Prior-Guided Multi-Scale Fusion Transformer for Face Attribute Recognition

Shaoheng Song[1,2,3], Huaibo Huang[2,3], Jiaxiang Wang[1], Aihua Zheng[1,*], and Ran He[2,3]

[1] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, China
[2] Center for Research on Intelligent Perception and Computing (CRIPAC)
[3] Institute of Automation, Chinese Academy of Sciences
Emails: soul951128@gmail.com, huaibo.huang@cripac.ia.ac.cn, Netizenwjx@foxmail.com, ahzeng214@foxmail.com, rhe@nlpr.ia.ac.cn

Abstract. Multi-label face attribute recognition (FAR) refers to the task of predicting a set of attribute labels for a facial image. However, existing FAR methods do not work well for recognizing attributes of different scales, since most frameworks use the features of the last layer and ignore the detailed information which is crucial for FAR.To solve this problem, we propose a prior-guided multi-scale fusion transformer, which possesses the ability to build the fusion among features of different scales with prior knowledge of attributes. First, we employ a unifying Graph Convolution Network (GCN) to model the relations between multiple attributes by the prior knowledge of facial labels and the statistical frequencies of co-occurrence between attributes. Second, we propose a multi-scale fusion module, which uses adaptive attention to fuse features from two adjacent layers, and then simultaneously fuse the features of different scales hierarchically to explore the multilevel relation. In addition, we utilize the transformer as a feature extraction module to achieve a global correlation among the acquired features. Experiments on a large-scale face attribute dataset verify the effectiveness of the proposed method both qualitatively and quantitatively.

Keywords: Face Attribute Recognition · Multi-Scale · Prior-Guided.

1 Introduction

The technology of face attribute recognition, which aims to predict a number of attributes in face images, has drawn extensive attention due to its potential applications such as face retrieval [31], face recognition, [2] *etc.* Despite the great achievements that have been made in this field, there still exist a variety of challenges to address. During our study, we summarize the difficulties in face

* Corresponding Author

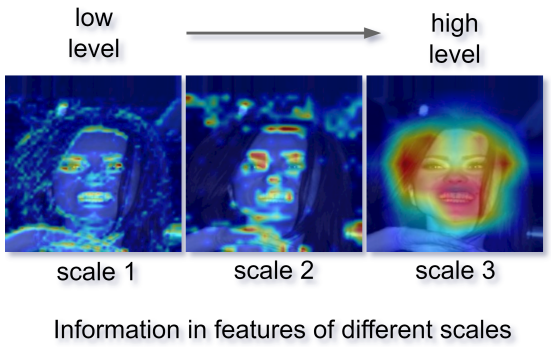

Fig. 1. The parts with warm colors represent where the network pays attention. a) Extracting features of different scales in the network. We can find that the network pays attention to some local details in the initial stage and gradually some global information in the later stage.

attribute recognition into three main points. First of all, as shown in Fig. 1, the information obtained from the last layer of the network mainly represents the high-level characteristics. To a certain extent, only using the last-layer feature may affect the ability of the network to capture the potential characteristics presented in low-level information. Second, there are some subtle correlations between attributes since the occurrence of some labels may affect each other in face images. Usually, with a great chance, *beard* comes together with *male*, and *receding hairline* indicates that a person is not *young*. Finally, the CNN models pay more attention to local information, while experiencing difficulty to capture global representations. So the lack of global relations between features may weaken the ability of representation learning.

In recent years, with the renaissance of CNN, some deep models have been applied to face attribute recognition and have made great progress. For instance, Liu *et al.* [15] solve the attribute recognition problem by learning independent classifiers for each attribute. Kalayeh *et al.* [10] use semantic segmentation to mine local clues to guide attribute prediction, which means to position the area where the attribute comes from. Cao *et al.* [1] consider both identity information and attribute relations. SSPL [25] captures the pixel-level and image-level semantic information. HFE [33] combines attribute and ID information to learn a fine-grained feature embedding. Nian *et al.* [21] use a decoupling matrix. Despite their achievements, the three challenges mentioned above remain not well addressed.

