# Partially Shared Multi-Task Convolutional Neural Network with Local Constraint for Face Attribute Learning

Jiajiong Cao, Yingming Li*, and Zhongfei Zhang
College of Information Science & Electronic Engineering
Zhejiang University, China
{jiajiong, yingming, zhongfei}@zju.edu.cn

## Abstract

*In this paper, we study the face attribute learning problem by considering the identity information and attribute relationships simultaneously. In particular, we first introduce a Partially Shared Multi-task Convolutional Neural Network (PS-MCNN), in which four Task Specific Networks (TSNets) and one Shared Network (SNet) are connected by Partially Shared (PS) structures to learn better shared and task specific representations. To utilize identity information to further boost the performance, we introduce a local learning constraint which minimizes the difference between the representations of each sample and its local geometric neighbours with the same identity. Consequently, we present a local constraint regularized multi-task network, called Partially Shared Multi-task Convolutional Neural Network with Local Constraint (PS-MCNN-LC), where PS structure and local constraint are integrated together to help the framework learn better attribute representations. The experimental results on CelebA and LFWA demonstrate the promise of the proposed methods.*

## 1. Introduction

Face attribute learning [26, 1, 2, 30, 37, 8, 18, 9] has attracted much attention in many real-world applications such as face identification and verification [33, 23, 32, 34, 27]. It aims to learn mid-level representations as the abstraction between the low-level features and the high-level labels. However, large-scale face attribute learning is still a very challenging problem as the faces captured in the wild are usually influenced by the variations of the factors such as illumination, pose, and expression.

Motivated by the success of convolutional neural network (CNN) [16, 35, 31, 6, 24, 10, 11], the deep CNN representations have been widely employed for face attribute learning. For example, Razavian *et al.* exploit a face recognition network to extract facial features and then train SVMs for attribute classification [30]. Further, Hand and Chellappa consider exploiting the attribute correlations to construct reliable deep architecture for attribute classification [9]. In particular, Multi-task deep CNN (MCNN) is introduced by sharing the lower layers of network for all the attributes and sharing the higher layers for closely related attributes through a split structure. Based on the assumption that many attributes are strongly correlated, MCNN divides all the 40 attributes into nine attribute groups so that similar attributes are within a group and high-level features are independently learned for each group.

Although MCNN has become the state-of-the-art performance by exploiting the complementary information for attributes, the interactions among different groups are restricted since they are independent after the split. The shared information vanishing among groups emerges when it reaches the high-level layers of MCNN. Consequently, it is difficult for attribute groups to effectively utilize the attribute relatedness from the beginning of the network to its end to boost the overall performance. From a perspective of multi-task learning, it is necessary to learn shared features among groups (tasks) throughout the network to model the sophisticated attribute relationships. Further, MCNN ignores the samples' identity information which implies their inter-dependence and encodes local geometric structure. As such local structure usually helps feature learning, the existing attribute learning methods that ignore the identity information may not be appropriate.

In this paper we investigate the multi-task face attribute learning problem with a new perspective of considering identity information and attribute relationships simultaneously. We hypothesize that combining identity information and task relationship modeling enables us to develop more accurate multi-task attribute learning algorithms. Our hypothesis is based on two insights: (1) efficient interactions among different attribute groups (tasks) help lead to more accurate attribute relationship modeling; and (2) informative identity labels further help boost the performance by

*Corresponding author

modeling local geometric structure for attribute learning.

First, we propose a novel Partially Shared Multi-task Convolutional Neural Network (PS-MCNN) in which task relation is captured by a Shared Network (SNet) and variability across different tasks is captured by Task Specific Networks (TSNets). Similar to MCNN, all the attributes are split into several groups according to spatial information and then the classification learning of each group can be regarded as an individual task. The key idea of PS-MCNN lies in sharing a common network for all the groups to learn shared features, and constructing group specific network for each group from the beginning of the architecture to its end to learn task specific features, which makes it different from the existing MCNN that learns shared features at low-level layers while learns task specific features at high-level ones. This way of multi-task learning helps effectively exploit the complementary information from different tasks while maximally preserving the specific information of specific tasks. Figure 1 gives two structure examples to show respective strategies for PS-MCNN and MCNN.

Furthermore, we incorporate identity information into PS-MCNN to improve the performance of multi-task face attribute learning. To achieve this goal, a simple approach is to treat identity as an additional face attribute. However, face identity and attribute information have different properties. For example, an attribute is an actual evidence that represents a specific identity and usually acts as a kind of mid-level description. Thus, this approach ignores such characteristic and fails to combine attribute and identity information seamlessly.

To incorporate the identity information into multi-task attribute learning effectively, we introduce local learning constraint which minimizes the difference between the representations of each sample and its local geometric neighbours with the same identity. Such constraint helps samples with the same identity have more attribute similarity. Consequently, we propose a Local Constraint Loss (LCLoss) and combine it with PS-MCNN to obtain Partially Shared Multi-task Convolutional Neural Network with Local Constraint (PS-MCNN-LC).

We conduct extensive evaluations to investigate the performances of PS-MCNN and PS-MCNN-LC. The experimental results on CelebA and LFWA show the superior performances of both networks over the competing models.

