. CelebA is split into 3 parts, including 162,770 images for training, 19,867 images for validation, and 19,962 images for testing.

LFWA (Huang et al., 2008) is another challenging facial attribute dataset. It consists of 13,143 facial images with the same attribute annotations as the CelebA dataset. Similar to CelebA, LFWA is divided into a training set (6263 images) and a test set (6880 images).

MAAD (Terhörst et al., 2020) is a newly-released massive facial attribute dataset. It is constructed based on the VGGFace2 database (Cao et al., 2018b) and consists of 3.3 M facial images with 123.9 M attribute labels of 47 attributes. In MAAD, 3,138,862 images and 169,178 images are used for training and testing, respectively.

In the first stage, we use the default training set (without labels) to train three auxiliary tasks for CelebA and LFWA. We randomly select 200,000 images from the training set (without labels) to train three auxiliary tasks for MAAD. In the second stage, we randomly choose a proportion of the training set (with labels) of CelebA, LFWA, or MAAD to fine-tune the FAR model. Moreover, we use the default validation and test sets of CelebA and LFWA, while we randomly select 20,000 images from the test set of MAAD, to evaluate the performance. All the experiments are performed 10 times, and the average recognition accuracy is reported.

4.2 Implementation Details

We use PreAct ResNet-18 (without pre-training) as the backbone of T_B and each region branch in R_B^k ($k \in \{W, U, M, L\}$) is comprised of four cascaded DCA modules. In PRT, the number of patches per side m is set to 3. Hence, there are $3 \times 3 = 9$ patches in total. Each facial image $\mathbf{I}$ in unlabeled facial data is first resized to 255×255, and then 9 patches with the size of 85×85 are cropped. Finally, a patch with the size of 64×64 is randomly cropped from each 85×85 patch and resized to 224×224. Such a way prevents the model from using low-level texture statistics, which are not advantageous for the FAR task. In PST and PCT, a patch with the size of 75×75 is randomly cropped from each facial image, and then resized to 224×224. In PCT, the number

Table 1 Group configuration of facial component labels and attribute labels in CelebA, LFWA, and MAAD

Groups	Component Labels	Attribute Labels in CelebA/LFWA	Attribute Labels in MAAD
Whole Group	Background, Skin	5_o_Clock Shadow, Attractive, Blurry, Chubby, Heavy Makeup, Male, Oval Face, Pale Skin, Straight Hair, Smiling, Wavy Hair, Young	Male, Young, Middle Aged, Senior, Asian, White, Black Shiny Skin, Wavy_Hair, 5_o_Clock_Shadow, Oval_Face, Square_Face, Round_Face, Chubby, Smiling, Heavy_Makeup, Attractive
Upper Group	Left Eyebrow, Right Eyebrow, Left Eye, Right Eye, Eye Glasses, Hair, Hat	Arched Eyebrows, Bags Under Eyes, Bald, Bangs, Black Hair, Blond Hair, Brown Hair, Bushy Eyebrows, Eyeglasses, Gray Hair, Narrow Eyes, Receding Hairline, Wearing Hat	Bald, Receding Hairline, Bangs, Black_Hair, Blond Hair, Brown Hair, Gray Hair, Obstructed Forehead, Fully Visible Forehead, Brown Eyes, Bags Under Eyes, Bushy Eyebrows, Arched Eyebrows, Wearing Hat, No Eyewear, Eyeglasses
Middle Group	Left Ear, Right Ear, Ear Ring, Nose	Big Nose, High Cheekbones, Pointy Nose, Rosy Cheeks, Sideburns, Wearing Earrings	Rosy Cheeks, Sideburns, High Cheekbones, Big Nose, Pointy Nose, Wearing Earrings
Lower Group	Mouth, Upper Lip, Lower Lip, Neck, Necklace, Cloth	Big Lips, Double Chin, Goatee, Mustache, Mouth Slightly Open, No Beard, Wearing Lipstick, Wearing Necklace, Wearing Necktie	No Beard, Mustache, Goatee, Double Chin,Mouth Closed, Big Lips, Wearing Necktie, Wearing Lipstick

of dominant facial components v is set to 9. The number of attributes C is 40 for CelebA and LFWA, and 47 for MAAD. The number of facial components J is 19.

We use PyTorch to implement SPL-Net, and all the experiments are performed on four GTX 2080 GPUs. For the first stage, the batch size is set to 40, and the model is trained for 80 epochs. The number of steps to update the discriminator D is set to 3. The values of λ_1 and λ_2 in Eq. (16) are empirically set to 0.05 and 0.50, respectively. For the second stage, the batch size is set to 128, and the model is trained for 60 epochs.

During training, the Adam optimizer (Kingma & Ba, 2014) is adopted with the initial learning rate of 1×10^{-4}, $\beta_1 = 0.500$, $\beta_2 = 0.999$ and the weight decay of 5×10^{-4}. The warm-up strategy is used to update the learning rate, 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. As mentioned in Sect. 3, both facial component labels in PCT and facial attribute labels in FAR are divided into four groups (i.e., a whole group, an upper group, a middle group, and a lower group) according to different spatial locations. The detailed group configuration is shown in Table 1. We use BiSeNetV2 (Yu et al., 2021), which is pre-trained on ImageNet and fine-tuned with only limited labeled data, to generate semantic masks for training auxiliary tasks. In particular, we select the same number of training data in CelebA-HQ (Karras et al., 2017) as that of limited labeled data in the facial attribute dataset. All experiments on speed analysis are performed by using a single NVIDIA GTX 2080 GPU.

4.3 Ablation Studies

To show the effectiveness of the proposed SPL-Net method, we conduct ablation studies to evaluate the influence of the DCA module, the whole region branch, MSS, different auxiliary tasks (i.e., PRT, PST, and PCT), the SME loss, adversarial training in MSS, the two-stage learning procedure, and critical parameters (including the number of patches and the number of dominant facial components) on the final recognition performance.

We evaluate the performance obtained by sixteen variants of the proposed method, including: 1) the baseline method that uses the PreAct ResNet-18 backbone and two FC layers to predict facial attributes; 2) the method (denoted "SPL_CBAM") that is the same as SPL_Net except that the channel and spatial attention blocks in the DCA module are replaced by those in CBAM; 3) the method (denoted

Table 2 The details of sixteen variants of SPL-Net

Variants	DCA	MSS	PRT	PST	PCT	AT	W	SME	TS
Baseline	–	–	–	–	–	–	–	–	–
SPL_CBAM	CBAM	✓	✓	✓	✓	✓	✓	✓	✓
SPL_w/o_whole	✓	✓	✓	✓	✓	✓	–	✓	✓
MSS	✓	✓	–	–	–	–	–	–	✓
SPL_R	–	–	✓	–	–	–	–	–	✓
SPL_S	–	–	–	✓	–	–	–	–	✓
SPL_C	✓	✓	–	–	✓	✓	✓	–	✓
SPL_C_w/o_A	✓	✓	–	–	✓	–	–	–	✓
SPL_RS	–	–	✓	✓	–	✓	✓	–	✓
SPL_RC	✓	✓	✓	–	✓	✓	✓	–	✓
SPL_SC	✓	✓	–	✓	✓	✓	✓	–	✓
SPL_w/o_A	✓	✓	✓	✓	✓	–	–	✓	✓
SPL_w/o_SME	✓	✓	✓	✓	✓	✓	✓	–	✓
SPL_L2	✓	✓	✓	✓	✓	L2	–	✓	✓
SPL_Semi	✓	✓	✓	✓	–	–	–	✓	One-stage
SPL-Net	✓	✓	✓	✓	✓	✓	✓	✓	✓

AT denotes adversarial training. W denotes the whole region branch. TS denotes the two-stage learning procedure