In this work, we propose a prior-guided multi-scale fusion transformer to capture the local and global representations in image features with attribute prior information. It consists of two sub-modules to progressively capture information hierarchically. First, we apply an attribute residual mapping module (ARMM) to capture the relations between attributes. Inspired by [30], following the GCN

[12] paradigm, we use the prior knowledge of facial labels and the statistical frequencies of co-occurrence between attributes to construct the graph. Then the obtained feature can enhance attribute-related regions in image features. Second, inspired by [26], we design a multi-scale fusion module (MFM) to enable the network to gradually fuse low-level and high-level features at the same time and then simultaneously utilize the features of different scales. In addition, we introduce Swin-Tranformer [14] to model global relations. Then pairwise relations can be fused into image features in a global way by performing message passing through each spatial patch. The two sub-modules are aggregated together to perform multilevel relations learning for face attribute recognition.

In summary, the contributions of this work are three-fold.

(1) We propose a multi-scale fusion module to jointly capture relations between low-level and high-level features for face attribute recognition.

(2) We propose an improved end-to-end architecture based on a transformer and prior information of attributes. The relations between attributes can be learned to strengthen the representations.

(3) Experiments show the superiority of the proposed method over recent methods and the effectiveness of our framework for face attribute recognition.

2 Related Work

2.1 Face Attribute Recognition

Face attribute recognition has risen in recent years. Rudd *et al.* [22] define face attribute recognition as a regression task. It applies a single DCNN to learn multiple attribute labels. Zhong *et al.* [35] use the mid-level features as the best representation for recognition. Then Hand *et al.* [6] branch out to multiple groups for modeling the attribute correlations due to many attributes being strongly correlated. Cao *et al.* [1] design a partially shared structure called PS-MCNN. Lu *et al.* [17] propose a network to learn shared features in a fully adaptive way, which incrementally widens the current design in a layer-wise manner. He *et al.* [9] utilize dynamic weights to guide network learning and Huang *et al.* In several latest works, HFE [33] combines attribute and ID information to learn fine-grained feature embeddings, then attribute-level and ID-level constraints are utilized to establish the hierarchical structure. SSPL [25] proposes a method, which captures semantic information of facial images in the pixel-level and image-level.

2.2 Graph Convolution Network

GCN [12] is used to process topological data. Recently, graph-based reasoning has been proved to be beneficial to a variety of vision tasks including multi-label classification [3], FVQA [36], zero-shot learning [29], social networks [32], etc. In recent years, image classification [3] and face attributes classification [21] propose to use GCN to learn the representations with attribute information.

2.3 Vision Transformer

At present, Transformer [27] is applied to the vision tasks based on the Vision Transformer (ViT) [5]. This demonstrates that pure Transformer-based architectures can also obtain relatively good results, promising the potential of handling the vision tasks and natural language processing (NLP) tasks under a unified Transformer. Recently, rather than focusing on a particular visual task, some works try to design a general vision Transformer backbone for general-purpose vision tasks.[4,14,28] these transformers have been proved effective in features extracting and perform well in downstream tasks.

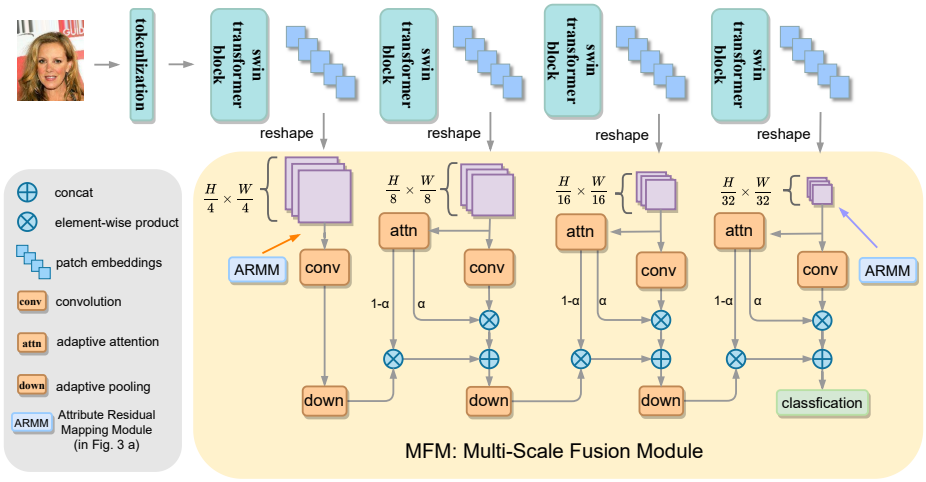

Fig. 2. Structure of prior-guided multi-scale fusion transformer. The inputs of the framework are face images and prior information of attributes. We use Swin-Transformer [14] to get four features of different scales in a global way. Then we take the reshaped features into the Multi-Scale Fusion Module (MFM) to fuse multi-scale features of high-level and low-level in a hierarchical manner. At the same time, we employ an Attribute Residual Mapping Module (ARMM) to map the prior information of attributes into features in the first and last layers of the network. The details of ARMM is described in Fig. 3. The orange arrow represents the output prior information from the first GCN [12] layer, and the purple arrow represents the output prior information from the last GCN [12] layer.