# 2. Partially Shared Network for Face Attribute Learning

## 2.1. Split structure

Multi-task learning [29, 19, 13, 20, 25, 38, 21, 5] aims at learning multiple tasks simultaneously. It assumes that related tasks interact with each other so that they can make usage of complementary information from each other to boost the overall performance. This is especially true for face attribute learning with deep architectures as there is a hierarchy of attributes. A common strategy of multi-task attribute learning is to use the split structure as shown in the right of Figure 1, which shares low-level features for all the tasks while high-level features are specifically learned for each task after the bifurcation. Though this structure has been used widely in deep learning methods [9, 21], it has two main drawbacks: (1) it needs intensive experiments to find the optimal split point, especially for deep networks; and (2) more importantly, interactions among different tasks at high-level layers are restricted since there are no shared layers after the bifurcation.

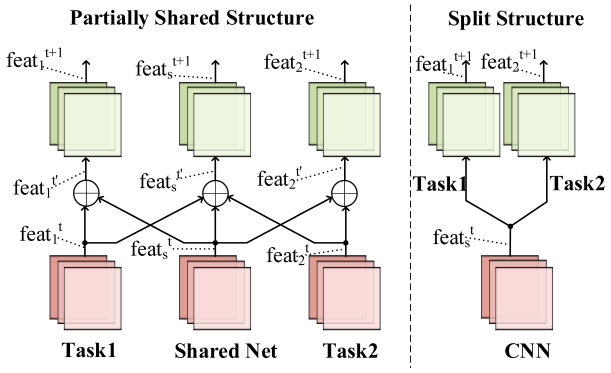


Figure 1. Partially Shared structure and split structure.

## 2.2. Partially shared structure

To overcome the above limitations of the split structure, we propose a novel network structure called Partially Shared (PS) structure for multi-task learning. PS structure consists of two types of networks: Task Specific Network (TSNet) and Shared Network (SNet). TSNet focuses on learning features for a specific task, while SNet learns informative representations which are shared for each task. SNet interacts with each TSNet through a simple connectivity pattern: to ensure maximum information flow between layers of SNet and TSNet, we connect the layers of TSNet with the layers of SNet. Each layer of SNet obtains additional inputs from the previous layers of TSNet and passes on its own feature-maps to the next layers of shared and task specific networks. As shown in Figure 1, the information flow can be formulated as follows:

$$
\begin{aligned}
feat_1^{t'} &= [feat_1^t, feat_s^t], feat_1^{t+1} = H^{t+1}(feat_1^{t'}), \\
feat_2^{t'} &= [feat_2^t, feat_s^t], feat_2^{t+1} = H^{t+1}(feat_2^{t'}), \\
feat_s^{t'} &= [feat_1^t, feat_2^t, feat_s^t], feat_s^{t+1} = H^{t+1}(feat_s^{t'}),
\end{aligned}
$$

where $H^{t+1}(\cdot)$ represents a sequence of operations: Conv-BN-ReLU. Different from the split structure, where

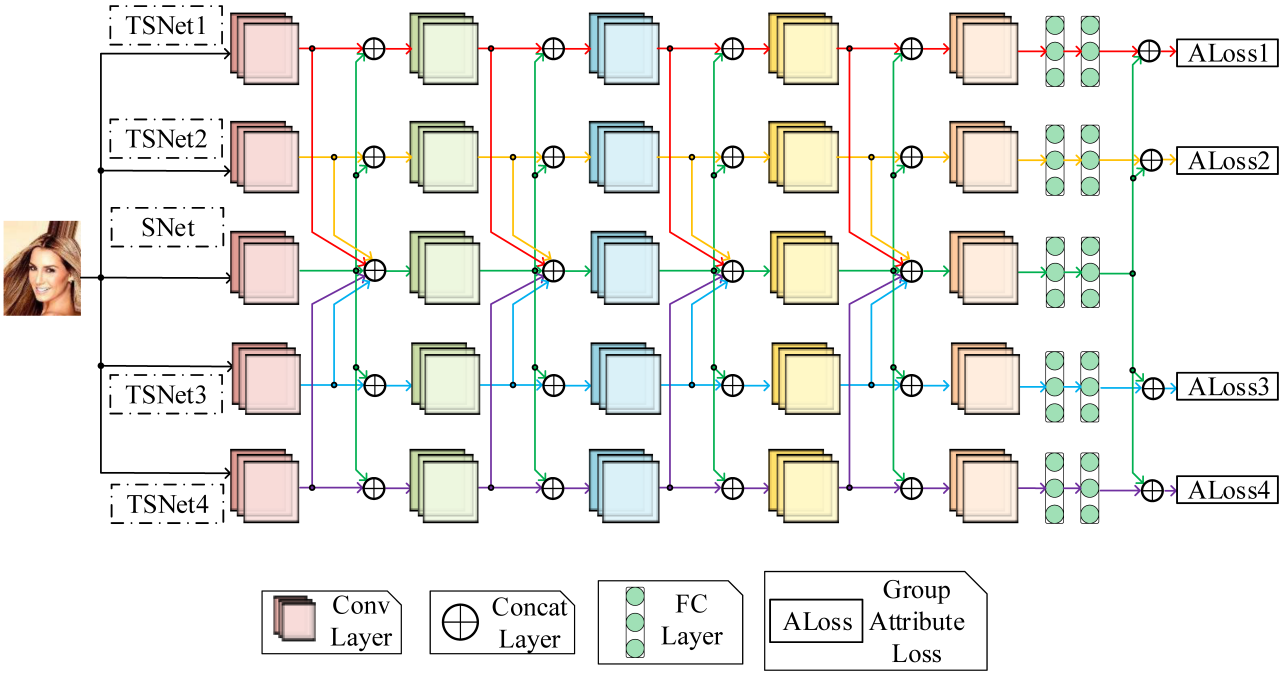


Figure 2. Architecture of PS-MCNN.