"SPL_w/o_whole") that is the same as SPL_Net except that the whole region branch is replaced by the aggregation (i.e., the feature fusion block) after adversarial training in the second stage; 4) the method (denoted "MSS") that is based on MSS and the FAR subnetwork; Note that both the baseline and MSS methods are directly trained by using limited labeled data. 5) the method (denoted "SPL_R") that only adopts PRT as the auxiliary task; 6) the method (denoted "SPL_S") that only adopts PST as the auxiliary task; 7) the method (denoted "SPL_C") that only adopts PCT as the auxiliary task; 8) the method (denoted "SPL_C_w/o_A") that only adopts PCT as the auxiliary task without using adversarial training; 9) the method (denoted "SPL_RS") that adopts PRT and PST as the auxiliary tasks; 10) the method (denoted "SPL_RC") that uses PRT and PCT as the auxiliary tasks; 11) the method (denoted "SPL_SC") that uses PST and PCT as the auxiliary tasks; 12) the method (denoted "SPL_w/o_A") that jointly trains PRT, PST, and PCT in an integrated network but without using adversarial training in MSS; 13) the method (denoted "SPL_w/o_SME") that jointly combines PRT, PST, and PCT but without using the SME loss in PCT; 14) the method (denoted "SPL_L2") that is the same as SPL_Net except that adversarial training is replaced by a simple contrastive learning method (based on the L2 loss); 15) the method (denote "SPL_semi") that jointly trains PRT, PST, and FAR in a semi-supervised manner; and 16) the proposed SPL-Net method.

The details of these variants are summarized in Table 2. The results obtained by these variants with the different proportions of labeled training data on CelebA, LFWA, and MAAD are given in Tables 3, 4, and 5, respectively.

Influence of the DCA Module We validate the effectiveness of DCA via replacing the attention blocks in the DCA module by those in CBAM. Experimental results show that SPL-Net can achieve slightly better performance than SPL_CBAM. This can be ascribed to the superiority of the SE block, which effectively recalibrates channel-wise feature responses by exploiting interdependencies between different channels.

Influence of the Whole Region Branch We validate the importance of the whole region branch in MSS. The adversarial training introduced in MSS encourages the whole region branch to extract features close to the aggregated features from the three local branches. Therefore, we can replace the whole region branch by the aggregation (i.e., the feature fusion block) after adversarial training in the second stage.

We can see that SPL_w/o_whole cannot achieve satisfactory performance. This is because the feature fusion block (consisting of only a simple convolutional layer and a batch normalization layer) cannot successfully learn powerful feature representations for classifying facial attributes, when limited labeled data are given. In contrast, the whole region branch involving cascaded DCA modules provides better feature extraction capability and can be more effectively fine-tuned by taking the whole facial images as inputs in the second stage.

Influence of the Multi-branch Shared Subnetwork (MSS) MSS includes a task-shared branch and four region branches (each branch is composed of cascaded DCA modules). As observed from Tables 3, 4, and 5, the MSS method obtains better performance than the baseline method on all the three

Table 3 Ablation studies: the recognition accuracy (%) obtained by sixteen variants of SPL-Net with the different proportions of labeled training data on the CelebA dataset

	CelebA					
Proportion	0.02%	0.2%	0.5%	1%	2%	100%
Number of labeled samples	33	325	843	1627	3225	162,770
Baseline	76.34	82.16	85.23	87.60	88.40	90.90
SPL_CBAM	79.03	86.95	88.15	88.77	89.53	91.68
SPL_w/o_whole	78.21	85.65	87.24	86.34	88.43	90.55
MSS	76.92	83.97	86.33	87.82	88.85	91.49
SPL_R	78.38	85.25	87.67	88.58	89.13	91.70
SPL_S	77.53	84.58	87.13	87.87	88.77	91.53
SPL_C	77.23	83.87	86.40	87.59	88.65	91.38
SPL_C_w/o_A	76.95	83.35	86.01	87.02	88.10	91.21
SPL_RS	78.52	85.77	87.97	88.97	89.75	91.70
SPL_RC	78.50	86.14	87.65	88.76	89.32	91.71
SPL_SC	78.04	85.15	87.32	88.01	88.98	91.60
SPL_w/o_A	78.89	86.68	88.09	87.66	88.23	91.66
SPL_w/o_SME	78.83	86.60	87.86	87.41	88.05	91.53
SPL_L2	78.01	85.84	87.21	86.95	87.43	89.64
SPL_semi	75.23	83.12	85.75	85.94	86.45	89.57
SPL-Net	**79.33**	**87.02**	**88.21**	**88.97**	**89.83**	**91.78**

The best results are boldfaced

Table 4 Ablation studies: the recognition accuracy (%) obtained by sixteen variants of SPL-Net with the different proportions of labeled training data on the LFWA dataset

	LFWA					
Proportion	0.5%	5%	10%	20%	50%	100%
Number of labeled samples	31	313	626	1252	3131	6263
Baseline	67.37	73.92	77.04	80.90	83.65	85.76
SPL_CBAM	71.72	78.89	82.01	84.23	85.80	86.58
SPL_w/o_whole	70.25	77.47	80.33	82.21	83.15	85.60
MSS	68.35	74.96	78.79	82.14	84.53	86.14
SPL_R	69.68	77.01	80.85	83.19	85.25	86.59
SPL_S	68.53	76.01	79.79	82.37	84.70	86.32
SPL_C	68.42	75.51	79.22	82.10	84.31	86.01
SPL_C_w/o_A	68.03	75.30	78.87	81.81	83.65	85.98
SPL_RS	70.15	77.45	81.59	83.31	85.23	86.47
SPL_RC	70.31	77.54	81.60	83.42	85.42	86.54
SPL_SC	69.25	76.15	79.31	82.52	84.88	86.46
SPL_w/o_A	71.30	78.34	81.05	82.84	83.51	85.01
SPL_w/o_SME	71.22	78.40	81.21	82.35	83.81	85.03
SPL_L2	70.58	77.46	80.47	81.21	81.63	83.45
SPL_semi	68.30	70.14	74.25	77.80	80.05	83.01
SPL-Net	**71.88**	**79.20**	**82.12**	**84.43**	**85.86**	**86.77**

The best results are boldfaced

datasets. More specifically, the MSS method improves the performance by 0.58% on CelebA, 0.98% on LFWA, and 0.57% on MAAD, when 0.02%, 0.5%, and 0.02% of labeled training data are respectively used. The above results show the effectiveness of MSS, which can extract region-specific features according to the regions of interest, for improving the FAR performance.

Influence of Different Auxiliary Tasks SPL-Net outperforms the MSS method by 2.41% on CelebA, 3.53% on LFWA, and 5.40% on MAAD when 0.02%, 0.5%, and 0.02% of labeled training data are respectively used. Generally, when a smaller proportion of labeled training data is employed, the improvements obtained by SPL-Net are more evident. In particular, SPL-Net outperforms the baseline method (2.99%, 4.51%, and 5.97% improvements on CelebA, LFWA, and MAAD,

Table 5 Ablation studies: the recognition accuracy (%) obtained by sixteen variants of SPL-Net with the different proportions of labeled training data on the MAAD dataset

	MAAD					
Proportion	0.02%	0.2%	0.5%	1%	2%	100%
Number of labeled samples	40	400	1000	2000	4000	200,000
Baseline	63.04	67.18	70.25	71.19	74.92	85.86
SPL_CBAM	68.83	73.69	76.21	77.88	79.05	85.67
SPL_w/o_whole	67.21	72.39	75.37	76.70	78.75	84.21
MSS	63.61	67.72	70.90	73.67	75.32	85.88
SPL_R	67.55	70.50	73.55	76.92	78.59	85.71
SPL_S	66.83	69.13	72.41	75.87	77.21	85.34
SPL_C	66.65	68.14	71.86	75.01	76.93	85.06
SPL_C_w/o_A	66.31	67.79	70.87	74.31	76.44	85.02
SPL_RS	67.83	71.56	74.30	77.15	79.01	85.92
SPL_RC	67.76	71.67	74.81	77.17	78.90	85.80
SPL_SC	67.13	70.55	73.34	75.99	78.15	85.13
SPL_w/o_A	68.45	72.41	75.83	76.14	78.20	84.89
SPL_w/o_SME	68.55	73.24	76.01	77.51	78.05	85.11
SPL_L2	67.13	71.53	74.81	75.57	76.25	84.14
SPL_semi	65.45	69.21	72.57	73.89	75.10	84.12
SPL-Net	**69.01**	**73.98**	**76.39**	**77.97**	**79.21**	**85.94**

The best results are boldfaced

respectively) when 0.02%, 0.5%, and 0.02% of labeled training data are respectively used. This validates the importance of exploiting the spatial-semantic relationship to ensure the performance of the SPL-Net method.