3 APPROACHES

In this paper, we propose a prior-guided multi-scale fusion transformer to simultaneously utilize features of different scales and prior information of attributes to process the feature learning in a global way for face attribute recognition. As shown in Fig. 2, our network consists of two main modules: 1) Attribute

Residual Mapping Module (ARMM), to combine the prior knowledge of related attributes and the image features in the deep and the shallow layers. 2) Multi-scale Fusion Module (MFM), to obtain the relations between low-level detailed features and high-level semantic features in different scales, which are extracted from a transformer. We shall elaborate on these two modules in the following two sections.

3.1 ARMM: Attribute Residual Mapping Module

1) GCN review

GCN [12] can capture the relationship between nodes in structured graph data in a semi-supervised manner. The graph is represented in the form of $\mathbf{G} = \{\mathbf{V}, \mathbf{A}\}$, where $\mathbf{V} \in \mathbb{R}^{N \times D}$ is the set of $\mathbf{N}$ data vectors in D dimension, and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is adjacency matrix. Then GCN can encode the pairwise relationship among data. The goal of GCN is to learn a function $f(\cdot, \cdot)$ on a graph $\mathbf{G}$, which takes initial node continuous representations $\mathbf{V}$ and an adjacency matrix $\mathbf{A}$ as inputs. And it updates the node features as $\mathbf{X}^{l+1} \in \mathbb{R}^{N \times D'}$ after spreading information through each layer. Every GCN layer can be formulated as:

$$\mathbf{X}^{l+1} = f(\mathbf{X}^l, \mathbf{A}) = \sigma(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{X}^l\mathbf{W}^l), \tag{1}$$

where $\mathbf{D} = diag(d_1, d_2 ... d_k)$ is a diagonal matrix with $\mathrm{d}_i = \sum_{j=1}^{n} \mathbf{A}_{ij}$. $\mathbf{W}^l \in \mathbb{R}^{D_l \times D_{l+1}}$ is a transformation matrix learned during training and σ denotes a non-linear operation, which is acted by LeakyReLU [18] for our purpose. Finally, $\mathbf{X}^{l+1} \in \mathbb{R}^{N \times D_{l+1}}$ denotes the output in the l+1-th layer.

2) Attribute Prior Information

We aim to input the internal relations between attributes obtained from the distribution of data as prior information into the network. First of all, the construction of a graph is necessary. Inspired by [30], to obtain the prior information of the label attributes, we extract the feature vector of each word related to the label from Google Corpus (GoogleNews-vectors-negative300). Followed by [13] and [20], since each attribute of the face is composed of multiple words, we sum the features of the words contained in each attribute label and take the average. Then we use the final vectors as graph nodes with prior information. According to this, we construct graph nodes as $V \in \mathbb{R}^{K \times D}$, where K denotes the total number of the labels.

In order to better propagate information between attributes, a correlation matrix is a key point. Then we get the matrix with statistical co-occurrence information by the distribution of samples in the training set. Following [3], we build this correlation matrix in a data-driven way. That is, we mine their relevant information based on the distribution of attributes within the dataset and compute the degree of semantic relevance of attributes. The attribute correlation dependency is modeled in the form of conditional probability between attributes. We denote the $P(\mathbf{V}_i \mid \mathbf{V}_j)$ as the probability of occurrence of attribute $\mathbf{V}_i$ when attribute $\mathbf{V}_j$ appears. To construct the correlation matrix, to begin with, we define the total number of occurrences of each attribute as N_i. Then we count

