Pattern Recognition 000 (2020) 107155

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

BLAN: Bi-directional ladder attentive network for facial attribute prediction

Xin Zheng[a], Huaibo Huang[b,c], Yanqing Guo[a,*], Bo Wang[a], Ran He[b,c]

[a] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
[b] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China
[c] National Laboratory of Pattern Recognition, CASIA, Center for Research on Intelligent Perception and Computing, CASIA, Center for Excellence in Brain Science and Intelligence Technology, CAS, China

ARTICLE INFO

Article history:
Received 13 March 2019
Revised 22 October 2019
Accepted 8 December 2019
Available online xxx

Keywords:
Deep facial attribute prediction
Bi-directional ladder attentive network (BLAN)
Residual dual attention module (RDAM)
Local mutual information maximization (LMIM)
Adaptive score fusion

ABSTRACT

Deep facial attribute prediction has received considerable attention with a wide range of real-world applications in the past few years. Existing works almost extract abstract global features at high levels of deep neural networks to make predictions. However, local features at low levels, which contain detailed local attribute information, are not well exploited. In this paper, we propose a novel Bi-directional Ladder Attentive Network (BLAN) to learn hierarchical representations, covering the correlations between feature hierarchies and attribute characteristics. BLAN adopts layer-wise bi-directional connections based on the autoencoder framework from low to high levels. In this way, hierarchical features with local and global attribute characteristics could be correspondingly interweaved at each level via multiple designed Residual Dual Attention Modules (RDAMs). Besides, we derive a Local Mutual Information Maximization (LMIM) loss to further incorporate the locality of facial attributes to high-level representations at each hierarchy. Multiple attribute classifiers receive hierarchical representations to produce local and global decisions, followed by a proposed adaptive score fusion module to merge these decisions for yielding the final prediction result. Extensive experiments on two facial attribute datasets, CelebA and LFWA, demonstrate that our BLAN outperforms state-of-the-art methods.

1. Introduction

Facial attributes represent intuitive semantic features that describe visual properties of face images [1,2], such as *smiling* and *eyeglasses*, contributing to numerous real-world applications, e.g., face verification [3,4], face recognition [5,6], and face retrieval [7,8]. Given a face image, facial attribute prediction aims to estimate whether desired attributes are present by learning discriminative feature representations and constructing accurate attribute classifiers.

Recently, deep convolutional neural networks (CNNs) have gained great popularity and have dramatically improved the performance of state-of-the-art algorithms in the field of facial attribute prediction. In general, deep facial attribute prediction methods can be categorized into two groups: part-based methods [9,10] and holistic methods [11,12]. Part-based methods first locate the positions of facial attributes and then extract features according to obtained location cues for the subsequent attribute prediction. In contrast, holistic methods learn attribute relationships and estimate facial attributes from the entire face images without any additional localization mechanism.

In this paper, we focus on holistic facial attribute prediction methods. The insight in this line of work lies in capturing shared and specific attribute features with customized architectures. Specifically, the customized networks learn shared features of all attributes across low-level layers. Then, these features flow to high-level layers, which resort to multiple split branches to predict attributes with different characteristics. However, in this process, only the high-level abstract features at the end of each branch take part in the final attribute prediction. The low-level shared information at low-level layers might vanish when arriving at the high-level layers [12]. Consequently, low-level features may not be fully explored and utilized.

Such deficiency of current holistic facial attribute methods prompts us to reconsider the relationship between the CNN network architecture and its extracted features at each level. Rather than capturing features with the commonality and specialty in deep networks, this paper considers leveraging the hierarchical structure of a deep network to learn the locality and globality of facial attribute features. Specifically, low-level CNN layers capture

* Corresponding author.
E-mail address: guoyq@dlut.edu.cn (Y. Guo).

https://doi.org/10.1016/j.patcog.2019.107155
0031-3203/

subtle and detailed face features, corresponding to the attributes that appear in local face regions, i.e., local facial attributes. As CNNs go deeper, more global and abstract information is explored to estimate the attributes that rely on the entire face to make predictions, i.e., global facial attributes. Therefore, the local and global natures of facial attributes can be significantly projected to the local and global feature representations, which are captured by low-level and high-level hierarchies of deep networks.

Taking such correlations between feature hierarchies and attribute characteristics, we design a novel Bi-directional Ladder Attentive Network (BLAN) to learn hierarchical feature representations from low-levels to high-levels, correspondingly to predict facial attributes with the locality and the globality. BLAN is constructed based on the autoencoder framework with multiple layer-wise bi-directional connections between its encoder and decoder. The encoder and decoder features learned at each level are fed into the proposed Residual Dual Attention Module (RDAM). RDAM adaptively interweaves these features to learn complementary information via residual connections. Besides, it employs dual channel-wise and spatial-wise attention to jointly learn what and where to focus, yielding richer attentive feature representations. To further improve the quality of learned interweaved representations at each level, Local Mutual Information Maximization (LMIM) loss is derived for incorporating the locality of input attributes into high-level representations. After that, multiple hierarchical classifiers operate on learned hierarchical attentive features with maximized mutual information to produce global and local decisions. Then, an adaptive score fusion module is followed to merge these multiple decisions at each level of BLAN, resulting in a further boost of the final performance. Extensive experiments on two facial attribute datasets CelebA and LFWA demonstrate that the proposed method outperforms state-of-the-art methods.

The main contributions are summarized as follows.

- We propose a novel Bi-directional Ladder Attentive Network (BLAN) which exploits the correlations between low-to-high hierarchy features and local-to-global facial attributes. Layer-wise bi-directional connections are designed based on the autoencoder framework to learn complementary features from the encoder and the decoder.
- Residual Dual Attention Module (RDAM) is developed to jointly learn dual channel-wise and spatial-wise attention for interweaving the encoder and decoder features. The residual connection ensures to capture complementary information.
- A Local Mutual Information Maximization (LMIM) loss is introduced to maximize the deep mutual information between input attentive attribute features and learned abstract representations, yielding improved features at each hierarchy.
- We present an adaptive score fusion strategy to merge local and global decisions from multiple hierarchical attribute classifiers for further boosting the performance of facial attribute prediction. Superior experimental results on two facial attribute datasets CelebA and LFWA demonstrate the effectiveness of the proposed BLAN.

2. Related work

2.1. Facial attribute prediction

Existing deep facial attribute prediction works can be generally grouped into two broad categories: part-based methods and holistic methods. We provide a detailed introduction about the two categories below, respectively.

Part-based methods extract feature representations from different positions of facial attributes. Each position corresponds to a single attribute classifier. Hence, the key of part-based methods exists in the localization mechanism, which further classifies part-based methods into two groups: separate auxiliary localization based methods and end-to-end localization based methods.

Typically, separate auxiliary localization methods utilize existing part detectors [41] or auxiliary localization algorithms to locate facial attributes in a separate and independent way. Zhang et al. [9] propose a PANDA model, which draws support from existing poselet part detectors [13]. Once poselet image patches are obtained, one CNN per poselet is trained to extract features from all patches. Kalayeh et al. [14] employ semantic segmentation as a separate auxiliary localization scheme to guide the prediction focusing on the naturally occurring areas of facial attributes. In contrast, Mahbub et al. [15] consider a more straightforward way. They resort to key points to segment faces into several images patches directly.

However, separated auxiliary localization based methods considerably rely on the accuracies of face detection, facial semantic segmentation, as well as facial landmark localization [16]. Thus, once these localization strategies are imprecise, or landmark annotations are unavailable, the performance of facial attribute prediction would be harmed significantly.

In contrast, end-to-end localization based methods exploit unified networks that discover position clues and predict attribute categories in an end-to-end manner. Liu et al. [17] first propose a cascaded deep learning framework for joint face localization and attribute prediction. Specifically, the cascaded network is made up of an LNet and an ANet, where LNet locates the entire face regions and ANet extracts high-level facial representations from located areas. However, the face regions located by LNet are too coarse to learn facial attribute related details. In light of this, Ding et al. [18] propose a cascade network to locate the regions that are only relevant to facial attributes. Specifically, a Face Region Localization network (FRL) is designed to locate attribute positions and generate image patches, where each patch is corresponding to one attribute. Then, a Parts and Whole (PaW) classification network is followed to make a binary classification on each image patch. Recently, Li et al. [10] design an AFFAIR network for learning a hierarchy of spatial transformation and predicting facial attributes without landmarks.