PRT exploits the spatial information of facial images based on self-supervised learning. Compared with SPL_S and SPL_C, SPL_RS and SPL_RC give higher accuracy on the CelebA, LFWA, and MAAD datasets. Moreover, SPL-Net also achieves better recognition accuracy than SPL_SC. The above results show the effectiveness of PRT, which takes advantage of spatial information to improve the FAR performance in the case of limited labeled data.

PST leverages semantic segmentation to extract the fine-grained semantic information from facial images. As shown in Tables 3, 4 and 5, SPL_RS and SPL_SC obtain higher accuracy than SPL_R and SPL_C, respectively. Introducing the pixel-level semantic information in the first stage is helpful to improve the final FAR performance in the second stage. In comparison with SPL_RC, SPL-Net achieves higher accuracy (e.g., 0.83%, 1.57%, and 1.25% improvements on CelebA, LFWA, and MAAD, respectively, when 0.02%, 0.5%, and 0.02% of labeled training data are respectively used). Hence, the pixel-level semantic segmentation is beneficial to boost the final FAR performance.

PCT capitalizes on the semantic relationship to identify facial components. SPL_RC and SPL_SC achieve higher accuracy than SPL_R and SPL_S, respectively. Compared with SPL_RS, SPL_w/o_A also improves the performance on CelebA, LFWA, and MAAD (i.e., 0.37%, 1.15%, and 0.62% improvements in terms of recognition accuracy on CelebA, LFWA, and MAAD when 0.02%, 0.5%, and 0.02% of labeled training data are adopted, respectively). Therefore, the image-level semantic information is also important to enhance the FAR performance with limited labeled data.

By combing PRT, PST, and PCT with adversarial training, SPL-Net gives the top performance among all the variants. Therefore,modeling the spatial-semantic relationship of facial images is advantageous for the FAR task.

Influence of the Spatial Mutual Exclusion (SME) Loss We evaluate the importance of the SME loss in Tables 3, 4 and 5. Compared with SPL_Net, SPL_w/o_SME obtains worse performance (0.50%, 0.66%, and 0.46% drop on CelebA, LFWA, and MAAD, respectively, when 0.02%, 0.5%, and 0.02% of labeled training data are respectively used). By minimizing the SME loss, SPL-Net explicitly enforces different local branches to focus on their corresponding regions, benefiting the model to extract region-specific features. This improves the performance of the FAR model when limited labeled data are used for fine-tuning.

Influence of the Adversarial Training Strategy From Tables 3, 4, and 5, in all six proportions on three datasets, SPL-Net outperforms SPL_w/o_A (e.g. 0.44% improvements on CelebA, 0.58% improvements on LFWA, and 0.56% improvements on MAAD, when 0.02%, 0.5%, and 0.02% of labeled training data are respectively used). Compared with SPL_C_w/o_A, SPL_C obtains higher performance. The above results demonstrate the importance of the adversarial training strategy adopted in MSS.

Moreover, we compare our adversarial learning with a simple contrastive learning method (we adopt the L2 loss

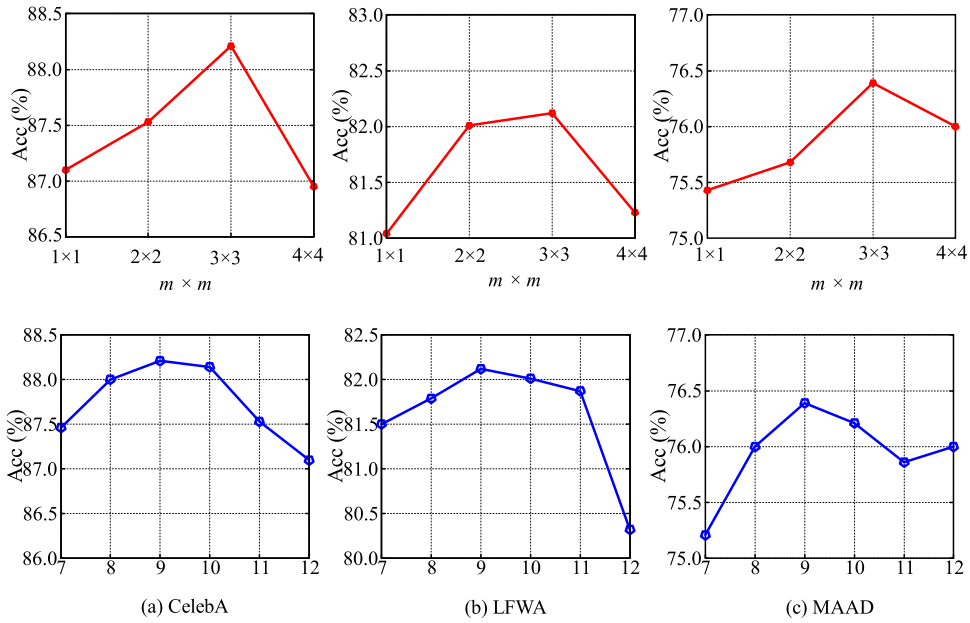

Fig. 5 Ablation studies: Influence of the number of patches (the first row) and the number of dominant facial components (the second row) on the final performance when 0.5%, 10% and 0.5% of the labeled training data of **a** CelebA, **b** LFWA, and **c** MAAD are used, respectively

between the features from the whole region branch and the aggregated features from the three local branches). We can see that our method with adversarial learning achieves much better performance than that with contrastive learning (i.e., SPL_L2). Adversarial training is a generative model, which matches the distribution of generated features from the whole region branch to the distribution of aggregated features from the three local branches. Adversarial training pursues distribution consistency, enabling different branches to learn diverse feature representations. Such a way benefits feature extraction of the FAR task. In contrast, contrastive learning only reduces the distances between two features to be as close as possible, limiting the diversity of local branches.

Influence of the Two-stage Learning Procedure SPL-Net adopts the two-stage learning procedure (i.e., performing auxiliary tasks with large-scale unlabeled data in the first stage and performing FAR with limited labeled data in the second stage). Alternatively, we can design a one-stage learning method that performs multi-task learning (based on the multi-branch architecture) in a semi-supervised manner. To be specific, two branches perform PRT and PST (trained with large-scale unlabeled data), while one branch performs FAR (trained with limited labeled data). In this manner, we can jointly train these tasks in a single stage.

From Tables 3, 4, and 5, we can see that SPL_semi based on one-stage learning achieves much worse results than SPL-Net based on two-stage learning. This is because the one-stage learning method does not fully exploit the spatial-semantic relationship of facial images. The joint learning of these tasks cannot effectively guide the model to extract discriminative features for predicting facial attributes. Note that multi-task learning can boost the performance in the case that multiple tasks are correlated or complementary to each other (Zhao et al., 2018). However, PRT, PST, and FAR are weak in terms of task relevance.

In contrast, the two-stage learning procedure follows the pre-training and fine-tuning paradigm. This shows the importance of obtaining a powerful pre-trained model, as validated in recent research (Chen et al., 2021).