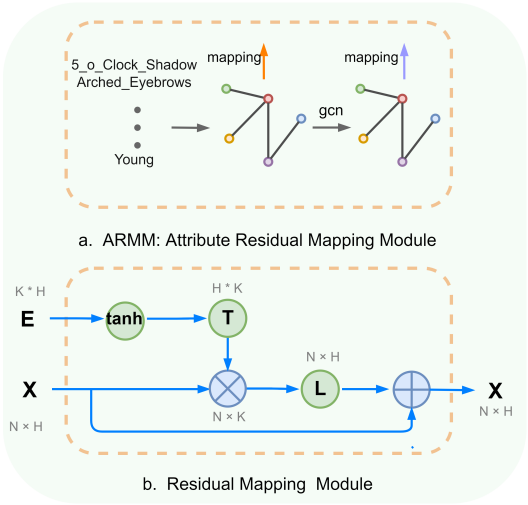

Fig. 3. a) Attribute Residual Mapping Module (ARMM). GCN is utilized to capture the relationship between the prior information of attributes extracted from Google Corpus. Then the information are mapping into the image features through a residual mapping module. The attribute features in first layer are mapped into head features, and the second are mapped into tail features. b) The residual mapping module. 'T', '$\oplus$', '$\otimes$' denote matrix transpose, sum and multiplication operations respectively. 'L' and 'tanh' are activate fuction. X and E are transformer feature and GCN feature. The shape of each tensor is marked in gray annotation.

the number of co-occurrences of every attribute pair and build a co-occurrence matrix $M \in \mathbb{R}^{K \times K}$, which K means the total number of face attributes. Then, we define the correlation matrix by the conditional probability matrix as:

$$[\mathbf{A}]_{ij} = \mathbf{M}_{ij} / \mathbf{N}_i, \tag{2}$$

where M_{ij} denotes the number of co-occurrences of i-th and j-th facial attributes and N_i denotes the occurrence times of i-th face attribute.

Since there are some uncommon co-occurrence relationships in the data, it may cause noise. We apply a threshold τ to filter the noisy conditional probabilities and obtain the robust matrix:

$$[\mathbf{A}]_{ij} = \begin{cases} 0 & if\ \mathbf{A}_{ij} < \tau \\ \mathbf{A}_{ij} & if\ \mathbf{A}_{ij} > \tau \end{cases} . \tag{3}$$

3) Residual Mapping Module

The module aims to map the prior information of the attributes into image features, which is processed from GCN mentioned above. So the network can apply the prior information to weighted related information for face attribute recognition. We use the module in the first and last layer of the network for

shallow and deep guidance. As shown in Fig. 3 b, the details of the module are as follows:

$$y = \sigma(X\phi(E)^T) + X, \tag{4}$$

here $\mathbf{X} \in \mathbb{R}^{N\times H}$ is transformer feature from a middle layer, N is the patch number and H is the dimension of the hidden feature. $\mathbf{E} \in \mathbb{R}^{K\times H}$ indicates the hidden attribute embeddings of GCN. $\sigma(\cdot)$ denotes a non-linear activation operation, T is transpose operation and $\phi(\cdot)$ means a Tanh function. Finally, we use a residual connection to add the original $\mathbf{X}$.

3.2 MFM: Multi-Scale Fusion Module

In the task of face attribute recognition, the information extracted form images is often inadequate, which brings big challenges. The existing methods have two potential issues that might limit face attribute recognition performance. First, CNN-based methods may focus on local regions, which might ignore the spatial relations, because attributes such as necklace and hair occupy irregular areas in image space, and sometimes pixels in these irregular areas may lack close spatial connections. Second, the features that recent methods use for face attribute recognition are only at a certain level. However, the scales of face attributes in images are different, such that hair takes up a lot of space and eyes occupy a very small area. Therefore, to enable the network to pay attention to global information and recognize attributes at different scales, we aim to use a transformer to capture long-distance relations in spatial and a multi-scale fusion module to process the features of different scales extracted from a transformer. Recently, transformer shines in the field of computer vision, because self-attention can capture the global relevant information in space in a parallel step. We apply Swin-Transformer [14] as a backbone to get features in a global way for face attribute recognition. Given an image $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, through the non-overlapped convolutional token encoder, we obtain patch tokens. Next, there are four Swin-Transformer blocks and each block contains multiple layers of multi-head self-attention mechanism. Then we obtained four image features of different scales, $\mathbf{X}_1 \in \mathbb{R}^{\frac{H\times W}{4\times 4}\times C1}$, $\mathbf{X}_2 \in \mathbb{R}^{\frac{H\times W}{8\times 8}\times C2}$, $\mathbf{X}_3 \in \mathbb{R}^{\frac{H\times W}{16\times 16}\times C3}$, $\mathbf{X}_4 \in \mathbb{R}^{\frac{H\times W}{32\times 32}\times C4}$, where the first dimension of the vector represents the number of patches and the second dimension of the vector represents different dimension of each patch. Additionally, $\mathbf{H}$ and $\mathbf{W}$ represent the height and weight of the features. As shown in Fig. 1, these features of different scales can focus on regions of different levels. The network can aggregate low-level detailed information and high-level semantic information by using these features at the same time. So the fused features used for face attribute recognition contain more sufficient information. For this purpose, we designed a multi-scale fusion module inspired by [26].