Nevertheless, end-to-end localization methods might cause redundant computations when many attributes might exist in the same facial area. Therefore, no matter what localization mechanism is adopted, it is still a challenge to avoid the adverse influences on subsequent prediction tasks.

Holistic methods extract features from the entire face images and predict facial attributes without any localization mechanism. What the most crucial issue that holistic methods concern about is modeling attribute relationships with customized architectures. In general, holistic facial attribute prediction networks share information across certain low-level layers and further split into multiple branches at high-level layers for specific attribute predictions. Specifically, shared layers ensure to learn general information among all attributes, whereas split forks extract attribute-specific features.

There exist two core challenges in holistic methods. One is appropriately assigning shared and attribute-specific information at different levels of networks. The other is excavating relationships among attributes for learning more discriminative feature representations [42].

MOON [19] first learns shared high-level features and then predicts multiple attributes simultaneously at the FC layer based on the 16-layer VGG [20]. Zhong et al. [21] replace high-level CNN features in MOON with mid-level ones for identifying the best representation over each attribute.

Compared with splitting networks at FC layers, Hand et al. [11] present a Multi-task deep CNN (MCNN), which branches several groups out at mid-level layers. Note that these groups are divided manually according to attribute semantics. However, there exists a problem that shared information from low-level layers may vanish after splitting in MCNN. In light of this, Cao et al. [12] rectify MCNN by deriving a Partially Shared structure (PS-MCNN) to learn shared and task-specific representations better. PS-MCNN divides attributes into four groups according to attribute locations. After that, it trains four corresponding Task-Specific Networks (TSNets) and one Shared Network (SNet) connected via the partially shared structure.

Notably, all the above methods design their networks by manual, Lu et al. [22] break this limitation and propose an automatically designed compact multi-task deep learning architecture, which fully shares features and learns discriminative representations adaptively. The automatically designed network starts with a thin multi-layer network. Then, it widens dynamically in a greedy manner, resulting in a more fast and compact model.

In summary, through grouping attributes manually according to semantic or locations [11,12], or automatically designing networks in an adaptive manner [22], attribute relationships can be projected to the high-level layers of networks for extracting abstract global features. However, the low-level features, which indicate more local and detailed context information of facial attributes, are not well addressed. More local and subtle features can be significantly captured by low-level layers, rather than high-level counterparts. There is a correspondence between feature representations captured by different hierarchies of deep networks and facial attributes with different local or global characteristics. Therefore, in this paper, we propose BLAN to discover and model such a correlation for yielding appealing facial attribute prediction performance.

2.2. Attention mechanism

Attention mechanism plays a vital role in improving the performance of CNNs in large-scale classification tasks [23,24]. On the one hand, attention selects to focus on a salient location with high activations [25]; On the other hand, attention strengthens the feature representations of different classification objectives over this location.

Wang et al. [25] propose a residual attention network to generate attention-aware features via a stacked encoder-decoder style attention module. Hu et al. [24] construct a Squeeze-and-Excitation (SE) block to model the interdependencies between channels of convolutional features explicitly. By stacking SE blocks, only minimal additional computations bring significant classification performance improvements. However, the SE block only emphasizes 'what' to focus for inferring the profitable channel attention, the spatial attention that concerns more about 'where' to focus is significantly ignored. In light of this, Woo et al. [26] design a Convolutional Block Attention Module (CBAM) for exploiting both spatial-wise and channel-wise attention, resulting in superior performance compared with using the single channel-wise attention. Besides, Zhu et al.[27] develop a Recurrent Attention Residual (RAR) module to select a residual component and refine context features from different depths of the network. RAR introduces residual learning technique [28] to capture complementary information. In the meantime, as many original features can be preserved as possible.

In light of these, we take both spatial-wise and channel-wise attention [26] into consideration and develop a novel Residual Dual Attention Module (RDAM). Specifically, channel-wise attention contributes to capturing the inter-channel relationships of features, whereas spatial-wise counterpart models the inter-spatial relationships. RDAM links the encoder features and the decoder features at each level of BLAN, analogous to a step on the ladder architecture.

2.3. Deep mutual information

Mutual Information (MI) indicates non-linear dependencies between random variables. It has been widely applied in the field of data science, such as information bottleneck [29] and feature selection [30]. Generally, MI can be formulated with Kullback-Leibler (KL) divergence between the joint distribution of two random variables (X and Z) and the product of their marginals, i.e., $I(X;Z) = D_{KL}(\mathbb{P}_{XZ}||\mathbb{P}_X \otimes \mathbb{P}_Z)$. Mutual information maximization is a general representation learning function to discover beneficial representations. Recently, several advances have broken through the limit of high dimensional computing difficulty faced by MI, leading to the further extension of such a conventional unsupervised representation objective to deep neural networks.

Belghazi et al. [31] first introduce Mutual Information Neural Estimation (MINE) working on continuous variables via back-propagation in deep neural networks. Further, they prove that MINE performs well when constructing bi-directional generative models. Nevertheless, the roles that MI plays in large-scale classification tasks are not discussed. Taking this point into consideration, Hjelm et al. [32] propose Deep InfoMax (DIM) to incorporate knowledge about locality and globality of the input to the deep MI for enhancing the representations for classification.

Based on DIM, we introduce Local Mutual Information Maximization (LMIM) loss to restrain the obtained features from containing as much label-related information as possible at each level of BLAN. By maximizing the deep mutual information, LMIM incorporates the input attentive facial attribute features into high-level abstract representations. In this way, the quality of features can be improved, and the performance of attribute prediction would be boosted.

In addition, our proposed BLAN bases on the autoencoder framework with layer-wise bi-directional hierarchical connections. In 2014, Rasmus et al. propose a similar ladder architecture [33], which has some applications in unsupervised learning and semi-supervised learning [33,34]. Lateral shortcut connections are adopted at every level of the model from the encoder to the decoder, i.e., single directional connections. The intuition behind this design is utilizing the encoder features to strength the reconstruction. Despite sharing the analogous name, our proposed network possesses the entirely distinct architecture and motivation. The proposed BLAN replaces the single directional connections with the bi-directional counterparts. Moreover, the features of both the encoder and the decoder contain facial attribute related details. Thus, they are equally treated for the subsequent attribute prediction task in our BLAN.

3. Bi-directional ladder attentive network

Given facial attribute images, the proposed BLAN first learns hierarchical feature representations from low-level layers to high-level layers under the autoencoder framework, corresponding to local and global features with the locality and the globality of facial attributes. Then, learned representations from both the encoder and the decoder at different hierarchies are fed into multiple residual dual attention modules for interweaving more discriminative attentive features. Next, these attentive features and the features at the end of the encoder are taken as inputs to multiple attribute classifiers. Under the constraints of proposed local mutual information maximization loss, classification loss, and reconstruction loss, these classifiers predict corresponding attributes at different levels and generate multiple decisions. Note that the scores of these classifiers are summed to produce another decision. After that, an adaptive score fusion module is adopted to integrate obtained multiple decisions, leading to a further boost of the final performance.

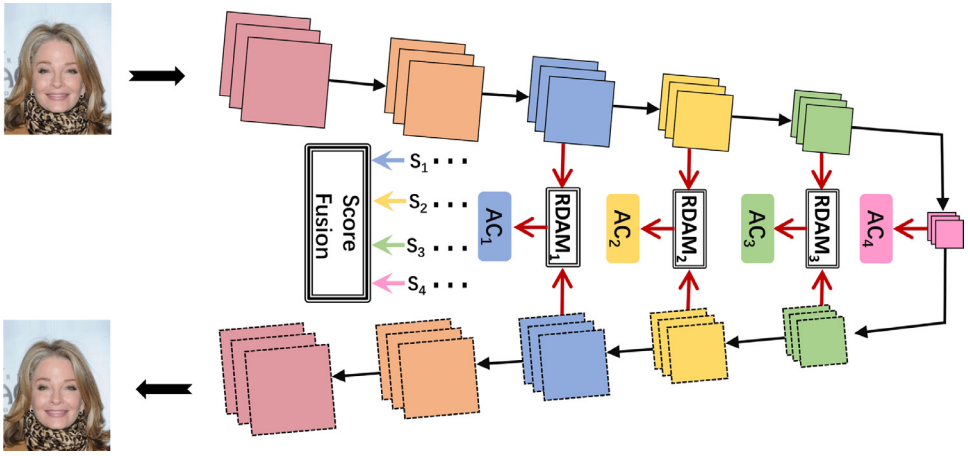


Fig. 1. The overall architecture of BLAN.