Influence of the Number of Patches $m \times m$ We evaluate the performance of SPL-Net with the different numbers of patches $m \times m$ (including 1×1, 2×2, 3×3, and 4×4) in PRT. The experimental results on CelebA, LFWA, and MAAD are shown in the first row of Fig. 5. SPL-Net achieves the best performance, when the number of patches $m \times m$ is set to 3×3. When the number of patches is larger, the semantically consistent facial components (such as the eye, nose, and mouth) are over-segmented into small patches. On the other hand, when the number of patches is smaller, the large patch involves many facial components. In both cases, the feature extraction capability of PRT to exploit the spatial information is adversely affected.

Influence of the Number of Dominant Facial Components v We further evaluate the influence of the number of dominant facial components in PCT on the final performance. The experimental results on three datasets are given in the second row of Fig. 5. Our SPL-Net method achieves the best recognition performance when the value of v is set to 9. Note that the input facial patch of PCT is randomly cropped from the facial image. Hence, some facial components involve only a few pixels. On the one hand, when the values of v are too large, the facial components with a small number of pixels are chosen as dominant facial components. On the other hand,

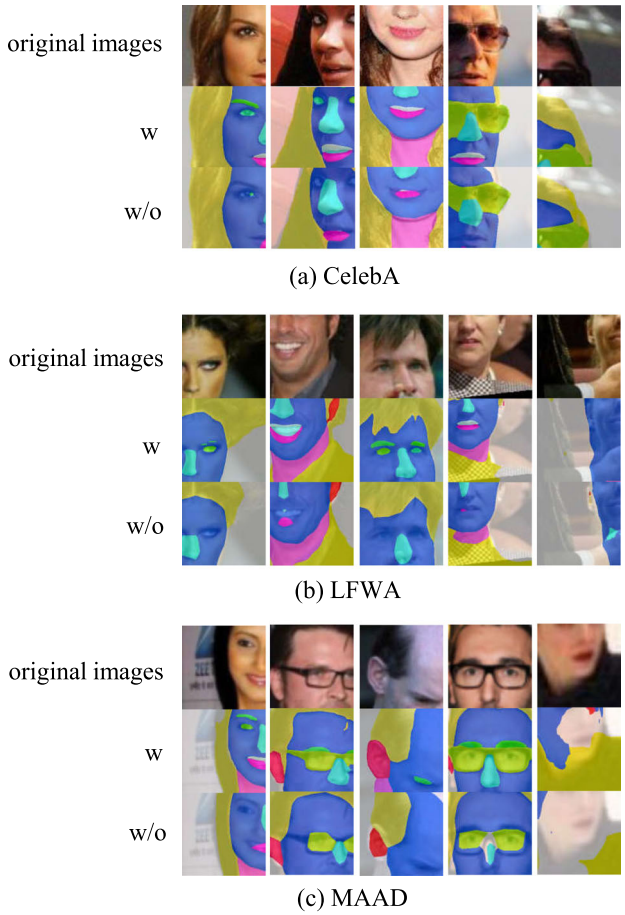

Fig. 6 Semantic masks generated by SPL-Net with (denoted w) and without the label smoothing strategy (denoted w/o) on **a** CelebA, **b** LFWA, and **c** MAAD

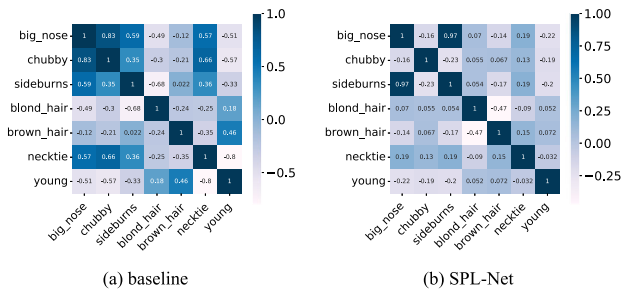

Fig. 7 The correlation maps of seven randomly selected facial attributes obtained by **a** the baseline and **b** SPL-Net on CelebA

when the values of v are too small, some dominant facial components are ignored. This is harmful to learn the image-level semantic information. Both cases lead to performance degradation.

4.4 Visualization

In this subsection, we visualize several examples of semantic masks generated by SPL-Net with and without the label smoothing strategy. The results are illustrated in Fig. 6.

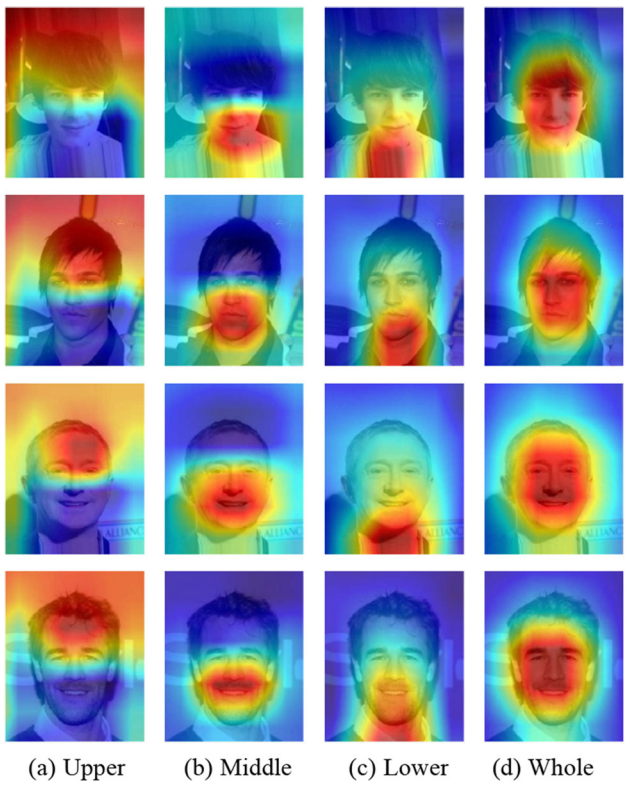

Fig. 8 Visualization of heat maps from different region branches of MSS: **a** the upper region branch, **b** the middle region branch, **c** the lower region branch, and **d** the whole region branch on CelebA

Moreover, we also plot the correlation maps of several randomly chosen facial attributes obtained by the baseline and SPL-Net methods, as shown in Fig. 7. We randomly choose seven facial attributes, and calculate the correlation map based on the predicted outputs of the trained models. Finally, we show the heat feature maps (we employ the attentive feature maps as done in Ruan et al. (2022)) from the four region branches of MSS in Fig. 8. Here, we employ 0.2% of labeled training data of CelebA to train SPL-Net and the baseline.

From Fig. 6, compared with SPL-Net without the label smoothing strategy, SPL-Net is able to generate the semantic masks with much less noise. This validates the importance of the label smoothing strategy. Based on accurate semantic masks, PST can capture the pixel-level semantic information more effectively. Such a manner is beneficial for training MSS. Notice that there are some false detected masks. For example, in the last column of LWFA in Fig. 6, the “ear" and “nose" are falsely detected since there are a small number of pixels for these facial components in the image. Meanwhile, in the last column of MAAD in Fig. 6, most pixels in the facial patches are classified as “face”, due to the lack of facial details caused by blurring. However, the false-detected masks have no significant influence on the final performance since the corresponding facial components are not dominant, and

the distorted facial details in blurry patches do not greatly contribute to the learning process of PST.

From Fig. 7, SPL-Net shows better correlation responses between facial attributes than the baseline method. For instance, the "blond_hair" attribute is negatively related to the "brown_hair" attribute (the correlation value is −0.47 obtained by SPL-Net and that is −0.24 by baseline), while the "sideburns" and "blond_hair" attributes are not so strongly correlated with each other (the correlation value is 0.054 obtained by SPL-Net and that is −0.68 by baseline).

In Fig. 8, the warm-toned parts of an image correspond to the regions with large values in the feature map, and vice versa. We can see the feature maps from different region branches focus on different facial regions. In particular, for the three local branches, their corresponding feature maps concentrate on local regions. For the whole region branch, its corresponding feature maps tend to pay attention to the whole facial regions. This can be ascribed to the MSS structure and the PCT subnetwork, which are supervised with the adversarial loss, the SME loss, and the classification loss.