Firstly, we reshape the extracted features to $C1\times\frac{H}{4}\times\frac{W}{4}, C2\times\frac{H}{8}\times\frac{W}{8}, C3\times\frac{H}{16}\times\frac{W}{16}, C4\times\frac{H}{32}\times\frac{W}{32}$ respectively. Then in order that both local and global information can be exploited simultaneously, we use a convolution module to capture the short-range context after reshaping operation. To predict more suitable selection weights under the scenario of fusion in adjacent features. We introduce

an adaptive attention module to let the network automatically select the area that needs attention in adjacent features. When obtaining an adaptive weight α, we use 1-α to select the previous level of information, and α to select the current level of information. Finally, adaptive pooling is applied to match the scale of the previous level to the current features. We define the adaptive attention α as:

$$\alpha = \mathbf{Attn}(\mathbf{X}^{l+1}), \tag{5}$$

where $\mathbf{X}^{l+1}$ is the feature of the l+1-th layer and **Attn** is adaptive attention. We define the hierarchical flow as:

$$\mathbf{H}^{l+1} = \mathbf{down}(\alpha \times \mathbf{Conv}(\mathbf{X}^{l+1}) + (1-\alpha) \times \mathbf{H}^{l}), \tag{6}$$

where $\mathbf{H}^{l+1}$ is the fused feature between the information in the l+1-th layer and the previous fused information, **down** is the adaptive pooling and **conv** is the convlution module to cupture the short-range context.

4 Experiments

4.1 Dataset.

The proposed method is evaluated on a largescale face attribute dataset. The CelebA [16] consists of 202,599 face images collected from 10,177 people. Each face includes 40 attribute labels. Following the standard protocol in [16], CelebA is partitioned into three non-overlapping parts: 160,000 images of first 8000 identities for training, 20,000 images of another 1000 identities for validation, and the rest for testing.

4.2 Evaluation Metrics.

For fair comparison, we utilize accuracy as our criteria in our study to evaluate our performance.

4.3 Implementation detail.

Similar to [31], the number of convolution layers in our GCN is set to 2. The base model of Swin-Transformer [14] is used as the backbone. The hyper-parameters τ is set to 0.1. The input shape of images is reshaped to 224 $\times$ 224 with the data augmentations of randomly flip and color enhancement. We train our reasoning model using the Adam [11] algorithm. A pre-trained model of Swin is used and the initial learning rate is set to 10^{-4}, which is gradually reduced to 10^{-7} after 6 epochs. Our model is trained on the CelebA [16] dataset and gets converged with 10 epochs and it takes three hours with one NVIDIA RTX 3090.

4.4 Comparison to the State-of-the-Arts.

TABLE 1 show accuracy evaluations on CelebA [16]. The proposed method shows the relatively better performance on the dataset measured by evaluation metric. *MOON* [22] only uses a deep regression model and gets an accuracy of 90.94%, which is a relatively low level. *MCNN-AUX* [6] branches out several forks corresponding to different attribute groups and achieves 91.26%. *Adaptive Weighted* [9] uses a validation loss which dynamically add learning weights to each attribute and achieves 91.80%. *Adaptive Sharing* [17] starts with a thin multi-layer network and dynamically widens it in a greedy manner during training and achieves 91.26%. *GAN and Dual-path* [8] complement face parsing map with real images and achieves 91.26%. *SSPL* [25] use a large pretrained model model but only achieves 91.77%. *BLAN* [34] uses a bidirectional structure and a multiscale approach and achieves 91.80%. *HFE* [33] combine attribute and ID information and achieves 92.17%. The accuracy are mostly lower than 92%. However, none of these methods can solve multi-scale problems. Our method can dynamically fuse relevant information using multi-scale information, which is an improvement compared to the previous methods and achieves a higher accuracy of 92.47%.