$feat_i^{t+1}, i \in \{1, 2\}$ is computed based on $feat_s^t$, PS structure learns $feat_i^{t+1}$ based on both task specific features $feat_i^t$ and shared features $feat_s^t$. Besides, the shared features, $feat_s^{t+1}$, is directly related to all the features at depth $t$, which enables SNet to extract informatively shared features.

### 2.3. Partially shared multi-task convolutional neural network

Based on the proposed PS structure, we introduce a novel Partially Shared Multi-task Convolutional Neural Network (PS-MCNN) for face attribute learning. Similar to MCNN, we split all the 40 attributes into four attribute groups including Upper, Middle, Lower, and Whole Image according to their corresponding locations. Then the attributes classification of each group can be considered as an individual attribute learning task. The detailed group configuration is listed bellow.

**Upper Group:** *Arched Eyebrows, Bags Under Eyes, Bald, Bangs, Black Hair, Blond Hair, Brown Hair, Bushy Eyebrows, Eyeglasses, Gray Hair, Narrow Eyes, Receding Hairline, Wearing Hat.*

**Middle Group:** *Big Nose, High Cheekbones, Pointy Nose, Rosy Cheeks, Sideburns, Wearing Earrings.*

**Lower Group:** *Big Lips, Double Chin, Goatee, Mustache, Mouth Slightly Open, No Beard, Wearing Lipstick, Wearing Necklace, Wearing Necktie.*

**Whole Image Group:** *5 o'Clock Shadow, Attractive, Blurry, Chubby, Heavy Makeup, Male, Oval Face, Pale Skin, Straight Hair, Smiling, Wavy Hair, Young.*

Figure 2 shows the architecture of PS-MCNN, where four TSNets are exploited for four respective attribute groups to learn task specific features and one SNet is constructed for shared representation learning. Meanwhile, TSNets and SNet are connected via the PS structures at each depth for better interactions between different tasks.

### 2.4. Design decisions

To fully utilize the effectiveness of PS-MCNN when applying it to a specific task such as face attribute learning, there are two important factors that need further exploration.

**Network Initializations:** Since PS-MCNN consists of five individual networks, as shown in Figure 2, there are many available initialization policies. For example, we can choose the same initialization for TSNets and SNet or we can train all the networks from scratch. It needs experiments to make the optimal choice for the face attribute learning scenario.

**The Number of Channels of SNet:** The number of channels of SNet determines the fraction of shared representations for each TSNet. There are two extreme situations. When the number of shared channels becomes zero, SNet disappears and PS-MCNN degenerates to four independent TSNets. On the other hand, when the fraction of shared channels becomes one, all the features are shared

features and PS-MCNN only includes an SNet.

## 3. Partially Shared Network with Local Constraint for Face Attribute Learning

Since the identity information usually exists in the training data in face attribute learning, we take advantage of this information to improve the attribute classification performance. In this section, we first introduce local learning constraint. Then we combine it with PS-MCNN to solve the problem of face attribute learning.

### 3.1. Local learning constraint

According to [7, 3], learning a good representation in a global way might not be a good strategy because it usually fails to consider the local geometric structure in the data. Therefore, if we can encode the geometric structure in a simple way and fuse it into the process of representation learning, then it would be more effective to find the appropriate embedding functions. In fact, it is common for face verification methods [33, 4, 28, 36, 17] to adopt a local constraint loss such as Contrastive Loss [33] or Triple Loss [28] besides the global classification loss. In addition, we also find that face attribute labels are highly similar for the same identity. It implies that attributes have certain local geometric structure through identity correlation, which is complementary to the traditional global attribute learning.

To achieve the goal of local learning, we propose a local learning constraint by first defining samples with the same identity as the geometric neighbours and then constraining the attribute representations of geometric neighbors to be close to each other. Consequently, samples with the same identity will have more attribute similarity. We formulate the loss function of the local learning constraint as bellow, referred to as LCLoss:

$$LCLoss = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} w_{i,j} ||feat_{si}^{t} - feat_{sj}^{t}||_2^2,$$

$$w_{i,j} = \begin{cases} 1, \text{if sample } i \text{ and sample } j \text{ have the same identity} \\ 0, \text{otherwise} \end{cases}.$$

### 3.2. Partially shared multi-task convolutional neural network with local constraint

We combine the local learning constraint and PS-MCNN as shown in Figure 3 and compute LCLoss on the concatenated features of the final layers to constrain all the four groups. Consequently, a Partially Shared Multi-task Convolutional Neural Network with Local Constraint (PS-MCNN-LC) is proposed and its objective function can be formulated as follows:

$$Obj = \sum_{i=1}^{4} ALoss_i + \lambda \times LCLoss,$$

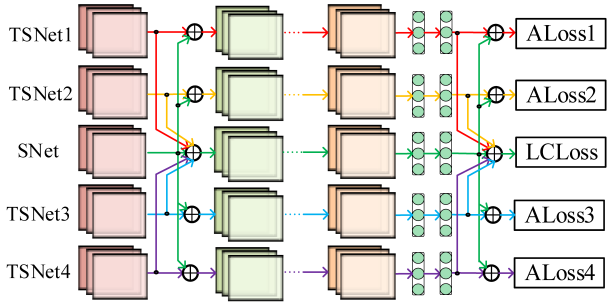


Figure 3. The architecture of PS-MCNN-LC, which shares the architecture of PS-MCNN but is different at the loss function.