4.5 Comparison with State-of-the-Art Methods

In this subsection, we compare the proposed SPL-Net method with several state-of-the-art methods, including five supervised FAR methods (DMM (Mao et al., 2020), SlimCNN (Sharma & Foroosh, 2020), AFFAIR (Li et al., 2018), PS-MCNN (Cao et al., 2018a), and FAN (He et al., 2018a)), five self-supervised learning methods (DeepCluster (Caron et al., 2018), JigsawPuzzle (Noroozi & Favaro, 2016), Rot (Gidaris et al., 2018), MoCo (He et al., 2020), and SimCLR (Chen et al., 2020)), and two semi-supervised learning methods (FixMatch (Sohn et al., 2020) and VAT (Miyato et al., 2018)), on the CelebA, LFWA, and MAAD datasets, respectively. Our previous SSPL method (Shu et al., 2021) is also evaluated for performance comparison. In particular, we evaluate two versions of SSPL (i.e., SSPL-w and SSPL-p), where SSPL-w and SSPL-p indicate that the facial parsing models are trained on the whole CelebA-HQ and limited labeled data of CelebA-HQ (same as SPL-Net), respectively. We retrained the models (including DeepCluster, JigsawPuzzle, Rot, MoCo, SlimCLR, FixMatch, and VAT) on a series of experiments according to the publicly available codes from their papers. Note that the results obtained by four state-of-the-art methods (DMM, AFFAIR, PS-MCNN, and FAN) are not listed on MAAD since their source codes are not publicly available. The results of these methods on CelebA and LFWA are taken from their respective papers.

For five supervised FAR methods, we only leverage the available labeled training data to train the FAR models. For self-supervised learning methods, we use all the unlabeled training data to obtain the pre-trained models in the pretext task, and then use the different proportions of labeled training data for fine-tuning in the downstream FAR task. For semi-supervised learning methods, we simultaneously train the models using both unlabeled and labeled training data. The accuracy obtained by all the competing methods with the different proportions of labeled training data on CelebA, LFWA, and MAAD are shown in Tables 6, 7, and 8.

We can observe that, compared with several state-of-the-art FAR methods (including DMM, SlimCNN, AFFAIR, and FAN), our SPL-Net method shows similar or better performance on the three datasets when 100% of labeled data are used to train the FAR models. State-of-the-art FAR methods are capable of extracting discriminative features for classifying facial attributes from large-scale labeled training data. Note that DMM predicts facial attributes based on a dynamic weighting scheme and an adaptive thresholding strategy. AFFAIR takes advantage of a unified transformation-localization architecture to capture a hierarchy of spatial transformations. Therefore, it can classify facial attributes without relying on landmark annotations or landmark detectors. PS-MCNN develops a network architecture consisting of four task-specific networks (TSNets) and a shared network (SNet) to extract features. FAN leverages abstraction images generated by GAN to locate facial parts. In contrast, SPL-Net makes full use of three auxiliary tasks, which can capture fine-grained spatial and semantic information for FAR. This demonstrates the effectiveness of the pre-trained MSS in the auxiliary tasks. Moreover, the proposed method achieves much better performance (from 79.90% to 87.02% on CelebA, from 70.90% to 79.20% on LFWA, and from 64.48% to 73.98% on MAAD) than Slim-CNN when only a small proportion of training data (i.e., 0.2%, 0.5%, or 0.2%) is used. This is because that we jointly train the auxiliary tasks to exploit the spatial-semantic relationship on unlabeled facial data. Therefore, effective semantic-aware global and local features can be extracted for the FAR task.

The SPL-Net method significantly outperforms the competing context-based self-supervised learning methods (i.e., DeepCluster, JigsawPuzzle, and Rot) under the small proportions of labeled training data. Compared with Rot, our method obtains 3.77%, 4.80%, and 2.92% improvements on CelebA, LFWA, and MAAD, when 0.2%, 5%, and 0.2% of labeled data are used, respectively. Notice that, when less labeled training data are used, the performance improvements obtained by our method are more evident than the competing self-supervised learning methods. These results indicate the good generalization ability of SPL-Net to perform FAR with limited labeled data. SPL-Net effectively exploits both spatial and semantic information on unlabeled facial data by leveraging three auxiliary tasks. Moreover, compared with contrastive learning-based self-supervised methods (i.e., MoCo and SimCLR), SPL-Net also achieves better accuracy. In particular, SPL-Net outper-

Table 6 The recognition accuracy (%) obtained by our proposed SPL-Net method and several state-of-the-art methods with the different proportions of labeled training data on the CelebA dataset

	CelebA					
Proportion	0.02%	0.2%	0.5%	1%	2%	100%
Number of labeled samples	33	325	843	1627	3225	162,770
DMM (Mao et al., 2020)	–	–	–	–	–	91.70
SlimCNN (Sharma & Foroosh, 2020)	67.32	79.90	80.20	80.96	82.32	91.24
AFFAIR (Li et al., 2018)	–	–	–	–	–	91.45
PS-MCNN (Cao et al., 2018a)	–	–	–	–	–	92.98
FAN (He et al., 2018a)	–	–	–	–	–	91.81
DeepCluster (Caron et al., 2018)	72.87	83.21	86.13	87.46	88.86	91.68
JigsawPuzzle (Noroozi & Favaro, 2016)	71.96	82.88	84.71	86.25	87.77	91.57
Rot (Gidaris et al., 2018)	73.82	83.25	86.51	87.67	88.82	91.69
MoCo (He et al., 2020)	78.34	85.09	87.44	88.43	89.06	91.66
SimCLR (Chen et al., 2020)	79.22	86.24	88.01	88.63	89.34	91.72
FixMatch (Sohn et al., 2020)	69.45	80.22	84.19	85.77	86.14	89.78
VAT (Miyato et al., 2018)	72.13	81.44	84.02	86.30	87.28	91.44
SSPL-w (Shu et al., 2021)	78.21	86.67	88.05	88.84	89.58	91.77
SSPL-p (Shu et al., 2021)	77.88	85.86	87.34	87.10	88.02	91.43
SPL-Net (Ours)	**79.33**	**87.02**	**88.21**	**88.97**	**89.83**	**91.78**

The best results are boldfaced

Table 7 The recognition accuracy (%) obtained by our proposed SPL-Net method and several state-of-the-art methods with the different proportions of labeled training data on the LFWA dataset

	LFWA					
Proportion	0.5%	5%	10%	20%	50%	100%
Number of labeled samples	31	313	626	1252	3131	6263
DMM (Mao et al., 2020)	–	–	–	–	–	86.56
SlimCNN (Sharma & Foroosh, 2020)	60.54	70.90	71.49	72.12	73.45	76.02
AFFAIR (Li et al., 2018)	–	–	–	–	–	86.13
PS-MCNN (Cao et al., 2018a)	–	–	–	–	–	87.36
FAN (He et al., 2018a)	–	–	–	–	–	85.20
DeepCluster (Caron et al., 2018)	63.97	74.21	77.42	80.77	84.27	85.90
JigsawPuzzle (Noroozi & Favaro, 2016)	63.32	73.90	77.01	79.56	83.29	84.86
Rot (Gidaris et al., 2018)	64.08	74.40	76.67	81.52	84.90	85.72
MoCo (He et al., 2020)	71.71	78.08	80.15	82.56	84.92	86.15
SimCLR (Chen et al., 2020)	70.49	78.63	80.66	82.73	85.44	86.24
FixMatch (Sohn et al., 2020)	62.87	71.42	72.78	75.10	80.87	83.84
VAT (Miyato et al., 2018)	62.96	72.19	74.42	76.26	80.55	84.68
SSPL-w (Shu et al., 2021)	71.64	78.68	81.65	83.45	85.43	86.53
SSPL-p (Shu et al., 2021)	70.43	76.23	89.26	82.87	84.01	86.21
SPL-Net (Ours)	**71.88**	**79.20**	**82.12**	**84.43**	**85.86**	**86.77**

The best results are boldfaced

forms the MOCO (0.99%, 0.17%, and 0.05% improvements on CelebA, LFWA, and MAAD, respectively) when 0.02%, 0.5%, and 0.02% of labeled training data are respectively used, while compared with SimCLR, the improvements are 0.11%, 1.39%, and 1.33% on CelebA, LFWA, and MAAD when 0.02%, 0.5%, and 0.02% of labeled training data are used, respectively.