Table 1. Comparison of mean accuracy on CelebA [16] dataset.

Method	CelebA [16]
MOON [22]	90.94
Adaptively Weighted [9]	91.80
MCNN-AUX [6]	91.26
Adaptive Sharing [17]	91.26
GAN and Dual-path [8]	91.81
Autoencoder [24]	90.14
Deep Multi-task [19]	91.70
HFE [33]	92.17
BLAN [34]	91.80
SSPL [25]	91.77
ours	**92.47**

4.5 Ablution Study.

TABLE 2 indicates the degree of contribution of each module in the whole network. In this experiment, the baseline is the pure Swin-Transformer [14]. With the multi-scale fusion module (MFM), the accuracy increase by about 0.26%. It can gradually combine low-level local features with high-level global features, which affect a lot on the final classification results. Then we map the prior

information of attributes into image features to extract sufficient information about the related attributes and the module increase accuracy by 0.16%. When using these two sub-modules at the same time, the accuracy improves by 0.41%, which achieves the relatively best.

Table 2. Ablution study on CELEBA [16] dataset with our method and ResNet50 [7] backbone.

Method	acc.	Method	acc.
baseline	92.06	ResNet50	91.90
+ MFM	92.32	+ MFM	92.15
+ ARMM	92.22	+ ARMM	92.00
+ MFM + ARMM	**92.47**	+ MFM + ARMM	**92.23**

In order to demonstrate the superiority of Swin-Transformer [14] over ResNet50 [7] and show the effectiveness of our method. We also take the ResNet50 [7] as the backbone to experiment. TABLE 2 shows the effect of the same module on ResNet50 [7]. The multi-scale module and the prior information of attributes improve the accuracy by 0.25% and 0.10% respectively and it also achieves the relatively best when using both the two modules at the same time. Based on the two results with different backbones, we find that Swin-transformer can learn the relevance of spatial features, that is, the global relationship between features.

In TABLE 3, we can find that our adaptive attention method is more effective than directly concatenating or weighed summing the features. Then with the adaptive attention method, the network can focus on the related features between adjacent layers, which can recognize attributes at different scales. However, the methods of directly concatenating or weighed summing are the static method, which may fuse unnecessary information.

Table 3. Comparison experiment between Adaptive Attention method, Concat method and Weighted Sum method.

Method	acc.
Weighted Sum	92.28
Concat Directly	92.32
Adaptive Attention	**92.47**

4.6 Qualitative Evaluation

As shown in Fig. 4, we discover that ResNet50 [7] can only pay attention to a whole local area, which produces a certain offset. And our method can make the

Table 4. Classification accuracy (%) of Ours (92.47%), BLAN (91.81%) [34] and MCNN-AUX (91.26%) [6] on CelebA [16] over 40 facial attributes.

Attributes	Ours	BLAN	MCNN-AUX	Attributes	Ours	BLAN	MCNN-AUX
5 o'clock Shadow	**95.19**	95.18	94.51	Male	**99.13**	98.32	98.17
Arched Eyebrows	**87.71**	84.74	83.42	Mouth Slightly Open	**94.47**	94.22	93.74
Attractive	82.63	**83.25**	83.06	Mustache	**97.04**	96.99	96.88
Bags Under Eyes	**86.15**	86.11	84.92	Narrow Eyes	**94.00**	87.78	87.23
Bald	**99.06**	99.02	98.90	No Beard	**96.64**	96.46	96.05
Bangs	**96.35**	96.26	96.05	Oval Face	**77.59**	76.86	75.54
Big Lips	**83.31**	72.59	71.47	Pale Skin	96.65	**97.25**	97.05
Big Nose	**85.30**	85.21	84.53	Pointy Nose	**78.38**	78.02	77.47
Black Hair	**91.91**	90.49	89.78	Receding Hairline	**95.11**	93.99	93.81
Blond Hair	95.65	**96.27**	96.01	Rosy Cheeks	**95.41**	95.36	95.16
Blurry	**96.82**	96.37	96.17	Sideburns	97.56	**98.04**	97.85
Brown Hair	85.92	**89.79**	89.15	Smiling	**94.00**	93.19	92.73
Bushy Eyebrows	**93.11**	93.08	92.84	Straight Hair	**85.85**	84.65	83.58
Chubby	**95.99**	95.88	95.67	Wavy Hair	**87.37**	85.35	83.91
Young	**89.09**	89.06	88.48	Necktie	**97.33**	97.20	96.51
Necklace	**89.78**	88.16	86.63	Lipstick	**94.40**	94.34	94.11
Hat	**99.17**	99.15	99.05	High Cheekbones	**89.36**	88.13	87.58
Heavy Makeup	**92.96**	92.04	91.55	Gray Hair	98.07	**98.35**	98.20
Goatee	97.06	**97.69**	97.24	Eyeglasses	**99.63**	99.63	99.63
Earrings	**91.38**	90.93	90.43	Double Chin	**96.82**	96.58	96.32