Table 1
Configuration of the basic BLAE.

Layer name	Output size	Encoder	Decoder
$conv_0$	128 × 128	7 × 7, 64, stride=2	
$block_1$	64 × 64	3 × 3, maxpool, stride=2 $\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$ upsample, scale factor=2
$block_2$	32 × 32	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$, stride = 2	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$ upsample, scale factor=2
$block_3$	16 × 16	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$, stride = 2	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$ upsample, scale factor=2
$block_4$	8 × 8	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$, stride = 2	-

We provide the overall architecture of the proposed BLAN in Fig. 1. BLAN consists of three main parts: the basic bi-directional ladder autoencoder BLAE, the residual dual attention modules $RDAM_i$ $(i = 1, \ldots, K-1; K = 4)$, and the attribute classifiers AC_i $(i = 1, \ldots, K; K = 4)$ with the adaptive score fusion module SF. In Fig. 1, the top half part with solid boxes denotes the encoder part, while the bottom half part with dashed boxes indicates the decoder part. The two parts jointly comprise the basic autoencoder framework BLAE. The adaptive score fusion module accepts the decisions represented by s_i $(i = 1, \ldots, K; K = 4)$ from K attribute classifiers for yielding the final prediction results. The details of each part are described respectively as below.

3.1. Basic bi-directional ladder autoencoder

BLAE contains an encoder, a decoder, and the bi-directional connections between them at different levels. The encoder employs Resnet-34 as the primary skeleton, where the last two layers are removed since they are customized for the ImageNet classification [28]. The decoder is a mirrored version of the encoder, but the last residual block of the encoder does not participate in the mirror operation. Upsampling layers are utilized to increase resolutions with scale factor 2. Hence, the convolutional residual blocks do not perform any resolution reduction. We provide the detailed configuration of BLAE in Table 1. Taking 3-channel RGB images with the resolution 256 × 256 as inputs, BLAE generates 64 × 64, 32 × 32, 16 × 16, and 8 × 8 four types of dimension features via $block_i$ $(i = 1, \ldots, K; K = 4)$. These outputs of internal convolutional blocks are pulled out at each level of the hierarchy for downstream multiple attention modules and attribute classifiers. Besides, BLAE is a supervised autoencoder that jointly classifies facial attributes and reconstructs input images. Consequently, the outputs of the decoder are 3-channel reconstructed RGB images, which share the same resolution with inputs. This network architecture design is mainly based on the following two considerations.

First, the decoder is introduced to enhance feature representations. During the process of the reconstruction, plenty of facial details can be fully captured to cater for lower reconstruction error. Therefore, taking the decoder features into account brings important complementary information, which might be ignored or vanished with the single encoder features. Besides, in terms of the reconstruction error, it can be taken as a specific form of regularization to some extent, leading to guaranteed generalization and uniform stability [35]. It has been proved theoretically and empirically that considering the reconstruction error never damages the performance of classification tasks and significantly improves the generalization ability [35].

Therefore, the reconstruction loss l_{REC} from the decoder is calculated by

$$l_{REC} = \|\tilde{\mathbf{x}} - \mathbf{x}\|_2^2. \tag{1}$$

We utilize the Mean Square Error (MSE) function in Eq. (1) to measure the distance between the reconstruction image $\tilde{\mathbf{x}}$ from the decoder and the original image $\mathbf{x}$, where $\|\cdot\|_2$ indicates the L_2-norm.

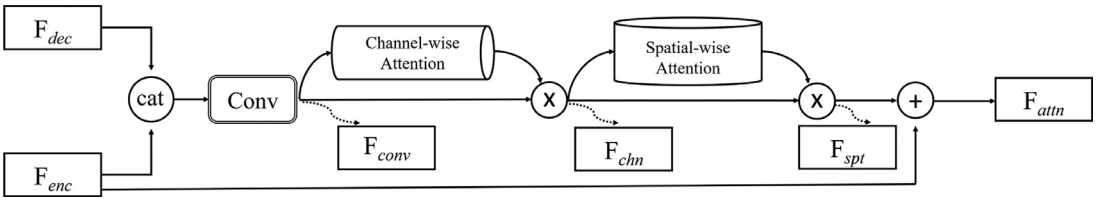


Fig. 2. The proposed residual dual attention module.

Second, pulling out features from both the encoder and the decoder bi-directionally at each level of BLAE helps to learn hierarchical representations. The motivation behind this design is that different layers of deep CNNs can learn distinct characteristics of features. Specifically, deeper layers learn abstract global context information, while detailed local context features can be captured at shallower layers. In terms of BLAE, we predict local facial attributes at low-level layers and estimate global facial attributes at high-level layers adaptively. There is no need to partition attribute groups according to globality and locality manually. At each level of the hierarchy, all attributes are classified so that we can select the best prediction for each one by the following ensemble strategy.

3.2. Residual dual attention module

We design an attention mechanism, termed Residual Dual Attention Module (RDAM), to learn attentive features from both the encoder and the decoder bi-directionally at each level of BLAN. To adaptively integrate more discriminative representations, features from block_1 ~ block_4 in Table 1 are propagated to corresponding attention modules RDAM_i $(i = 1, \ldots, K-1; K = 4)$. Therefore, RDAM modules become the steps on the ladder architecture of BLAN. All these attention modules share the same architecture but distinct parameter configurations. The overall architecture of RDAM is provided in Fig. 2.

Given the encoder feature F_{enc} and the decoder feature F_{dec} for each RDAM, they are first concatenated and fed into a convolutional layer. Then, the channel-wise attention and spatial-wise attention are executed dually in a sequential manner to model inter-channel and inter-spatial relationships of features, respectively. After that, the residual learning technique is introduced by adding the encoder features to maintain the original features for generating the ultimate attentive features. Note that in Fig. 2, we adopt the sequential connection configuration of the two types of attention, although they can exchange the order or be paralleled.

In terms of the channel-wise attention, it concerns what is meaningful in the given image by taking each channel of a feature map as a feature detector. Besides, the average pooling and the max pooling are utilized simultaneously via a shared network, where the former captures the extent of the target attribute and the latter congregates clues related to discriminative attribute features. Woo et al. [26] have proved that such joint utility of the two pooling operations performs better than using each independently.

Formulaically, let F_{conv} denote the output of the convolutional layer in Fig. 2, and *MLP* denotes the Multi-Layer Perceptron (MLP) with one hidden layer. Then, the output of channel-wise attention module can be written as

$$F_{chn} = F_{conv} \otimes \sigma\left(MLP(AvgPool(F_{conv}))\right) + MLP(MaxPool(F_{conv})). \quad (2)$$

Different from the channel-wise attention, spatial-wise attention captures where the informative regions exist. The average pooling and max pooling are also adopted whose outputs are concatenated and convolved via a convolution layer, denoted as $Conv_F$.

Table 2
Configurations of attribute classifiers.

Classifiers	Layers	Configurations	Output Size
AC_1	Conv-BN-Relu_1	5×5, stride=4	$32 \times 16 \times 16$
	Conv-BN-Relu_2	1×1, stride=1	$16 \times 16 \times 16$
AC_2	Conv-BN-Relu_1	3×3, stride=2	$64 \times 16 \times 16$
	Conv-BN-Relu_2	1×1, stride=1	$16 \times 16 \times 16$
AC_3	Conv-BN-Relu_1	3×3, stride=2	$128 \times 8 \times 8$
	Conv-BN-Relu_2	1×1, stride=1	$64 \times 8 \times 8$
AC_4	Conv-BN-Relu_1	3×3, stride=2	$256 \times 4 \times 4$
	Conv-BN-Relu_2	1×1, stride=1	$256 \times 4 \times 4$

Therefore, the output of spatial-wise attention module can be written as

$$F_{spt} = F_{chn} \otimes \sigma\left(Conv_F([AvgPool(F_{chn}); MaxPool(F_{chn})])\right). \quad (3)$$

Note that in both Eqs. (2) and (3), $\sigma(\cdot)$ indicates the sigmoid function and $\otimes$ denotes the element-wise multiplication.

As a result, the output of each RDAM can be denoted as $F_{attn} = F_{spt} + F_{enc}$, where the residual operation is adopted to preserve as many original encoder features as possible. Then, the obtained F_{attn} is fed into an attribute classifier for generating corresponding prediction scores.