Compared with those semi-supervised learning methods, our SPL-Net method achieves considerably higher accuracy in the case of limited labeled data. Among the competing semi-supervised learning methods, FixMatch simulta-

Table 8 The recognition accuracy (%) obtained by our proposed SPL-Net method and several state-of-the-art methods with the different proportions of labeled training data on the MAAD dataset

	MAAD					
Proportion	0.02%	0.2%	0.5%	1%	2%	100%
Number of labeled samples	40	400	1000	2000	4000	200,000
SlimCNN (Sharma & Foroosh, 2020)	58.23	64.48	65.04	65.88	66.45	83.00
DeepCluster (Caron et al., 2018)	62.37	71.53	73.57	76.02	78.69	85.92
JigsawPuzzle (Noroozi & Favaro, 2016)	60.18	65.84	65.74	74.14	76.04	85.34
Rot (Gidaris et al., 2018)	67.42	71.06	75.35	77.09	78.95	85.81
MoCo (He et al., 2020)	68.96	71.87	75.59	77.83	78.88	85.82
SimCLR (Chen et al., 2020)	67.68	72.28	76.27	78.02	79.23	85.84
FixMatch (Sohn et al., 2020)	63.97	68.74	69.23	72.01	73.52	80.93
VAT (Miyato et al., 2018)	64.23	69.88	71.34	73.91	75.34	82.18
SSPL-w (Shu et al., 2021)	68.82	72.46	76.24	77.99	79.30	85.88
SSPL-p (Shu et al., 2021)	67.15	71.21	75.83	76.02	77.92	85.34
SPL-Net (Ours)	**69.01**	**73.98**	**76.39**	**77.97**	**79.21**	**85.94**

The best results are boldfaced

neously introduces consistency regularization and proxy-labeling strategies, while VAT explores unlabeled data by minimizing the distances between images and transformed versions of these images. However, these methods focus on holistic features, and thus they cannot effectively model the spatial relationship, which plays a critical role for FAR. On the contrary, SPL-Net learns the spatial-semantic correlation of facial images and extracts fine-grained features, leading to superior performance.

It is worth pointing out that both SSPL-w and SSPL-p are based on ResNet-50, while SPL-Net uses a smaller backbone (ResNet-18) with cascaded attention blocks. In addition, SPL-Net and SSPL-p adopt limited labeled data of CelebA-HQ for training the facial parsing model, while SSPL-w leverages the whole CelebA-HQ for the training. However, both SSPL-w and SSPL-p do not fully consider the characteristics of FAR that facial attributes involve global and local attributes. In contrast, SPL-Net adopts MSS with four region branches to exploit the region-specific information for different attributes and model the attribute group relationship to boost the performance. Such a way benefits the model to predict global and local attributes in the FAR task. Therefore, SPL-Net can achieve higher performance than SSPL-w and SSPL-p. The above results validate the importance of exploiting the characteristics of facial attributes in designing the network architecture for FAR with limited labeled data.

Compared with SSPL-w and SSPL-p, the performance improvements of SPL-Net are not very significant on the three facial attribute datasets. This can be ascribed to the following four factors. First, the imbalanced class data distribution (Huang et al., 2019) (e.g., the imbalance ratios between the minority classes and the majority classes on the CelebA dataset are up to 1:43) exists in facial attribute datasets. Second, many facial attributes, especially for subjective attributes, have ambiguous annotations in these datasets (Yan et al., 2022). Third, some facial attributes may not be provided with positive samples due to limited labeled data. Fourth, SPL-Net employs much less annotated data than SSPL-w to train the facial parsing model, resulting in inferior semantic masks for learning the auxiliary tasks. This can affect the representation capability of the pre-trained model obtained in the first stage. The above factors greatly increase the training difficulty, making it extremely challenging to significantly improve the accuracy on these datasets.

We further report the accuracy obtained by each attribute, to more comprehensively evaluate different methods at one round of test. The results are given in Table 9, where 0.2% of labeled training data on CelebA are used. Experimental results clearly show that SPL-Net improves the accuracy corresponding to global attributes (such as the "Attractive" and "Young" attributes) and local attributes (such as the "Mouth_Open" and "Bangs" attributes), compared with the other competing methods. In particular, SPL-Net outperforms SSPL (both SSPL-w and SSPL-p) on most of facial attributes, showing the effectiveness of SPL-Net for FAR with limited labeled data. Generally, it is easier to identify objective attributes (such as the "Male" and "Hat" attributes) than subjective attributes (such as the "Oval_Face" and "Pointy_Nose" attributes). This is mainly because subjective attributes often appear in a subtle form, which makes the FAR model more difficult to learn the decision boundary.

We also observe that some attributes (such as the "Bald" attribute) are not chosen (i.e., only negative samples of these attributes are provided for the training) when a small proportion of labeled data are selected. However, the model can still predict these attributes. This can be ascribed to the pow-

Table 9 The recognition accuracy (%) obtained by each attribute when 0.2% of the labeled training data of CelebA are used

Attributes / Methods	Attractive	Mouth_ Open	Smiling	Lipstick Lipstick	High_ Cheekbones	Male	Heavy_ Makeup	Wavy_ Hair	Oval_ Face	Pointy_ Nose	Arched_ Eyebrows	Black_ Hair	Big_ Lips	Big_ Nose
Positive Samples	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SlimCNN (Sharma & Foroosh, 2020)	52	51	53	53	53	60	59	63	70	**71**	71	73	**67**	78
DeepCluster (Caron et al., 2018)	75	66	74	90	67	69	67	64	70	**71**	72	73	**67**	79
JigsawPuzzle (Noroozi & Favaro, 2016)	69	56	63	79	63	79	76	69	70	**71**	71	73	**67**	78
Rot (Gidaris et al., 2018)	72	51	51	87	52	87	82	68	70	**71**	73	72	**67**	79
MoCo (He et al., 2020)	74	80	87	**91**	**82**	93	**87**	67	**71**	**71**	**77**	73	**67**	80
SimCLR (Chen et al., 2020)	75	73	85	**91**	78	92	**87**	66	70	**71**	74	73	**67**	79
FixMatch (Sohn et al., 2020)	63	54	73	78	66	86	71	67	69	66	70	77	59	78
VAT (Miyato et al., 2018)	62	54	59	70	59	71	69	67	70	69	71	73	**67**	78
SSPL-w (Shu et al., 2021)	74	85	87	89	**82**	92	86	73	68	68	75	**83**	**67**	78
SSPL-p (Shu et al., 2021)	74	83	87	89	**82**	92	86	70	68	68	75	79	**67**	78
SPL-Net (Ours)	**76**	**87**	**88**	**91**	**82**	**94**	86	**73**	**71**	**71**	75	**83**	**67**	**81**

Attributes / Methods	Young	Straight_ Hair	Brown_ Hair	Bags_Under_ Eyes	Earrings	No_ Beard	Bangs	Blond_ Hair	Bushy_ Eyebrows	Necklace	Narrow_ Eyes	5_ Shadow	Receding_ Hairline
Positive Samples	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SlimCNN (Sharma & Foroosh, 2020)	76	78	82	79	79	85	84	87	87	**86**	**85**	**90**	91