where $ALoss_i$ is the attribute loss of the $i$-th TSNet and $\lambda$ is the weight of LCLoss. In the formulation, LCLoss plays as a local learning regularization and helps PS-MCNN-LC model local geometric structure. The regularization parameter $\lambda$ controls the strength of the local learning constraint. For example, when $\lambda$ becomes nearly zero, the LCLoss has limited impact on attribute learning. On the other hand, if $\lambda$ becomes too large, PS-MCNN-LC pays more attention to identity learning and the attribute prediction performance may be harmed.

## 4. Ablative Analysis

We discuss design decisions in Sec. 4.1 and Sec. 4.2 to fix the model hyper parameters. Then, based on the fixed model, the effectiveness of PS architecture is verified in Sec. 4.3 from two aspects. Finally complexity analysis are performed in Sec. 4.4. The experimental details of ablation studies are listed as bellow.

**Datasets:** CelebA dataset [18], is split into training set ($160k$ images), validation set ($20k$ images) and test set ($20k$ images) according to the suggested configuration. For a better comparison, we also test the performance on LFWA dataset [12].

**Network Architecture:** Both PS-MCNN and PS-MCNN-LC consist of 5 networks with the same architecture including four TSNets and one SNet. The detailed parameters are listed in Table 1. Note that Table 1 shows the number of channels of TSNets, while the number of channels of SNet will be discussed independently in Sec. 4.2

### 4.1. Network initializations

We have tried three configurations for network initializations: (1) SNet and TSNets are pretrained on face recognition task on the training set of CelebA, referred to as FR; (2) SNet and TSNets are trained from scratch; (3) SNet is pretrained on face recognition task while each TSNet is pretrained on the attribute prediction task within the group on CelebA, referred to as TS. As illustrated in Table 2, configuration (3) performs the best, since FR initialization

Table 1. Network architecture of TNets in PS-MCNN. SNet also shares this structure but differs on the number of the channels. Note that batch normalization [14] and ReLU [22] are adopted after each convolution layer and fully-connected layer.

| Layer Name | Output Size | Layer Parameters |
|---|---|---|
| Conv1 | $192 \times 160$ | $3 \times 3 \times 32$ |
| Pooling1 | $96 \times 80$ | $2 \times 2$ MAX |
| Conv2 | $96 \times 80$ | $3 \times 3 \times 64$ |
| Pooling2 | $48 \times 40$ | $2 \times 2$ MAX |
| Conv3 | $48 \times 40$ | $3 \times 3 \times 128$ |
| Pooling3 | $24 \times 20$ | $2 \times 2$ MAX |
| Conv4 | $24 \times 20$ | $3 \times 3 \times 256$ |
| Pooling4 | $12 \times 10$ | $2 \times 2$ MAX |
| Conv5 | $12 \times 10$ | $3 \times 3 \times 128$ |
| Pooling5 | $6 \times 5$ | $2 \times 2$ MAX |
| fc6 | – | 512 |
| fc7 | – | 512 |

Table 2. Performances of PS-MCNN under different network initializations.

| Init. \ Groups | Upp. | Mid. | Low. | Who. | Ave. |
|---|---|---|---|---|---|
| SNet(FR) +TSNets(FR) | 6.21 | 10.15 | 7.04 | 9.22 | 7.89 |
| Scratch | 6.02 | 9.99 | **6.60** | 8.72 | 7.62 |
| SNet(FR) +TSNets(TS) | **5.98** | **9.87** | 6.64 | **8.64** | **7.51** |

helps SNet extract better shared representations while TS initialization encourages TSNets to preserve more task specific features. However, when FR initializations are applied to both SNet and TSNets, TSNets pay more attention to identity-related areas while the attributes corresponding to the other areas are not well learned [30].

## 4.2. Number of channels of SNet

We have tried six different numbers of channels as shown in Table 3. There are two extreme situations: (1) number of channels of SNet is zero and PS-MCNN degenerates to four TSNets, which is called Inde. Group; and (2) there are only shared representations but no task specific features resulting in PS-MCNN to be a single SNet, which is referred to as SNet in Table 3. The performances of the two extreme situations are significantly worse than those of regular PS-MCNNs. Since the former has no shared features and the latter has no task specific features, both of the situations are not suitable for modeling attribute relationships.