3.3. Attribute classifiers

Attribute classifiers AC_i $(i = 1, \ldots, K; K = 4)$ of BLAN accept features from each attention module RDAM_i $(i = 1, \ldots, K-1; K = 4)$, as well as the feature map at the end of the pipeline of the encoder, i.e., the output of block_4, respectively. Each AC_i $(i = 1, \ldots, K; K = 4)$ yields the prediction scores for all attributes. Then, an adaptive score fusion module is followed to merge these scores from all hierarchies for producing the final result.

AC_i $(i = 1, \ldots, K; K = 4)$ is made up of two Conv-BN-ReLU units and two fully connection (FC) layers. There are no max pooling layers so that the downsampling operations are implemented by convolutional layers. After passing Conv-BN-ReLU units, the sizes of feature maps reduce by half, and the numbers of channels have different scale reductions for different classifiers. The detailed configurations of all attribute classifiers are listed in Table 2. Then, the outputs of Conv-BN-ReLU units of all classifiers are fed into two FC layers, denoted as FC_1 and FC_2 with the dropout operation between them for reducing the overfitting. FC_1 reduces the dimensionality of feature maps and projects them into 512-dimension feature vectors. FC_2 outputs $2N$ prediction scores, where N denotes the number of attributes. Note that each attribute implements the binary classification in FC_2.

We adopt the categorical cross entropy function with the softmax operation to calculate the classification loss $l_{CLS}^{(i)}$ for each AC_i $(i = 1, \ldots, K; K = 4)$, which can be written as

$$l_{CLS}^{(i)} = \sum_{j=1}^{N} softmax\left(\mathbf{s}_j, \mathbf{y}_j\right), \quad (4)$$

where $\mathbf{s}_j$ and $\mathbf{y}_j$ are the prediction score and corresponding attribute label of the jth facial attribute, respectively.

3.3.1. Local mutual information maximization

As mentioned above, deep mutual information estimation Deep InfoMax (DIM) [32] contributes to learning profitable representations for downstream tasks. Typically, DIM has two maximization objectives: global DIM and local DIM. Since facial attributes represent subtle details of faces that exist in local parts of images, the average MI maximization between the high-level representations and local regions of the inputs would benefit the attribute classification task. This constraint encourages the learned representations to involve as much label-related information of local image patches as possible, resulting in high-quality and profitable features.

In light of this, we introduce the local version of Deep InfoMax and construct Local Mutual Information Maximization (LMIM) loss. By incorporating the locality of the input attribute features into high-level representations at each level of BLAN, the quality of feature representations customized for facial attribute prediction can be significantly improved.

Given a face image, the basic BLAE first takes it as the input, and the output of each block_i is fed into an RDAM_i at a certain level. Then, the obtained feature map F_{attn_i} at this hierarchy inputs a corresponding attribute classifier AC_i, yielding a 512-dimension feature vector T_i at each FC_1 layer. After that, we utilize the most straightforward concat-and-convolve architecture according to [32] to construct LMIM loss. Two steps are executed: (1) concatenating the replicated T_i with the feature map F_{attn_i} at every location; (2) distinguishing the 'real' pair $[F_{attn_i};\ T_i]$ with the 'fake' one through the binary cross entropy loss. Note that the 'fake' pair $[F_{attn_i}{}';\ T_i]$ is produced by pairing the feature vector with a feature map from another image.

Formulaically, let $\mathcal{I}_\varphi(F_{attn_i};\ T_i)$ denote the MI between the feature map $F_{attn_i} := \{F_{attn_i}{}^{(u)}\}_{u=1}^{M\times M}$ and the global feature vector T_i, where φ is the hyperparameter of the attribute classifier. Therefore, local mutual information maximization loss $l_{LMIM}^{(i)}$ can be written as

$$l_{LMIM}^{(i)} = \arg\max_{\varphi} \frac{1}{M^2} \sum_{u=1}^{M} \mathcal{I}_\varphi\left(F_{attn_i}{}^{(u)};\ T_i\right). \tag{5}$$

Note that there are many methods for computing the mutual information $\mathcal{I}_\varphi(F_{attn_i}{}^{u};\ T_i)$, such as Jensen-Shannon MI estimator [36] and Noise-Contrastive Estimation (NCE) [37,38]. In this paper, in terms of facial attribute prediction, we calculate the binary cross entropy loss of the 'real' $[F_{attn_i};\ T_i]$ and 'fake' $[F_{attn_i}{}';\ T_i]$ pairs as the estimation of MI.

Consequently, MI maximization constraint models the relationships between local attribute feature maps and high-level abstract representations at each hierarchy, resulting in high-quality features for achieving promising facial attribute prediction performance.

3.3.2. Adaptive score fusion

Score-level fusion strategy aims to merge scores from multiple predictors so that an ensemble of networks can be implemented at the score level. As a result, the final prediction performance would be significantly enhanced. Remarkably, multiple classifiers in BLAN provide us an opportunity to further boost the performance by means of the score fusion strategy.

After obtaining multiple prediction scores from all attribute classifiers, we derive an adaptive score fusion strategy to combine local and global decisions from all levels of BLAN. Specifically, given scores $\mathbf{s}_i \in \mathbb{R}^{N\times 1}$ from attribute classifiers AC_i $(i = 1, \ldots, K; K = 4)$, we first sum them up as another basic score, denoted as $\sum \mathbf{s}_i \in \mathbb{R}^{N\times 1}$. Then, the final facial attribute prediction result $\mathbf{p} \in \mathbb{R}^{N\times 1}$ can be computed by

$$\mathbf{p} = softmax\left[\mathbf{W}_s \otimes \left(\mathbf{s}_i, \sum \mathbf{s}_i\right)\right], \tag{6}$$

where $(\cdot,\cdot)$ and $\otimes$ denote the operations of matrix concatenation and multiplication, respectively. Instead of setting in advance artificially, $\mathbf{W}_s \in \mathbb{R}^{(K+1)\times N}$ is the weight matrix learned during the training process in an adaptive manner. We also adopt the categorical cross entropy to compute the fused score loss l_{SF}, which can be denoted as

$$l_{SF} = \sum_{j=1}^{N} softmax\left(\mathbf{p}_j, \mathbf{y}_j\right). \tag{7}$$

Here, $\mathbf{p}_j$ is the jth attribute score from the adaptive score fusion module.

In summary, the total loss to be optimized in BLAN is

$$l_{TOT} = \sum_{i=1}^{K} \left(l_{CLS}^{(i)} + \alpha l_{LMIM}^{(i)}\right) + \beta l_{REC} + \gamma l_{SF}, \tag{8}$$

where α, β, and γ are hyperparameters for weighting local mutual information maximization loss l_{LMIM}, reconstruction loss l_{REC}, and adaptive score fusion loss l_{SF}, respectively.

4. Experiments

In this section, we systemically conduct experiments on two facial attribute datasets: CelebA and LFWA [17]. First, we introduce their descriptions and test protocols. Second, the implementation details involving training schemes, hyperparameter configurations, and attention settings are provided. Third, we compare and discuss our BLAN with state-of-the-art methods. Then, we experimentally illustrate the effectiveness of the hierarchical features learned by BLAN. Finally, the in-depth analysis of the proposed model is performed from the perspectives of ablative study, hyperparameter sensitivity, generalization ability, and model complexity, respectively.

4.1. Datasets and protocols

Celeb-Faces Attribute Dataset (CelebA) is a large-scale facial attribute dataset with large pose variations and background clutter. It is collected by labeling images selected from Celeb-Faces [39]. There are total 10,177 identities, 202,599 face images with 40 binary attribute annotations per image. Following the standard protocol in [17], CelebA is partitioned into three parts: 160,000 images of first 8000 identities for training, 20,000 images of another 1000 identities for validation, and the rest for testing.

Labeled Faces in the Wild Attribute Dataset (LFWA) consists of 13,233 images from 5749 people collected via news sources online. Specifically, there are 1680 people with two or more images. Each image shares the same annotated 40 attributes as CelebA [17]. As for the protocol, half samples (6,263 images) are for training and the remaining half for testing. Besides, we adopt the classification accuracy as the quantitative metric for all experiments.

4.2. Implementation details

We use the Pytorch platform and conduct all experiments on the NVIDIA Titan X GPU. In our experiments, the number of attributes N is set to 40, and the number of attribute classifiers K is set to 4. All input images are resized to 256×256 for CelebA and 224×224 for LFWA, following [12]. The Adam algorithm with the mini-batch size of 64, is adopted to optimize the proposed BLAN. Moreover, we set the learning rate as 0.001 with the linear decay by 10. For the adaptive score fusion module, we adopt the SGD algorithm with the learning rate of 0.0001 and the same mini-batch size.