Table 9 continued

Attributes Methods	Young	Straight_ Hair	Brown_ Hair	Bags_Under_ Eyes	Earrings	No_ Beard	Bangs	Blond_ Hair	Bushy_ Eyebrows	Necklace	Narrow_ Eyes	5_ Shadow	Receding_ Hairline
DeepCluster (Caron et al., 2018)	76	**79**	82	**80**	79	85	84	87	87	**86**	**85**	**90**	**92**
JigsawPuzzle (Noroozi & Favaro, 2016)	77	**79**	81	79	79	85	85	89	87	**86**	**85**	**90**	**92**
Rot (Gidaris et al., 2018)	76	**79**	82	**80**	79	85	84	87	87	**86**	**85**	**90**	**92**
MoCo (He et al., 2020)	77	**79**	82	**80**	80	88	84	87	87	**86**	**85**	**90**	**92**
SimCLR (Chen et al., 2020)	76	**79**	82	**80**	79	85	84	87	87	**86**	**85**	**90**	**92**
FixMatch (Sohn et al., 2020)	76	67	79	76	78	86	49	89	87	74	79	**90**	91
VAT (Miyato et al., 2018)	76	**79**	82	**80**	79	85	84	89	87	**86**	**85**	**90**	**92**
SSPL-w (Shu et al., 2021)	80	**79**	**84**	**80**	**81**	90	91	93	**89**	83	**85**	**90**	**92**
SSPL-p (Shu et al., 2021)	80	74	82	79	79	88	89	91	87	83	**85**	**90**	**92**
SPL-Net (Ours)	**82**	**79**	**84**	**80**	**81**	**91**	**93**	**94**	**89**	**86**	**85**	**90**	**92**

Attributes Methods	Necktie	Eyeglasses	Rosy_Cheeks	Goatee	Chubby	Sideburns	Blurry	Hat	Double_Chin	Pale_Skin	Gray_Hair	Mustache	Bald	Average
Positive Samples	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	–
SlimCNN (Sharma & Foroosh, 2020)	93	**94**	**93**	95	**95**	**96**	**95**	**96**	**95**	**96**	**97**	**96**	**98**	79.85

Table 9 continued

Attributes Methods	Necktie	Eyeglasses	Rosy_Cheeks	Goatee	Chubby	Sideburns	Blurry	Hat	Double_Chin	Pale_Skin	Gray_Hair	Mustache	Bald	Average
DeepCluster (Caron et al., 2018)	93	**94**	**93**	95	**95**	**96**	**95**	**96**	**95**	**96**	**97**	**96**	**98**	83.38
JigsawPuzzle (Noroozi & Favaro, 2016)	93	**94**	**93**	95	**95**	**96**	**95**	**96**	**95**	**96**	**97**	**96**	**98**	83.01
Rot (Gidaris et al., 2018)	93	**94**	**93**	95	**95**	**96**	**95**	**96**	**95**	**96**	**97**	**96**	**98**	83.05
MoCo (He et al., 2020)	93	**94**	**93**	95	**95**	**96**	**95**	**96**	**95**	**96**	**97**	**96**	**98**	86.43
SimCLR (Chen et al., 2020)	93	**94**	**93**	95	**95**	**96**	**95**	**96**	**95**	**96**	**97**	**96**	**98**	85.19
FixMatch (Sohn et al., 2020)	93	93	92	95	**95**	**96**	**95**	**96**	**95**	**96**	**97**	**96**	**98**	80.25
VAT (Miyato et al., 2018)	93	**94**	**93**	95	**95**	**96**	**95**	**96**	**95**	**96**	**97**	**96**	**98**	81.50
SSPL-w (Shu et al., 2021)	**95**	**94**	**93**	**96**	**95**	**96**	**95**	**96**	**95**	**96**	**97**	**96**	**98**	86.55
SSPL-p (Shu et al., 2021)	94	**94**	**93**	95	**95**	**96**	**95**	**96**	**95**	**96**	**97**	**96**	**98**	85.88
SPL-Net (Ours)	**95**	**94**	**93**	**96**	**95**	**96**	**95**	**96**	**95**	**96**	**97**	**96**	**98**	**87.14**

The best results are boldfaced

Table 10 The number of parameters and FLOPs obtained by different methods on the CelebA dataset

Methods		Params (M)	FLOPs (G)
SimCLR	–	35.298	6.231
SSPL	First stage	29.063	18.343
	Second stage	23.590	8.385
SPL-Net	first stage	26.354	10.328
	Second stage	21.721	4.132

Table 11 The inference time and speed obtained by different methods on the CelebA dataset

Methods	Inference time (ms)	Speed (FPS)
SimCLR	12.17	82.15
SSPL	23.98	41.69
SPL-Net	10.56	94.70

The inference time and speed are measured in milliseconds (ms) and frames per second (FPS), respectively

erful pre-trained model and the potential correlation among attributes (e.g., the "Bald" attribute and the "Male" attribute are highly correlated). Moreover, the number of positive samples with respect to these attributes is small in the test set (for example, the "Bald" attribute has 423 positive samples and 19,539 negative samples). Note that all the competing methods are evaluated under the same settings (i.e., we use the same randomly selected labeled training set and the same test set at each round of test).

4.6 Computational Complexity

In this subsection, we analyze the computational complexity of our proposed SPL-Net method. We also evaluate SSPL and the SlimCLR method for a comparison. We use the number of parameters (Params) and Floating-Point operations (FLOPs) to evaluate the memory consumption and computational cost of the model, respectively. Moreover, we adopt the inference time and speed to measure the latency. We take the CelebA dataset (0.2% of the labeled training data) for performance evaluation.

Table 10 gives the number of parameters and FLOPs obtained by SPL-Net, SSPL, and SlimCLR. SSPL has more parameters and higher FLOPs than SPL-Net. This is because SSPL adopts the larger ResNet-50 as the backbone. Both SSPL and SPL-Net have higher memory consumption and computational cost (in terms of Params and FLOPs) than SimCLR, since they involve the two-stage learning procedure. However, the second stage (i.e., the fine-tuning stage based on limited labeled data) in SPL-Net has fewer parameters and smaller FLOPs than SimCLR.

The inference time and speed obtained by SPL-Net, SSPL, and SlimCLR are reported in Table 11. We can observe that the proposed SPL-Net obtains smaller inference time than the other two competing methods. The inference speed of SPL-Net is also faster than those of SSPL and SlimCLR. Although the training complexity of SPL-Net is high, it still obtains real-time inference speed. Therefore, SPL-Net can be applicable in practice.

5 Conclusion

In this paper, we have proposed a novel SPL-Net method to perform FAR with limited labeled data effectively. The SPL-Net method involves a two-stage learning procedure. For the first stage, three auxiliary tasks (PRT, PST, and PCT) are jointly developed to exploit the spatial-semantic information on large-scale unlabeled facial data, and thus a powerful pre-trained MSS is obtained. For the second stage, only a few number of labeled facial data are leveraged to fine-tune the pre-trained MSS and an FAR model is finally learned. Extensive experiments on the CelebA, LFWA, and MAAD datasets have demonstrated the effectiveness of our proposed method in comparison with several state-of-the-art methods to address FAR in the case of limited labeled data.

Acknowledgements This work was partly supported by the National Natural Science Foundation of China under Grants 62071404, U21A20514, and 61872307, by the Open Research Projects of Zhejiang Lab under Grant 2021KG0AB02, by the Natural Science Foundation of Fujian Province under Grant 2020J01001, and by the Youth Innovation Foundation of Xiamen City under Grant 3502Z20206046.

Data Availability Statement The datasets that support the findings of this study are available in: CelebA: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, LFWA: http://vis-www.cs.umass.edu/lfw/, MAAD: https://github.com/pterhoer/MAAD-Face, CelebA-HQ: https://github.com/nperraud/download-celebA-HQ

References

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C. (2019). Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249.

Cao, J., Li, Y., Zhang, Z. (2018a). Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4290–4299).