Table 3. Performances of PS-MCNNs with SNets of different numbers of channels (w.r.t. the number of TSNets in Table 1).

| Num. \ Groups | Upp. | Mid. | Low. | Who. | Ave. |
|---|---|---|---|---|---|
| Inde. Group | 7.23 | 11.42 | 7.83 | 10.08 | 8.84 |
| 1/8 | 6.28 | 10.00 | 7.14 | 9.01 | 7.91 |
| 1/4 | **6.16** | 10.00 | **7.01** | 8.87 | **7.77** |
| 1/2 | 6.22 | 9.96 | 7.07 | **8.80** | 7.81 |
| 1 | 6.24 | **9.95** | 7.02 | 8.84 | 7.82 |
| SNet | 7.51 | 11.73 | 7.99 | 10.25 | 9.07 |

## 4.3. Effectiveness of PS Architecture

To eliminate the influence of other factors of the network (e.g., attribute grouping), we train a PS-MCNN and an Inde. Group (4 independent TSNets) on 4 closely related attributes (Chubby, Double Chine, Mustache, No Beard), each of which corresponds to a TSNet. We refer them as Inde. Group (single) and PS-MCNN (single) respectively in Table 4. By comparing the error rates of Inde. Group (single) and PS-MCNN (single), we see that without attribute grouping, the performance gain is still significant. Therefore, the main performance gain does come from the proposed partially shared structure instead of other factors.

Table 4. Error rates on the 4 closely related attributes.

| Method | Chubby | Double Chin | Mustache | No Beard |
|---|---|---|---|---|
| Inde. Group (single) | 4.58 | 3.82 | 3.28 | 3.95 |
| PS-MCNN (single) | 2.99 | 2.47 | 2.11 | 2.77 |

Further, PS architecture requires SNet to connect to TSNets at every layer so that it enhances the information exchange between attribute groups as much as possible. We also investigate other possible designs to reduce the parameters of SNet. In particular, we provide the experimental results on PS-MCNN with a reduced SNet in Table 5, which fuses and broadcasts feature from TSNets every two layers. As we see from Table 5, the overall performance of such designed model drops due to the reduction of information exchange between different attribute groups. Therefore, the current design of SNet is suitable for releasing the potential ability of PS structure.

Table 5. Error rates on PS-MCNNs with different designs of SNets.

| Method | Upp. | Mid. | Low. | Who. | Ave. |
|---|---|---|---|---|---|
| Original | 5.98 | 9.87 | 6.64 | 8.64 | 7.51 |
| Reduced | 6.15 | 10.21 | 6.97 | 9.02 | 7.80 |

### 4.4. Complexity Analysis

We analyze the computational cost of MCNN [9] and PS-MCNN with the floating-point operations (FLOPs) in the number of multiply-adds, which is commonly used in previous works [10, 11]. Though the actual runtime may be influenced by other factors including coding quality and GPU bandwidth, such cost analysis provides an estimation of the speed upper bound. As shown in Table 6, PS-MCNN consumes about $22\%$ less FLOPs than MCNN and has only $1/20$ parameters of MCNN. Therefore, PS-MCNN is a much more practical network. We clarify the design differences between PS-MCNN and MCNN below.

Table 6. Complexity analysis on MCNN and PS-MCNN. For the analysis of MCNN, we refer to the setups of [9].

| Method | Input Size $w \times h$ | No. Param. $\times 10^6$ | No. FLOPs $\times 10^9$ | Trunk Depth |
|---|---|---|---|---|
| MCNN | $227 \times 227$ | 320 | 8.6 | 5 |
| PS-MCNN | $192 \times 160$ | 16 | 6.7 | 7 |

MCNN adopts $7 \times 7$ and $5 \times 5$ kernels for the first two convolutional layers individually, which leads to a large number of FLOPs. It directly feeds $14 \times 14 \times 500$ feature maps to the fully-connected layer for each group leading to a large number of parameters. On the contrary, PS-MCNN uses $3 \times 3$ kernels throughout the network and inputs $6 \times 5 \times 128$ feature maps to the fully-connected layers. Therefore, though PS-MCNN has $5$ subnetworks, each is carefully designed to reduce the numbers of parameters and FLOPs while maximizing its learning capacity by adding more layers.

## 5. Experiments

We now present experiments with PS-MCNN and PS-MCNN-LC for face attribute learning. We use the same datasets and network architectures as discussed in Sec. 4. Detailed configurations of PS-MCNN and PS-MCNN-LC are described as follows.

**Training Setups:** We use aligned images of CelebA dataset with a 64 batch size for training. We set the initial learning rate as 0.001 and decrease it two times during training. All the networks are trained on the Caffe [15] platform with the standard mini-batch SGD method [16].

In particular, we first train SNet with only identity loss on CelebA. Then, we remove the last fully connected layer of SNet and link it to the $4$ TSNets. Finally, the whole PS-MCNN is trained with attribute classification loss.

**Network Configurations:** According to Table 2 and Table 3, we choose 1/4 channels of TSNets for SNet while initializing the four TSNets with TS initialization configuration and SNet with FR initialization configuration.

### 5.1. Baselines

We compare PS-MCNN and PS-MCNN-LC with three strong baselines. LNets+ANet [18] is the first deep learning method for face attribute learning. MCNN [9] is the best state-of-the-art method, which adopts a split architecture in the network. For better comparison, we also report performance of four independent TSNets, referred to as Inde. Group. Results are reported in Table 7.

### 5.2. Comparisons with baselines

According to Table 7, PS-MCNN achieves an average error rate of **7.78%**, better than all the competing methods. In particular, compared with MCNN, the best state-of-the-art method, the error rate on CelebA is reduced by **11.0** percent. After adopting the LCLoss, the error rate is further reduced by **19.8** percent. The performance on LFWA is also significantly improved.

**Comparison with LNets+ANet:** PS-MCNN outperforms LNets+ANet by more than **30** percent. We attribute it to the naive architecture of ANet, which shares all features until the last fully-connected layer. Therefore, ANet pays very limited attention to task specific feature learning.