Table 3
Classification accuracy (%) comparisons on CelebA and LFWA.

Methods	CelebA	LFWA
PANDA [9]	85.00	81.00
LNet + ANet [17]	87.30	84.00
Mid-level CNN Features [21]	89.80	85.90
MOON [19]	90.94	–
SOMP-branch-32 [22]	90.74	–
MCNN [11]	91.26	86.27
PaW [18]	91.23	–
AFFAIR [10]	91.45	86.31
BLAN (w/o OS)	91.73	86.09
BLAN (w/ OS)	**91.80**	**87.13**

Training schemes. During the training, we propose an overfitting suppression scheme to prevent our BLAN getting into the adverse overfitting problem. Specifically, the proposed scheme contains three strategies: (1) Insert a dropout layer with dropout probability of 0.5 between two FC layers in each attribute classifier [21]; (2) Initialize the encoder of BLAN with the weights from Resnet-34, which is pretrained on the large-scale ImageNet dataset for the classification task [28]; Note that for the remaining modules of BLAN, including the decoder, we impose the same weight initialization scheme as in [28] and train them from scratch. (3) Randomly mirror training images for further data augmentation. Besides, we also train our BLAN from scratch without imposing any overfitting suppression strategy. In the following experiments, such two different training schemes are denoted as 'w/ OS' and 'w/o OS', respectively.

Hyperparameter configurations. We empirically set the hyperparameters α, β, and γ in Eq. (8). The principle is all the loss functions are optimized in a balanced way to avoid model biasing any of them. Note that the weight of classification loss l_{CLS} is always set as 1. This configuration ensures that attribute classification is the most primary and essential task. Moreover, we set γ as 1 due to its indistinctive effect on the performance. As for α and β, we adjust them within the range of {1, 0.1, 0.01}.

Attention settings. Since our attention module RDAM contains both channel-wise attention and spatial-wise attention, we adopt three attention settings according to different connection ways, i.e., sequential channel-wise attention and spatial-wise attention, its variant version by exchanging the order, and parallel channel-wise attention and spatial-wise attention.

4.3. Comparison with state-of-the-art methods

We list the average classification accuracies of our proposed BLAN over 40 facial attributes and compare them with state-of-the-art methods in Table 3. Comparative methods include: PANDA [9], LNet + ANet [17], Mid-level CNN Features [21], MOON [19], SOMP-branch-32 [22], MCNN [11], PaW [18], and AFFAIR [10]. The best results are shown in bold.

As shown in Table 3, our proposed BLAN outperforms state-of-the-art methods with and without the overfitting suppression scheme. Compared with part-based methods, i.e., PANDA, LNet+ANet, PaW, and AFFAIR, first, our BLAN has the largest increase by 6.8% when compared with PANDA. Then, although LNet+ANet and PaW methods pay more attention to detailed attribute areas, they are still beaten by the proposed BLAN. That is because the suboptimal performance of localization mechanisms might cause adverse effects on the downstream attribute prediction task. In contrast, BLAN concentrates more on exploring and utilizing the correspondence between low-high-level features and local-global attributes. No auxiliary localization networks are built in this process. As a result, the negative influences of upstream tasks can be avoided.

In contrast to AFFAIR, our BLAN (w/o OS) yields comparable results over LFWA dataset. This is because AFFAIR introduces prior information by dividing 40 facial attributes into 8 groups according to attribute locations. However, after imposing the overfitting suppression strategies, our BLAN (w/ OS) improves the classification accuracy by about 0.8%. We attribute this result to BLAN's adaptive learning mechanism without any human knowledge. Thus, better generalization ability can be achieved.

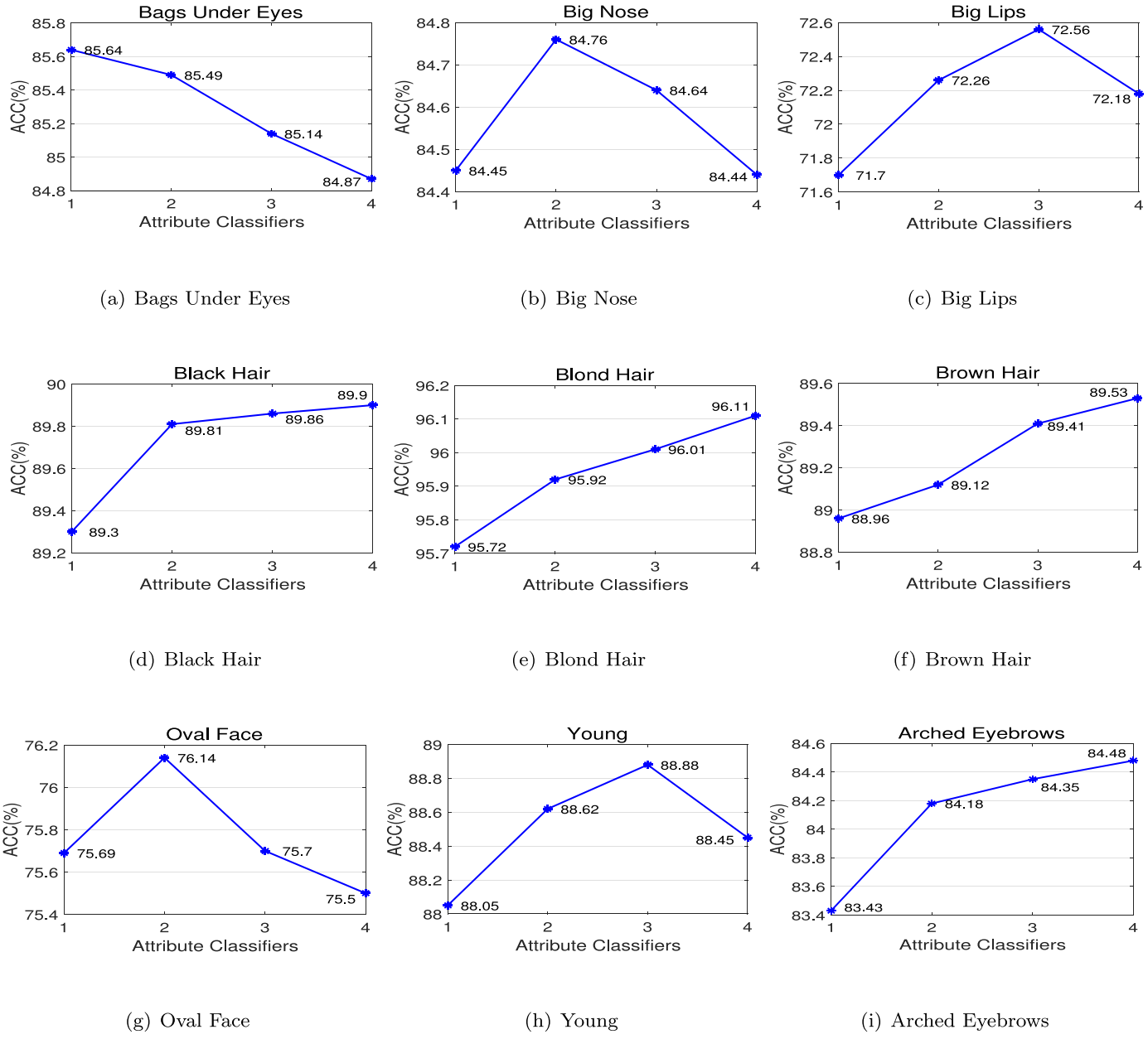
Compared with holistic methods, i.e., Mid-level CNN Features, MOON, SMOP-branch-32, and MCNN, the proposed BLAN achieves appealing improvements none the less. In contrast to Mid-level CNN Features, the hierarchical attentive features learned by BLAN perform better in identifying facial attributes. Since mid-level CNN features treat local and global features indiscriminately, this method predicts local and global attributes at the same network level. This limitation becomes the main reason of its inferior performance. Consequently, using single low-level or high-level features would impair the accuracy of attribute prediction. At this point, the advantage of BLAN's hierarchical representations can be highlighted. Besides, BLAN also outperforms the automatically designed SMOP-branch-32, as SMOP-branch-32 pursues fast model speed and compact compression at the expense of prediction accuracy. When comparing with MOON and MCNN that both consider attribute relationships, our BLAN has higher performance with different extents. Analogous to AFFAIR, MCNN manually partitions 40 facial attributes into 9 groups according to locations. However, the performance of attribute prediction might be over-restricted by current groups, especially when different individuals may give inconsistent grouping results. This result emphasizes one more the significant contribution of BLAN's adaptive learning for enhancing the model generalization.