Cao, Q., Shen, L., Xie, W., Parkhi, O. M, Zisserman, A. (2018b). Vggface2: A dataset for recognising faces across pose and age. In *Proceedings of the IEEE international conference on automatic face and gesture recognition* (pp. 67–74).

Caron, M., Bojanowski, P., Joulin, A., Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision* (pp. 132–149).

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W. (2021). Pre-trained image processing Transformer.

In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 12299–12310).

Chen, J. C., Ranjan, R., Sankaranarayanan, S., Kumar, A., Chen, C. H., Patel, V. M., Castillo, C. D., & Chellappa, R. (2018). Unconstrained still/video-based face verification with deep convolutional neural networks. *International Journal of Computer Vision, 126*(2), 272–291.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T. S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5659–5667).

Chen, T., Kornblith, S., Norouzi, M., Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the international conference on machine learning* (pp. 1597–1607).

Egger, B., Schönborn, S., Schneider, A., Kortylewski, A., Morel-Forster, A., Blumer, C., & Vetter, T. (2018). Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision, 126*(12), 1269–1287.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3146–3154).

Gao, J., Wang, J., Dai, S., Li, L. J., Nevatia, R. (2019). Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 9508–9517).

Gidaris, S., Singh, P., Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. In *International conference on learning representations*.

Hand, E., Chellappa, R. (2017). Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *Proceedings of the AAAI conference on artificial intelligence* (pp.1–7).

He, K., Zhang, X., Ren, S., Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, K., Zhang, X., Ren, S., Sun, J. (2016b). Identity mappings in deep residual networks. In *Proceedings of the European conference on computer vision* (pp. 630–645).

He, K., Fu, Y., Zhang, W., Wang, C., Jiang, Y. G., Huang, F., Xue, X. (2018a). Harnessing synthesized abstraction images to improve facial attribute recognition. In *Proceedings of the international joint conference on artificial intelligence* (pp. 733–740).

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9729–9738).

He, R., Wu, X., Sun, Z., & Tan, T. (2018). Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(7), 1761–1773.

He, X., Wang, P., Zhao, Z., Zhao, Y., Su, F. (2019). Mtcnn with weighted loss penalty and adaptive threshold learning for facial attribute prediction. In *Proceedings of the IEEE international conference on multimedia and expo workshops* (pp. 180–185).

Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Huang, C., Li, Y., Loy, C. C., & Tang, X. (2019). Deep imbalanced learning for face recognition and attribute prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(11), 2781–2794.

Huang, G. B., Mattar, M., Berg, T., Learned-Miller, E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: *Workshop on faces in'Real-Life'Images: Detection, alignment, and recognition*.

Huang, H., Li, Z., He, R., Sun, Z., Tan, T. (2018). Introvae: Introspective variational autoencoders for photographic image synthesis. In *Advances in neural information processing systems* (pp. 52–63).

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W. (2019). Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 603–612).

Jing, L., & Tian, Y. (2021). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(11), 4037–4058.

Kalayeh, M. M., Gong, B., Shah, M. (2017). Improving facial attribute prediction using semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6942–6950).

Karras, T., Aila, T., Laine, S., Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

Kingma, D. P., &Ba, J. (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Li, J., Zhao, F., Feng, J., Roy, S., Yan, S., & Sim, T. (2018). Landmark free face attribute prediction. *IEEE Transactions on Image Processing, 27*(9), 4651–4662.

Li, Y., Wang, R., Liu, H., Jiang, H., Shan, S., Chen, X. (2015). Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In *Proceedings of the IEEE international conference on computer vision* (pp. 3819–3827).

Liu, Z., Luo, P., Wang, X., Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (pp. 3730–3738).

Mahbub, U., Sarkar, S., & Chellappa, R. (2018). Segment-based methods for facial attribute detection from partial faces. *IEEE Transactions on Affective Computing, 11*(4), 601–613.

Mao, L., Yan, Y., Xue, J. H., Wang, H. (2020). Deep multi-task multi-label cnn for effective facial attribute classification. *IEEE Transactions on Affective Computing*.

Misra, I., & Maaten, L. V. D. (2020). Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6707–6717).

Miyato, T., Maeda, S., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(8), 1979–1993.

Nguyen, H. M., Ly, N. Q., Phung, T. T. (2018). Large-scale face image retrieval system at attribute level based on facial attribute ontology and deep neuron network. In *Proceedings of the Asian conference on intelligent information and database systems* (pp. 539–549).

Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European conference on computer vision* (pp. 69–84).

Qi, G. J., & Luo, J. (2020). Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Rao, Y., Lu, J., & Zhou, J. (2019). Learning discriminative aggregation network for video-based face recognition and person re-identification. *International Journal of Computer Vision, 127*(6), 701–718.

Ruan, D., Mo, R., Yan, Y., Chen, S., Xue, J. H., & Wang, H. (2022). Adaptive deep disturbance-disentangled learning for facial expression recognition. *International Journal of Computer Vision, 130*(2), 455–477.

Rudd, E. M., Günther, M., Boult, T. E. (2016). Moon: A mixed objective optimization network for the recognition of facial attributes. In

Proceedings of the European conference on computer vision (pp. 19–35).

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *Advances in Neural Information Processing Systems, 29*, 2234–2242.

Sharma A. K., Foroosh H. (2020). Slim-cnn: A light-weight cnn for face attribute prediction. In *Proceedings of the IEEE international conference on automatic face and gesture recognition* (pp. 329–335).

Shu, Y., Yan, Y., Chen, S., Xue, J. H., Shen, C., Wang, H. (2021). Learning spatial-semantic relationship for facial attribute recognition with limited labeled data. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 11916–11925).

Sohn, K., Berthelot, D., Li, C. L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., Raffel, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685.

Song, L., Zhang, M., Wu, X., He, R. (2018). Adversarial discriminative heterogeneous face recognition. In *Proceedings of the AAAI conference on artificial intelligence* (pp.1–7).

Song, L., Cao, J., Song, L., Hu, Y., He, R. (2019). Geometry-aware face completion and editing. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2506–2513).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Tang, Y., Wang, J., Wang, X., Gao, B., Dellandréa, E., Gaizauskas, R., & Chen, L. (2017). Visual and semantic knowledge transfer for large scale semi-supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(12), 3045–3058.

Terhörst, P., Fährmann, D., Kolf J. N., Damer, N., Kirchbuchner, F., Kuijper, A. (2020). Maad-face: A massively annotated attribute dataset for face images. arXiv preprint arXiv:2012.01030.

Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X. (2020). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 12275–12284).

Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T. S. (2018). Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7268–7277).

Woo, S., Park, J., Lee, J. Y., Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European conference on computer vision* (pp. 3–19).

Wu, H., & Prasad, S. (2017). Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing, 27*(3), 1259–1270.

Yan, Y., Xu, Y., Xue, J.-H., Lu, Y., Wang, H., Zhu, W. (2022). Drop loss for person attribute recognition with imbalanced noisy-labeled samples. *IEEE Transactions on Cybernetics*.

Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., & Sang, N. (2021). BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision, 129*(11), 3051–3068.

Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L. (2019). S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 1476–1485).

Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D. (2017.a) mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L. (2014). Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1637–1644).

Zhang, S., He, R., Sun, Z., & Tan, T. (2017). Demeshnet: Blind face inpainting for deep meshface verification. *IEEE Transactions on Information Forensics and Security, 13*(3), 637–647.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision* (pp. 286–301).

Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C. C., Lin, D., Jia, J. (2018). Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European conference on computer vision* (pp. 267–283).

Zhao, X., Li, H., Shen, X., Liang, X., Wu, Y. (2018). A modulation module for multi-task learning with applications in image retrieval. In *Proceedings of the European conference on computer vision* (pp. 401–416).

Zheng, X., Guo, Y., Huang, H., Li, Y., & He, R. (2020). A survey of deep facial attribute analysis. *International Journal of Computer Vision, 128*(8), 2002–2034.