**Comparison with MCNN:** MCNN uses the split structure, where lower layers are shared with all the attributes while higher layers are specifically learned for groups. Therefore, the lower layers only learn shared features while the higher layers only learn task specific features. On the contrary, PS-MCNN aims to learn shared features in all layers while preserving the task specific features as well. As a result, PS-MCNN is able to model better attribute relationships to improve the classification performance.

**Comparison with Inde. Group:** We compare Inde. Group with PS-MCNN to verify the necessity of SNet. As illustrated in Table 7, the error rate of Inde. Group is 1.07% and 1.83% higher on average than those of PS-MCNN and PS-MCNN-LC, respectively. We attribute this to the absence of shared features. Learning complementarily shared representations for all the tasks is essential to improving the generalization performance for multi-task learning. In particular, both PS-MCNN and PS-MCNN-LC construct a single network SNet to achieve this goal. By learning shared features throughout the networks, PS-MCNN encourages the four tasks to boost the performance by utilizing the features from each other.

To better understand the influence of the number of the shared feature channels, we show feature maps of the conv2 layer of two samples in Table 8. With the SNet, Upper Group puts more weights on the other areas besides the upper region so as Lower Group, which helps each task utilize information from the other groups to model better attribute relationships. For example, *Male* in Whole Image Group and *No Beard* in Lower group are strongly correlated. The SNet is able to help model these attribute relationships to

Table 7. Prediction error rates on CelebA and LFWA of different methods.

| Attr. \ Meth. | CelebA | | | | | LFWA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | LNets +ANet | MCNN | Inde. Group | PS -MCNN | PS-MCNN -LC | LNets +ANet | MCNN | Inde. Group | PS -MCNN | PS-MCNN -LC |
| 5 Shadow | 9 | 5.59 | 5.52 | 4.17 | **3.40** | 16 | 22.30 | 13.42 | **20.41** | 21.83 |
| Arch. Eyebrows | 21 | 16.50 | 16.33 | 15.39 | **14.23** | 18 | **14.10** | 16.92 | 16.51 | 16.47 |
| Attractive | 19 | 17.06 | 17.42 | 16.59 | **15.61** | 17 | 19.56 | **16.30** | 16.67 | 18.16 |
| Bags Un. Eyes | 21 | 15.02 | 14.94 | 13.85 | **12.71** | 17 | 16.49 | **11.60** | 14.77 | 13.26 |
| Bald | 2 | 1.13 | 1.10 | 0.67 | **0.59** | 12 | 8.01 | 11.89 | 7.61 | **7.40** |
| Bangs | 5 | 3.96 | 4.06 | 2.90 | **2.00** | 12 | 10.01 | 10.85 | 9.24 | **8.55** |
| Big Lips | 32 | 29.80 | 28.90 | 28.53 | **26.87** | 25 | 20.79 | 22.23 | 20.00 | **17.30** |
| Big Nose | 22 | 15.50 | 15.51 | 14.42 | **13.60** | 19 | 15.33 | 14.92 | 13.85 | **13.52** |
| Black Hair | 12 | 10.13 | 10.41 | 9.54 | **8.34** | 10 | 7.65 | 10.46 | 7.24 | **7.04** |
| Blond Hair | 5 | 4.03 | 4.61 | 2.97 | **2.07** | 3 | 3.55 | 3.76 | 3.60 | **1.49** |
| Blurry | 6 | 3.92 | 4.16 | 2.77 | **2.00** | 26 | 14.70 | 14.92 | 13.49 | **12.80** |
| Brown Hair | 20 | 11.01 | 11.24 | 9.82 | **8.97** | 23 | 19.06 | 18.76 | 17.55 | **18.13** |
| Bushy Eyebrows | 10 | 7.20 | 8.32 | 6.40 | **5.49** | 18 | 14.89 | 14.92 | **13.73** | 14.28 |
| Chubby | 8 | 4.34 | 4.33 | 2.78 | **2.34** | 27 | 23.10 | 22.08 | **21.27** | 21.89 |
| Double Chin | 8 | 3.59 | 3.49 | 2.26 | **1.71** | 22 | 18.83 | 18.20 | 17.38 | **13.30** |
| Eyeglasses | 1 | 0.37 | 0.44 | 0.21 | **0.15** | **5** | 8.78 | 7.43 | 7.44 | 7.22 |
| Goatee | 5 | 2.70 | 2.77 | 2.34 | **2.26** | 22 | 17.48 | 18.86 | 16.11 | **15.89** |
| Gray Hair | 3 | 1.80 | 2.23 | 1.56 | **1.34** | 16 | 10.96 | 14.91 | 10.15 | **8.96** |
| Heavy Makeup | 10 | 8.66 | 8.40 | 7.26 | **6.69** | 5 | 4.16 | 6.30 | 3.69 | **3.40** |
| H. Cheekbones | 13 | 12.45 | 12.64 | 11.43 | **10.50** | 13 | 11.75 | 14.15 | **10.94** | 11.23 |
| Male | 2 | 1.84 | 1.86 | 1.28 | **1.19** | 6 | 6.34 | 5.79 | 5.78 | **4.82** |
| Mouth S. O. | 8 | 6.26 | 6.23 | 5.02 | **4.01** | 18 | 16.53 | 16.57 | **15.29** | 15.40 |
| Mustache | 5 | 3.07 | 3.07 | 1.77 | **1.44** | 8 | 6.47 | 6.98 | 6.02 | **5.53** |
| Narrow Eyes | 19 | 12.84 | 12.81 | 11.86 | **10.93** | 19 | 17.63 | 16.83 | **16.16** | 16.49 |
| No Beard | 5 | 3.89 | 3.65 | 2.36 | **1.97** | 21 | 17.87 | 18.09 | **16.38** | 17.99 |
| Oval Face | 34 | 24.19 | 24.17 | 23.37 | **22.57** | 26 | 22.62 | 22.49 | **21.25** | 22.10 |
| Pale Skin | 8 | 2.98 | 2.92 | 1.76 | **1.16** | 6 | 6.59 | 11.68 | 6.14 | **5.03** |
| Pointy Nose | 28 | 22.53 | 22.81 | 21.90 | **20.68** | 22 | 12.48 | 15.06 | **11.57** | 12.48 |
| Receed. Hairline | 11 | 6.19 | 6.41 | 4.62 | **4.15** | 15 | 13.74 | 13.35 | 12.70 | **12.50** |
| Rosy Cheeks | 10 | 4.87 | 5.38 | 3.55 | **3.08** | 22 | 12.48 | 16.30 | **10.85** | 11.19 |
| Sideburns | 4 | 2.18 | 2.23 | 1.87 | **1.78** | 23 | 17.27 | 21.63 | 15.89 | **15.58** |
| Smiling | 8 | 7.34 | 7.65 | 6.27 | **5.15** | 9 | 8.25 | 8.98 | 7.39 | **7.30** |
| Straight Hair | 27 | 16.61 | 16.59 | 15.44 | **14.04** | 24 | 21.28 | 20.73 | **19.70** | 20.35 |
| Wavy Hair | 20 | 16.08 | 16.78 | 15.12 | **13.61** | 24 | 18.04 | 22.49 | 17.09 | **16.65** |
| Wear. Earings | 18 | 9.68 | 9.94 | 8.14 | **7.34** | 6 | 5.29 | 6.38 | 4.62 | **4.46** |
| Wear. Hat | 1 | 0.96 | 0.93 | 0.67 | **0.57** | 12 | 9.80 | 9.47 | 9.13 | **8.79** |
| Wear. Lipstick | 7 | 6.05 | 6.01 | 5.10 | **4.30** | 5 | 5.11 | 7.25 | 4.50 | **4.30** |
| Wear. Necklace | 29 | 13.18 | 12.97 | 12.22 | **11.02** | 12 | 10.34 | 10.43 | 9.22 | **9.08** |
| Wear. Necktie | 7 | 5.47 | 3.42 | 2.26 | **1.48** | 21 | 19.50 | 19.36 | 17.94 | **17.82** |
| Young | 13 | 11.52 | 11.20 | 10.51 | **9.46** | 14 | 14.63 | 14.39 | 13.23 | **13.12** |
| Ave. | 13 | 8.75 | 8.85 | 7.78 $\pm$ 0.15 | **7.02** $\pm$ **0.25** | 16 | 13.73 | 14.11 | 12.88 $\pm$ 0.18 | **12.64** $\pm$ **0.23** |