4.4. Effectiveness of hierarchical learning

BLAN employs multiple attribute classifiers to make predictions, of which each learn corresponding local or global attribute features. To demonstrate the effectiveness of such hierarchical learning, we report the classification accuracies of attribute classifiers AC_1 ~ AC_4 over several facial attributes.

As shown in Fig. 3, on the whole, different attribute classifiers are dedicated to predicting facial attributes with different characteristics. For an arbitrary attribute, different classifiers have distinct performance. The top three subgraphs in Fig. 3 represent the prediction performance of facial attributes (a) Bags Under Eyes, (b) Big Nose, and (c) Big Lips, respectively. We can observe that the best prediction accuracies are achieved over attribute classifiers AC_1, AC_2, AC_3, respectively. That is to say, Bags Under Eyes can be well predicted only with the shallowest local features whereas Big Nose and Big Lips resort to mid-level features of BLAN. In contrast, hair color attributes (d) Black Hair, (e) Blond Hair, and (f) Brown Hair, ask for abstract global features at the highest level, as shown in the middle row of Fig. 3. Since hair colors represent high-level semantic and context information captured well by high hierarchies, AC_4 takes the heaviest responsibility of predicting these attributes.

For attributes (g) Oval Face and (h) Young, researchers might regard them as global attributes according to human knowledge. That means they should be estimated through high-level abstract features. However, our experiments in the bottom subfigures of Fig. 3 illustrate that the two attributes can be well predicted with mid-level features by AC_3, even low-level features by AC_2. Furthermore, (i) Arched Eyebrows might be clustered into local attribute groups by the artificial partition. That means this attribute should have been predicted with low-level features. However, our BLAN comes to the exact opposite conclusion, that is, (i) Arched Eyebrows requires high-level features to make an accurate prediction via AC_4.



(a) Bags Under Eyes (b) Big Nose (c) Big Lips

(d) Black Hair (e) Blond Hair (f) Brown Hair

(g) Oval Face (h) Young (i) Arched Eyebrows

Fig. 3. The illustration of the effectiveness of BLAN's hierarchical learning.

In summary, through hierarchical learning, BLAN adaptively captures attribute features with different local and global characteristics. As a result, each attribute classifier accepts corresponding representations at a certain level for yielding appealing prediction performance.

4.5. In-depth model analysis

4.5.1. Ablative analysis

We perform adequate ablative experiments over different component variants to explore the contribution of each part in BLAN. The detailed results are reported in Table 4. We decompose BLAN into five components: (1) Baseline (Enc): the encoder, (2) LMIM: local mutual information maximization loss, (3) Dec: the decoder, (4) RDAM: residual dual attention module, (5) SF: adaptive score fusion module. Note that (1)+(3)= BLAE, i.e., the basic bi-directional ladder autoencoder. Index 1 ~ 5 denote different combinations of the above five components, while index 6 indicates the overall model BLAN.

Table 4
Ablative analysis on CelebA.

Index	Baseline (Enc)	LMIM	Dec	RDAM	SF	ACC (%)
1	√					88.89
2	√	√				90.14
3		√	√			89.83
4	√	√	√			90.93
5	√	√	√	√		91.51
6	√	√	√	√	√	91.73

As shown in Table 4, first, we take the performance of the encoder as the baseline, obtaining 88.89% classification accuracy. Then, we verify the effect of LMIM loss in index 2. It can be observed that 1.25% performance improvement demonstrates the effectiveness of the proposed mutual information maximization constraint. After that, all attributes are estimated over the decoder in index 3 to illustrate its contribution. We can observe that Enc and Dec have comparable performance, although Dec is a little weaker

Table 5
Classification accuracy (%) on LFWA with different attention settings.

Attention settings		w/o OS	w/ OS
Sequential	Chn-Spt	86.09	**87.03**
	Spt-Chn	86.24	86.96
Parallel Chn ≃ Spt		86.30	86.74

Table 6
Classification accuracy (%) on LFWA with different hyperparameters.

Hyperparameters		α for l_{LMIM}		
		1	0.1	0.01
β for l_{REC}	1	86.84	86.91	86.97
	0.1	87.07	**87.13**	87.06
	0.01	86.93	87.03	86.84

than Enc. On the one hand, these results manifest that both Enc and Dec are capable of learning hierarchical feature representations; On the other hand, a small amount of information loss during the reconstruction would slightly hamper the decoder's performance. Nevertheless, the decoder still significantly serves as the auxiliary and complementary component of the encoder for yielding hierarchical reconstruction representations.

Combining Enc with Dec, we test the performance of the basic BLAE with LMIM constraint in index 4. It improves nearly 1.1% accuracy compared with the single Enc and the single Dec. We attribute this result to BLAN's layer-wise bi-directional connections at different levels. Note that BLAE only simply concatenates the features from Enc and Dec. In contrast, the proposed RDAM interweaves the two types of features through the channel-wise attention and the spatial-wise attention and enhances the performance up to 91.51% in index 5. At last, multiple classifiers of BLAN create a favorable condition for the proposed SF module to yield the best performance in index 6.

Besides, to illustrate the effects of different attention connection ways inside RDAM, we conduct corresponding experiments with the whole BLAN model over LFWA dataset. All the experiments consider three attention settings (i.e., sequential channel-wise attention and spatial-wise attention (Chn-Spt), its order-exchanged version (Spt-Chn), and parallel variant (Chn ≃ Spt)) under two situations (i.e., w/o OS and w/ OS).

As shown in Table 5, Chn-Spt mechanism w/o OS achieves the best performance. Besides, its variant Spt-Chn yields the comparable result with the parallel setting Chn ≃ Spt. Nevertheless, the performance of the two sequential connection settings is not much different. That means the order of channel-wise attention and spatial-wise attention hardly affects the attribute prediction performance. Moreover, there is an interaction between channel-wise attention and spatial-wise attention. The former discovers what is meaningful in the given facial attribute images, whereas the latter explores where the meaningful parts of attribute images exist. We attribute the inferior performance of the parallel setting to its inadequate interaction relationship modeling.

In contrast, when no overfitting suppression strategies are imposed, Chn-Spt scheme produces the poorer performance, whereas its variant Spt-Chn and parallel version Chn ≃ Spt show comparable results. We believe such experimental results are closely related to the random network initialization, which makes the prediction performance slightly sensitive to different configurations.

In conclusion, different attention settings of RDAM have only slight effects on the facial attribute prediction performance. Hence, in following experiments, we adopt sequential Chn-Spt attention setting for all RDAM modules due to its superior performance with the overfitting suppression scheme.

4.5.2. Hyperparameter sensitivity analysis

We test the hyperparameter sensitivity of BLAN over LFWA dataset when α and β vary in the range of {1, 0.1, 0.01}. All the experimental results in Table 6 are obtained with the proposed overfitting suppression scheme.

As shown in Table 6, the best hyperparameter configuration is $\alpha = \beta = 0.1$. For fixed $\beta = 1$, when α is reduced from 1 to 0.01, the classification accuracy raises gradually from 86.84% to 86.97%, obtaining about 0.1% increase. Then, for fixed $\beta = 0.1$ and $\beta = 0.01$, the highest classification accuracy yields when $\alpha = 0.1$. Since LMIM loss contributes to incorporating the low-level locality of facial attribute features into the high-level abstract representations, the hyperparameter α reflects the ability of this feature incorporation. If α is too large, excessive incorporations make it difficult to distinguish between low-level and high-level features, which runs counter to BLAN's design principle. When α becomes too small, the local features at low-levels cannot be well merged into high-level representations. In this way, the contribution of LMIM loss would be significantly limited. In contrast, for fixed α, all the best accuracies produce when $\beta = 0.1$. That means, if reducing β to 0.01, reconstructed features from the decoder might have limited impact on the attribute prediction. When increasing β to 1, the overall optimization might not balance well among multiple losses.

In summary, when α and β vary within the range of {1, 0.1, 0.01}, the difference between the highest (87.13%) and the lowest (86.84%) results is approximate 0.3%. Therefore, we believe that facial attribute prediction results of BLAN are hardly sensitive to both α and β within specific ranges.