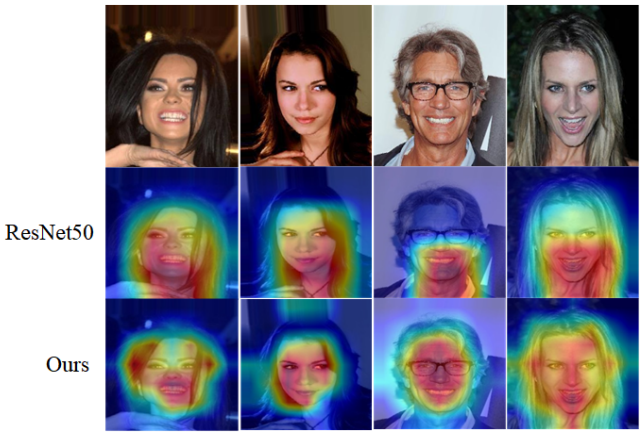

Fig. 4. We use Grad-Cam [23] to show the area network pay attention. The parts with warm colors represent where the network pays attention. The second and the third line show the results on the input images with ResNet50 [7] and our framework respectively.

network focus on some characteristic areas and the details are better detected. Additionally, our method pays more attention to the face area and may not be affected too much by the background part. These clearly demonstrate the effectiveness of the proposed solution.

As shown in Fig. 5, it can be seen from the figure that attributes about *Male*, *Mouth_Slightly_Open*, *No_Beard*, *Smiling*, and *young* have strong connections with other attributes. And these attributes in turn correspond to the sequence numbers 20, 21, 24, 31 and 39 in the figure. For example, at row 21 and column 31, the dark blue shows a strong connection between *Smiling* and *Mouth_Slightly_Open*.

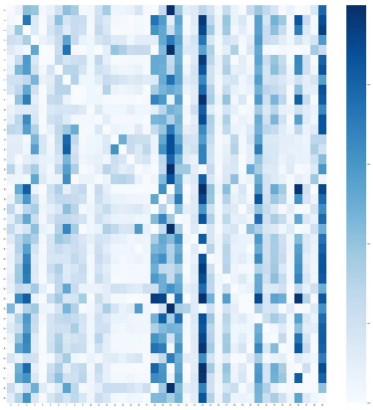

Fig. 5. The visualization of the prior relationship between attributes. The duck blue means the strong relation and the light blue means the weak relation. The serial number represents the corresponding attribute. Since the same attribute does not appear twice in labels of a picture, the diagonal line is not highlighted.

4.7 Quantitative Evaluation

TABLE 4 reports the accuracy of each attribute in CelebA [16]. We take three levels of models to make predictions, which are MCNN-AUX 91.26% [6], ours 92.47% and BLAN 91.81% [34], because the accuracy of each attribute is not reported in most previous methods. At the accuracy of attributes *Arched Eyebrow*, *Big Nose*, *High Cheekbone*, *Narrow Eye*, *Earring* and *Necklac*, ours has an advantage of more than 2%. Facts have proved that our method can have good results in recognizing some small attributes. And ours still maintain good accuracy in attributes with large scale. Compared to BLAN [34], the accuracy of the attribute *Big lips* is improved by about 11%. On the other hands, by performing multi-scale fusion reasoning, the proposed method has recognized more detailed attributes while making fewer mistakes and the accuracy of other attributes does not fluctuate greatly.