get a higher performance.

**The Effectiveness of LCLoss:** PS-MCNN-LC outperforms PS-MCNN significantly on CelebA dataset because LCLoss captures the local geometric structure based on the identity information. The parameter $\lambda$ reflects the strength of local learning constraint. Different $\lambda$ values can lead to dramatically different performances of PS-MCNN-LC according to Table 9. In particular, when $\lambda$ becomes nearly zero, the LCLoss has limited impact on attribute learning so that PS-MCNN-LC performs similarly to PS-MCNN. On the other hand, if $\lambda$ becomes too large, PS-MCNN-LC pays more attention to identity learning and the attribute prediction performance is inevitably harmed.

Overall, the current setup of identity constraint is a sim-

Table 8. Feature maps with and without SNet. *Upper Inde.* or *Lower Inde.* refers to the feature maps without SNet while *Upper* or *Lower* refers to the results with an SNet having $1/4$ channels of a TSNet.

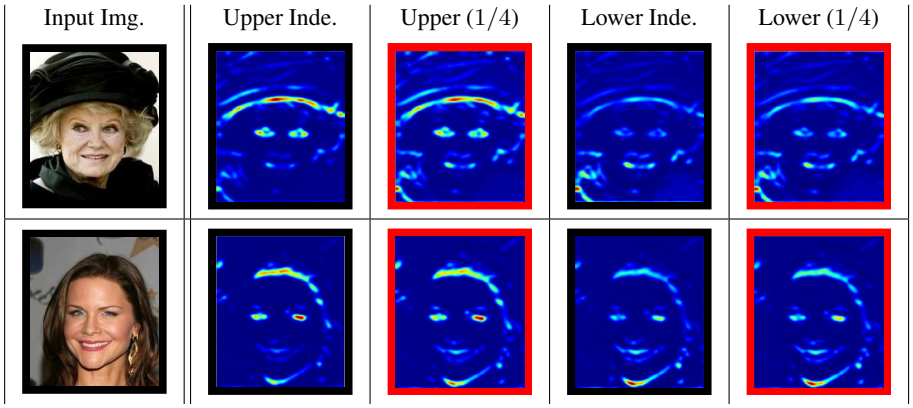


Table 9. PS-MCNN performances under different $\lambda$.