4.5.3. Model generalization analysis

The best performance of BLAN over each attribute on CelebA and LFWA is reported in Table 7. Note that LFWA has much poorer performance than CelebA. That may because that LFWA dataset has fewer training images and more complex backgrounds, leading to more severe overfitting. In light of this, we conduct a series of experiments to illustrate the generalization ability of the proposed BLAN. Corresponding results are reported in Tables 8 and 9.

First, we consider the effects of different train/test proportions on LFWA dataset. As shown in Table 8, two divisions of dataset are tested, i.e., 80%:10% and 50%:50%. The former follows the same train/test proportion as CelebA, whereas the latter is the commonly-used test protocol of LFWA dataset as we mentioned in Section 4.1. All the experiments are conducted under two situations: w/ OS and w/o OS.

We can observe that overfitting indeed hampers the performance of facial attribute prediction, no matter what train/test proportions are provided. Furthermore, under the 50%:50% train/test proportion, the classification accuracy on LFWA dataset experiences about 1% increase. This result significantly reflects the effectiveness of our proposed overfitting suppression scheme. On the one hand, adding dropout layers is a beneficial measure to reduce the overfitting. By randomly dropping units in each training batch, BLAN would not simply fit input samples. On the other hand, compared with training from scratch with random weights, BLAN initialized by the weights of a pretrained model would show better generalization ability and faster training speed [22], which mitigate the adverse effect caused by insufficient training data [40].

Besides, the 80%:10% train/test proportion shows better performance compared with the 50%:50% partition, whether or not the overfitting suppression scheme is imposed. That means more training data can alleviate the overfitting problem to some extent. However, under the 80%:10% partition, our overfitting suppression strategies only improve approximate 0.4% classification accuracy. That is because more training data limits the function of overfitting suppression strategies. The same conclusion can be drawn

Table 7
Classification accuracy (%) of BLAN on CelebA and LFWA over 40 facial attributes.

Attributes	CelebA	LFWA	Attributes	CelebA	LFWA
5 o'clock Shadow	95.18	79.42	Male	98.32	94.24
Arched Eyebrows	84.74	83.07	Mouth Slightly Open	94.22	83.53
Attractive	83.25	80.94	Mustache	96.99	93.78
Bags Under Eyes	86.11	84.04	Narrow Eyes	87.78	83.44
Bald	99.02	92.85	No Beard	96.46	83.72
Bangs	96.26	91.26	Oval Face	76.86	78.24
Big Lips	72.59	80.45	Pale Skin	97.25	91.74
Big Nose	85.21	84.91	Pointy Nose	78.02	84.75
Black Hair	90.49	92.65	Receding Hairline	93.99	87.33
Blond Hair	96.27	97.57	Rosy Cheeks	95.36	87.56
Blurry	96.37	87.06	Sideburns	98.04	83.53
Brown Hair	89.79	82.66	Smiling	93.19	91.70
Bushy Eyebrows	93.08	86.25	Straight Hair	84.65	82.15
Chubby	95.88	77.37	Wavy Hair	85.35	82.12
Double Chin	96.58	82.95	Earrings	90.93	95.04
Eyeglasses	99.70	93.02	Hat	99.15	90.96
Goatee	97.69	84.64	Lipstick	94.34	95.06
Gray Hair	98.35	89.22	Necklace	88.16	90.35
Heavy Makeup	92.04	95.92	Necktie	97.20	83.91
High Cheekbones	88.13	88.95	Young	89.06	86.64
			Average	**91.80**	**87.13**

Table 8
Train/Test proportions tests on LFWA.

Train/Test Proportions	50% : 50%	80% : 10%
w/o OS	86.09	87.60
w/ OS	87.03	88.18

Table 9
Cross-dataset evaluation.

OS	Train/Test	$LFWA_{test}$	$CelebA_{test}$
w/o OS	$LFWA_{train}$	86.09	78.55
	$CelebA_{train}$	73.34	91.73
w/ OS	$LFWA_{train}$	87.13	**78.86**
	$CelebA_{train}$	**73.69**	91.80

from the results in the last two rows of Table 3. The performance on CelebA dataset does not see much improvement with the proposed overfitting suppression scheme. Due to CelebA's relatively sufficient training data, overfitting is no longer the main limitation of the model performance. In summary, the smaller the training data is, the model is prone to overfitting easier, and the proposed overfitting suppression strategies would make a more significant difference.

Second, we evaluate the facial attribute prediction performance cross two datasets in Table 9 (w/ and w/o OS) to further demonstrate BLAN's generalization. Two cases is tested: (1) all LFWA images for training and CelebA test set for testing; (2) CelebA train set for training and LFWA test set for testing.

We can observe that training on a dataset and testing on the other dataset indeed significantly hamper the performance of facial attribute prediction (declining 12.94% for CelebA test set and 13.44% for LFWA test set even with the overfitting suppression scheme). This is because there is a domain gap between the two datasets. Even labeled with the same types of facial attributes, the distributions of the two datasets do not match to each other accurately. Besides, more complicated backgrounds of LFWA images compared with CelebA would hinder the model's prediction performance as well. This explains why training over the entire LFWA and testing on CelebA (w/ OS: 78.86%, w/o OS: 78.55%) achieves better performance than training on CelebA and testing on LFWA (w/ OS: 73.69%, w/o OS: 73.34%). Besides, it is observed that after suppressing the overfitting, the performance over the cross datasets is improved approximately 0.3% on both CelebA and LFWA, which further illustrates the effectiveness of the proposed overfitting suppression scheme.

The generalization ability of BLAN considerably relies on its model design. On the one hand, BLAN adopts the adaptive hierarchical learning and the attribute relationship modeling strategies, which significantly contribute to improving the model generalization [12]. Then, the proposed LMIM loss further strengthens the connections between low-level features and high-level representations, leading to further improvement of generalization ability. On the other hand, rather than learning in the single-label paradigm, BLAN performs multi-attribute joint learning to guarantee its generalization ability [12].

4.5.4. Model complexity analysis

To elaborate the complexity of the proposed BLAN, we compare it with three state-of-the-art facial attribute prediction methods from the perspectives of classification accuracy, test speed per image, and the number of network parameters. Corresponding results over CelebA test set are listed in Table 10.

In terms of classification accuracy, our proposed BLAN achieves the best performance. Compared with MOON, BLAN has faster test speed and fewer parameters. SOMP-branch-32 sacrifices classification accuracy but has the least number of parameters and the fastest test speed due to its automatic network design. Compared with this method, BLAN has around 1% accuracy improvement but three times slower. In contrast to Paw, BLAN has twice as many parameters but about 0.8% boost. We believe this is acceptable since BLAN costs massive computations at both low levels and high levels for capturing local and global attribute feature representations.

In conclusion, BLAN contains more submodules to explore the correlations between feature hierarchies and attribute characteristics. That means BLAN indeed uses more computations. Neverthe-

Table 10
Comparisons of accuracy, speed and parameters on CelebA test set.

Methods	Accuracy(%)	Test Speed (ms)	Parameters (millions)
MOON [19]	90.94	33	119.73
SOMP-branch-32 [22]	90.74	9.6	1.49
PaW [18]	91.07	–	19
BLAN	**91.81**	29.42	38.34

less, its test speed of 29.42 ms can cater to the general real-time computation requirements.

5. Conclusion and future works

In this paper, we study the facial attribute prediction problem by exploiting the correlations between hierarchical features and attributes with the locality and the globality characteristics. We have proposed a novel Bi-directional Ladder Attentive Network (BLAN) to learn hierarchical representations at different levels of an autoencoder framework. Layer-wise bi-directional connections between the encoder and the decoder ensure to capture richer local and global attribute representations by merging the original features and the reconstruction features. The designed Residual Dual Attention Module (RDAM) shows the excellent ability in interweaving features from the channel level and the spatial level for learning more discriminative representations. Besides, the derived Local Mutual Information Maximization (LMIM) loss further incorporates the locality of the input attribute features to the high-level representations and produces high-quality features. Meanwhile, the proposed adaptive score fusion module performs well in merging multiple global and local decisions from all hierarchies for further boosting the performance. Extensive experiments on CelebA and LFWA verify the promise of the proposed BLAN.

BLAN excels at capturing the correlations between feature hierarchies and underlying task characteristics. This network architecture makes it adaptive to other multi-task issues beyond the facial attribute prediction in future works. Besides, since BLAN contains the reconstruction module, we are going to expand it to unsupervised learning fields.