5 Conclusion

In this paper, we propose a prior-guided multi-scale fusion transformer for face attribute recognition to capture long-distance relations of features in spatial in an end-to-end manner. We also introduce a learnable weight to perform effective soft selection of adjacent features and apply a hierarchical approach to fuse them, which can get enough information to predict attributes at different scales. Additionally, we introduce prior information of attributes into feature learning, which can make the network focus on correlations between attributes. Extensive experimental results on a real dataset demonstrate the effectiveness and the generalization ability of our method in dealing with face attribute recognition.

References

1. Cao, J., Li, Y., Zhang, Z.: Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4290–4299 (2018)
2. Chen, B.C., Chen, Y.Y., Kuo, Y.H., Hsu, W.H.: Scalable face image retrieval using attribute-enhanced sparse codewords. IEEE Transactions on Multimedia **15**(5), 1163–1173 (2013). https://doi.org/10.1109/TMM.2013.2242460
3. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5177–5186 (2019)
4. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. arXiv preprint arXiv:2107.00652 (2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Hand, E.M., Chellappa, R.: Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
8. He, K., Fu, Y., Zhang, W., Wang, C., Jiang, Y.G., Huang, F., Xue, X.: Harnessing synthesized abstraction images to improve facial attribute recognition. In: IJCAI. pp. 733–740 (2018)
9. He, K., Wang, Z., Fu, Y., Feng, R., Jiang, Y.G., Xue, X.: Adaptively weighted multi-task deep network for person attribute classification. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1636–1644 (2017)
10. Kalayeh, M.M., Gong, B., Shah, M.: Improving facial attribute prediction using semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

13. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning. pp. 1188–1196. PMLR (2014)
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
15. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (December 2015)
16. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)
17. Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., Feris, R.: 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. pp. 5334–5343 (2017)
18. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. Proc. icml (2013)
19. Mao, L., Yan, Y., Xue, J.H., Wang, H.: Deep multi-task multi-label cnn for effective facial attribute classification. IEEE Transactions on Affective Computing (2020)
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
21. Nian, F., Chen, X., Yang, S., Lv, G.: Facial attribute recognition with feature decoupling and graph convolutional networks. IEEE Access **7**, 85500–85512 (2019)
22. Rudd, E.M., Günther, M., Boult, T.E.: Moon: A mixed objective optimization network for the recognition of facial attributes. In: European Conference on Computer Vision. pp. 19–35. Springer (2016)
23. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
24. Sethi, A., Singh, M., Singh, R., Vatsa, M.: Residual codean autoencoder for facial attribute analysis. Pattern Recognition Letters **119**, 157–165 (2019)
25. Shu, Y., Yan, Y., Chen, S., Xue, J.H., Shen, C., Wang, H.: Learning spatial-semantic relationship for facial attribute recognition with limited labeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11916–11925 (2021)
26. Tao, A., Sapra, K., Catanzaro, B.: Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821 (2020)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
28. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvtv2: Improved baselines with pyramid vision transformer. arXiv preprint arXiv:2106.13797 (2021)
29. Wang, X., Ye, Y., Gupta, A.: Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6857–6866 (2018)
30. Wang, Y., He, D., Li, F., Long, X., Zhou, Z., Ma, J., Wen, S.: Multi-label classification with label graph superimposing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12265–12272 (2020)

31. Wang, Z., He, K., Fu, Y., Feng, R., Jiang, Y.G., Xue, X.: Multi-task deep neural network for joint face recognition and facial attribute prediction. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. pp. 365–374 (2017)
32. Wu, L., Sun, P., Hong, R., Fu, Y., Wang, X., Wang, M.: Socialgcn: An efficient graph convolutional network based model for social recommendation. arXiv preprint arXiv:1811.02815 (2018)
33. Yang, J., Fan, J., Wang, Y., Wang, Y., Gan, W., Liu, L., Wu, W.: Hierarchical feature embedding for attribute recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13055–13064 (2020)
34. Zheng, X., Huang, H., Guo, Y., Wang, B., He, R.: Blan: Bi-directional ladder attentive network for facial attribute prediction. Pattern Recognition **100**, 107155 (2020)
35. Zhong, Y., Sullivan, J., Li, H.: Leveraging mid-level deep representations for predicting face attributes in the wild. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 3239–3243. IEEE (2016)
36. Zhu, Z., Yu, J., Wang, Y., Sun, Y., Hu, Y., Wu, Q.: Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. arXiv preprint arXiv:2006.09073 (2020)