| $\lambda$ \ Groups | Upp. | Mid. | Low. | Who. | Ave. |
|---|---|---|---|---|---|
| 0 | 6.16 | 10.00 | 7.01 | 8.87 | 7.77 |
| $10^{-4}$ | 5.78 | 9.60 | 6.46 | 8.49 | 7.39 |
| $10^{-3}$ | 5.52 | **9.42** | **6.12** | 8.05 | **7.00** |
| $10^{-2}$ | **5.45** | 9.49 | 6.19 | **7.95** | 7.04 |
| $10^{-1}$ | 6.12 | 9.96 | 7.01 | 8.92 | 7.81 |

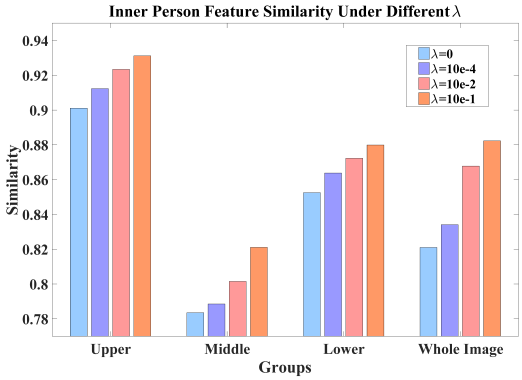


Figure 4. Inner person feature similarity under different $\lambda$ of four groups. The larger the $\lambda$, the more similar the features of the same identity.

ple and effective way of modeling the attribute similarity of the samples with the same identity. Further, such universal identity constraint may be not effective all the time and more flexible schemes need to be investigated.

To better illustrate the influences of LCLoss, we show group feature cosine similarities on the fc7 layer for $400k$ same identity image pairs under different $\lambda$ in Figure 4. Though the average similarities of the four groups vary, the similarities increase as $\lambda$ increases. However, the same person may have several different attribute labels. When a large $\lambda$ forces the features of different images of the same person becoming highly similar, PS-MCNN-LC fails to learn the attribute label variety for the same identity, which harms the final performance as indicated in Table 9. Therefore, it is a tradeoff choosing an appropriate $\lambda$ when considering both attribute label consistency and attribute label variety for the different samples with the same identity.

## 6. Conclusion

In this paper, we investigate the face attribute learning problem by considering the identity information and attribute relationship modeling simultaneously. In particular, we first introduce a PS-MCNN, in which four TSNets and one SNet are connected by PS structures to learn better shared and task specific features. To utilize identity information, we introduce an LCLoss which minimizes the difference between the features of each sample and its local geometric neighbours with the same identity. Consequently, we present a PS-MCNN-LC, where the PS structure and local constraint are integrated together to help the framework learn better attribute features. The experimental results on CelebA and LFWA demonstrate the promise of the proposed methods.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 61702448, 61672456), the Key R&D Program of Zhejiang Province (No. 2018C03042), and the Fundamental Research Funds for the Central Universities (No. 2017QNA5008, 2017FZA5007). We would like to thank the reviewers for their constructive comments.

## References


[1] T. Berg and P. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 955–962, 2013.

[2] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1543–1550. IEEE, 2011.

[3] D. Cai, X. He, X. Wu, and J. Han. Non-negative matrix factorization on manifold. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pages 63–72, 2008.

[4] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. *CoRR*, abs/1704.01719, 2017.

[5] X. Chu, W. Ouyang, W. Yang, and X. Wang. Multi-task recurrent neural network for immediacy prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3352–3360, 2015.

[6] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[7] Q. Gu and J. Zhou. Local learning regularized nonnegative matrix factorization. In *International Jont Conference on Artifical Intelligence*, pages 1046–1051, 2009.

[8] M. Günther, A. Rozsa, and T. E. Boult. AFFACT - alignment free facial attribute classification technique. *CoRR*, abs/1611.06158, 2016.

[9] E. M. Hand and R. Chellappa. Attributes for improved attributes: A multi-task network for attribute classification. *arXiv preprint arXiv:1604.07360*, 2016.

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

[11] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Technical Report 07-49, University of Massachusetts, Amherst*, 2007.

[13] W. Huang, G. Song, H. Hong, and K. Xie. Deep architecture for traffic flow prediction: deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems*, 15(5):2191–2201, 2014.

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[18] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. pages 3730–3738, 2015.

[19] Y. Luo, D. Tao, B. Geng, C. Xu, and S. J. Maybank. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE Transactions on Image Processing*, 22(2):523–536, 2013.

[20] A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 343–351, 2013.

[21] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016.

[22] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[23] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[25] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil. Multilinear multitask learning. In *International Conference on Machine Learning*, pages 1444–1452, 2013.

[26] E. M. Rudd, M. Günther, and T. E. Boult. MOON: A mixed objective optimization network for the recognition of facial attributes. *CoRR*, abs/1603.07027, 2016.

[27] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[28] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[29] M. L. Seltzer and J. Droppo. Multi-task learning in deep neural networks for improved phoneme recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6965–6969. IEEE, 2013.

[30] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[32] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[33] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. *Advances in neural information processing systems*, pages 1988–1996, 2014.

[34] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[36] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[37] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014.

[38] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.