Acknowledgements

This work is supported in part by the the State Key Development Program (Grant No. 2016YFB1001001), in part by the National Natural Science Foundation of China (NSFC) under Grant U1736119 and Grant U1936117, as well as the Fundamental Research Funds for the Central Universities under Grant DUT18JC06.

References

[1] N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar, Attribute and simile classifiers for face verification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2009, pp. 365–372.
[2] J. Wang, Y. Cheng, R. Schmidt Feris, Walk and learn: facial attribute representation learning from egocentric video and contextual data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2295–2304.
[3] Y. Li, L. Song, X. Wu, R. He, T. Tan, Learning a bi-level adversarial network with global and local perception for makeup-invariant face verification, Pattern Recognit. 90 (2019) 99–108.
[4] S. Zhang, R. He, Z. Sun, T. Tan, DeMeshnet: blind face inpainting for deep meshface verification, IEEE Trans. Inf. ForensicsSecur. (TIFS) 13 (3) (2018) 637–647.
[5] R. He, X. Wu, Z. Sun, T. Tan, Wasserstein CNN: learning invariant features for NIR-VIS face recognition, IEEE Trans. Pattern Anal. Mach.Intell. (TPAMI) (2018).
[6] R. He, T. Tan, L. Davis, Z. Sun, Learning structured ordinal measures for video based face recognition, Pattern Recognit. 75 (2018) 4–14.
[7] Y. Li, R. Wang, H. Liu, H. Jiang, S. Shan, X. Chen, 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 (ICCV), IEEE, 2015, pp. 3819–3827.
[8] H.M. Nguyen, N.Q. Ly, T.T. Phung, Large-scale face image retrieval system at attribute level based on facial attribute ontology and deep neuron network, in: Asian Conference on Intelligent Information and Database Systems, Springer, 2018, pp. 539–549.
[9] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev, PANDA: pose aligned networks for deep attribute modeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 1637–1644.
[10] J. Li, F. Zhao, J. Feng, S. Roy, S. Yan, T. Sim, Landmark free face attribute prediction, IEEE Trans. Image Process. 27 (9) (2018) 4651–4662.
[11] E.M. Hand, R. Chellappa, Attributes for improved attributes: a multi-task network utilizing implicit and explicit relationships for facial attribute classification., in: Proceedings of the 31st (AAAI) Conference on Artificial Intelligence, 2017, pp. 4068–4074.
[12] J. Cao, Y. Li, Z. Zhang, 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 (CVPR), 2018, pp. 4290–4299.
[13] L. Bourdev, J. Malik, Poselets: body part detectors trained using 3D human pose annotations, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2009, pp. 1365–1372.
[14] M.M. Kalayeh, B. Gong, M. Shah, Improving facial attribute prediction using semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 4227–4235.
[15] U. Mahbub, S. Sarkar, R. Chellappa, Segment-based methods for facial attribute detection from partial faces, IEEE Trans. Affective Comput. (2018).
[16] P. Perakis, T. Theoharis, I.A. Kakadiaris, Feature fusion for facial landmark detection, Pattern Recognit. 47 (9) (2014) 2783–2793.
[17] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3730–3738.
[18] H. Ding, H. Zhou, S.K. Zhou, R. Chellappa, A deep cascade network for unaligned face attribute classification, in: Proceedings of the Conference on Artificial Intelligence(AAAI), 2018.
[19] E.M. Rudd, M. Günther, T.E. Boult, MOON: a mixed objective optimization network for the recognition of facial attributes, in: European Conference on Computer Vision (ECCV), Springer, 2016, pp. 19–35.
[20] O.M. Parkhi, A. Vedaldi, A. Zisserman, et al., Deep face recognition, in: Proceedings of the British Machine Vision Conference 2015, (BMVC), vol. 1, 2015, p. 6.
[21] Y. Zhong, J. Sullivan, H. Li, Leveraging mid-level deep representations for predicting face attributes in the wild, in: Image Processing (ICIP), 2016 IEEE International Conference on, IEEE, 2016, pp. 3239–3243.
[22] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, R. Feris, Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2017, p. 6.
[23] K. Li, Z. Wu, K.-C. Peng, J. Ernst, Y. Fu, Tell me where to look: guided attention inference network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9215–9223.
[24] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 7132–7141.
[25] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 6450–6458.
[26] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, CBAM: convolutional block attention module, in: European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[27] L. Zhu, Z. Deng, X. Hu, C. Fu, X. Xu, J. Qin, P. Heng, Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection, in: European Conference on Computer Vision (ECCV), 2018, pp. 122–137.
[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[29] N. Slonim, N. Tishby, Agglomerative information bottleneck, in: Advances in Neural Information Processing Systems (NIPS), 2000, pp. 617–623.
[30] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach.Intell. (TPAMI) (8) (2005) 1226–1238.
[31] M.I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, R.D. Hjelm, A.C. Courville, Mutual information neural estimation, in: Proceedings of the Conference on Machine Learning (ICML), 2018, pp. 530–539.
[32] R.D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, A. Trischler, Y. Bengio, Learning deep representations by mutual information estimation and maximization, in: International Conference on Learning Representations (ICLR), 2019.
[33] A. Rasmus, T. Raiko, H. Valpola, Denoising autoencoder with modulated lateral connections learns invariant representations of natural images, in: International Conference on Learning Representations Workshop Track Proceedings (ICLRW), 2015.
[34] C.K. Sønderby, T. Raiko, L. Maaløe, S.K. Sønderby, O. Winther, Ladder variational autoencoders, in: Advances in Neural Information Processing Systems (NIPS), 2016, pp. 3738–3746.
[35] L. Le, A. Patterson, M. White, Supervised autoencoders: improving generalization performance with unsupervised regularizers, in: Advances in Neural Information Processing Systems (NIPS), 2018, pp. 107–117.
[36] S. Nowozin, B. Cseke, R. Tomioka, f-GAN: training generative neural samplers using variational divergence minimization, in: Advances in Neural Information Processing Systems (NIPS), 2016, pp. 271–279.
[37] M. Gutmann, A. Hyvärinen, Noise-contrastive estimation: a new estimation principle for unnormalized statistical models, in: Proceedings of International Conference on Artificial Intelligence and Statistics, 2010, pp. 297–304.
[38] M.U. Gutmann, A. Hyvärinen, Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics, J. Mach. Learn. Res. 13 (Feb) (2012) 307–361.

[39] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Advances in Neural Information Processing Systems (NIPS), 2014, pp. 1988–1996.
[40] N. Sarafianos, T. Giannakopoulos, C. Nikou, I.A. Kakadiaris, Curriculum learning of visual attribute clusters for multi-task classification, Pattern Recognit. 80 (2018) 94–108.
[41] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen, et al., Deep Learning for Generic Object Detection: A Survey, International Journal of Computer Vision (2019) 1–58.
[42] L. Liu, P. Fieguth, G. Zhao, R. Chellappa, M. Pietikainen, et al., From BoW to CNN: Two Decades of Texture Representation for Texture Classification, International Journal of Computer Vision 127 (1) (2019) 74–109.

Xin Zheng received the B.E. degree in Integrated Circuit Design and Integration System, Dalian University of Technology, in 2017. She is currently a Master Student in the School of Information and Communication Engineering, Dalian University of Technology. Her research interests are in computer vision and pattern recognition.

Huaibo Huang received the B.E. degree in Measurement and Control Technology and Instrument from Xi'an Jiaotong University in 2012, and the M.E. degree in Optical Engineering from Beihang University in 2016. He is currently a Ph.D. student in the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), CASIA, Beijing, China. His current research interests include computer vision and pattern recognition.

Yanqing Guo received the B.S. degree and Ph.D. degree in Electronic Engineering from Dalian University of Technology of China, in 2002 and 2009, respectively. He is currently a professor with School of Information and Communication Engineering, Dalian University of Technology. His research interests include multimedia security and forensics, digital image processing, deep learning and machine learning.

Bo Wang received the Ph.D. degree from Dalian University of Technology, China, in 2010. He is currently an Associate Professor with the School of Information and Communication Engineering, Dalian University of Technology. His research interests include image forensics and image steganalysis.

Ran He received the BE and MS degrees in computer science from Dalian University of Technology, and the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2001, 2004, and 2009, respectively. Since September 2010, he has been with the National Laboratory of Pattern Recognition, where he is currently an associate professor. He currently serves as an associate editor of Neurocomputing (Elsevier) and serves on the program committees of several conferences. His research interests include information theoretic learning, pattern recognition, and computer vision.