

Attribute-Aware Convolutional Neural Networks for Facial Beauty Prediction

Luojun Lin¹, Lingyu Liang¹, Lianwen Jin^{1*} and Weijie Chen²

¹School of Electronic and Information Engineering, South China University of Technology, China

²Hikvision Research Institute, China

linluojun2009@126.com, lianglysky@gmail.com, lianwen.jin@gmail.com, chenweijie5@hikvision.com

Abstract

Facial beauty prediction (FBP) aims to develop a machine that automatically makes facial attractiveness assessment. To a large extent, the perception of facial beauty for a human is involved with the attributes of facial appearance, which provides some significant visual cues for FBP. Deep convolution neural networks (CNNs) have shown its power for FBP, but convolution filters with fixed parameters cannot take full advantage of the facial attributes for FBP. To address this problem, we propose an *Attribute-aware Convolutional Neural Network* (AaNet) that modulates the filters of the main network, adaptively, using parameter generators that take beauty-related attributes as extra inputs. The parameter generators update the filters in the main network in two different manners: filter tuning or filter rebirth. However, AaNet takes attributes information as prior knowledge, that is ill-suited to those datasets merely with task-oriented labels. Therefore, imitating the design of AaNet, we further propose a *Pseudo Attribute-aware Convolutional Neural Network* (P-AaNet) that modulates filters conditioned on global context embeddings (pseudo attributes) of input faces learnt by a lightweight pseudo attribute distiller. Extensive ablation studies show that the AaNet and P-AaNet improve the performance of FBP when compared to conventional convolution and attention scheme, which validates the effectiveness of our method.

1 Introduction

Since the golden rules in the era of Leonardo Da Vinci, decoding the human perception of facial beauty has been a significant research. Recently, many efforts have been devoted to facial beauty prediction (FBP) based on data-driven computation [Zhang *et al.*, 2016; Liang *et al.*, 2018; Lin *et al.*, 2018]. FBP is an essential component for various applications, such as face beautification [Li *et al.*, 2015], makeup recommendation [Liu *et al.*, 2014], and personal social recommendation [Rothe *et al.*, 2016], to name a few.

*Corresponding author: Lianwen Jin.

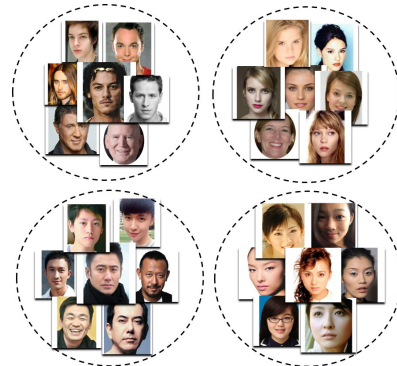


Figure 1: Faces with different attributes from the SCUT-FBP5500 database, which can be grouped into Caucasian/Asian male/female. It can be observed that the faces with the same attributes share appearance similarity but vary greatly across different attributes.

To a large extent, the perception of facial beauty for a human is involved with the attributes of facial appearance, such as gender, race, and age. For example, male and female faces are quite different in shape and skin texture because of sexual dimorphism [Perrett *et al.*, 1998]. In other words, faces with the same attributes usually share some common patterns in appearance and tend to be within the same cluster, e.g., male faces are more likely to grow a beard, and Caucasian faces usually have lighter skin tones. In contrast, faces with different attributes tend to be in different clusters, as shown in Fig.1. For intuitive illustration, we adopt t-SNE to visualize the feature distribution of a conventional convolutional neural network (CNN), namely AlexNet [Krizhevsky *et al.*, 2012], which is trained on SCUT-FBP5500 for FBP. The visualization is also adopted for the AaNet, as shown in Fig.2. We can observe that the features of faces with the same gender or ethnicity are inclined to be in the same cluster, which indicates that facial attributes can be good visual cues to guide FBP.

Previous CNN-based methods for FBP always used a fixed set of filters to learn the mapping from a raw image to a beauty score, without explicitly considering facial attributes. To further improve the performance, one simple solution is to ensemble several CNN models, which are trained separately for the specific attribute. However, this may lead to considerable complexity of model and computation. Therefore, it would be significant to develop a CNN-component

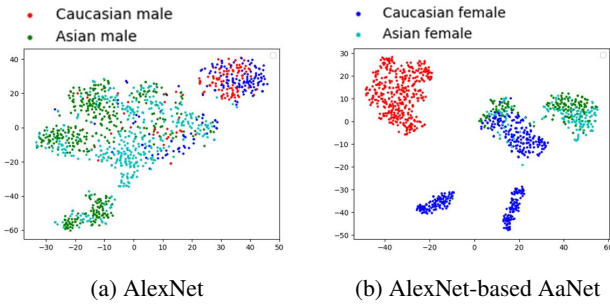


Figure 2: t-SNE is adopted to visualize the distribution of the deep features (the outputs of the last pooling layer) of the trained FBP models such as AlexNet (a) and AaNet (b). It shows that faces with the same attributes tend to be within the same cluster, and vice versa. The attribute-clustering property is more obvious for AaNet because of the explicit introduction of attributes information.

that is adaptive for different attributes.

In this paper, we propose an *Attribute-aware Convolutional Neural Network (AaNet)*, whose filter parameters is controlled adaptively by facial attributes. The framework of AaNet is demonstrated in Fig.3. The main network contains multiple *adaptive convolutional layers*, whose weights are updated by *parameter generators* via taking the attributes embedding as input. The embedding is generated by a shallow embedding network conditioned on attributes information (e.g., gender and race). Consequently, AaNet can adjust filter parameters and generate features based on the attributes, which can handle the facial variations across different attributes. Furthermore, we propose two different filter updating schemes, called filter tuning and filter rebirth, for adaptive convolutional layers. Filter tuning modulates the filters by generating residuals adding to the original filters, while filter rebirth discards the original filters and generates new filters in each feedforward.

The AaNet depends on the given attribute information to generate adaptive filters. However, it is ill-suited to those datasets that have merely task-oriented labels and no extra attribute label. To this end, we further propose a *Pseudo Attribute-aware Convolutional Neural Network (P-AaNet)* that modulates the filter parameters conditioned on each input face itself; and thus, it does not require related attributes as prior knowledge. The P-AaNet architecture is designed by mimicking AaNet after introducing a newly lightweight subnetwork called *pseudo attributes distiller*. The distiller extracts attribute-like knowledge for each input that guides the parameter generators to adjust the filters in the main network, as shown in Fig.4. Actually, P-AaNet can be considered a special case of AaNet, with a self-adaptive mechanism to solve the variations brought by each input face.

We verify the effectiveness of the proposed methods using several FBP benchmark datasets. We conducted extensive experiments on the SCUT-FBP5500 [Liang *et al.*, 2018] database to explore the properties of AaNet and P-AaNet. The results illustrate the effectiveness of each component of our networks under different prior knowledge, embedding dimensions, filter updating schemes and network topology. We

also made comparison with effective self-adaptive method, *squeeze and excitation* module [Hu *et al.*, 2018], for verification, and the results indicate the superiority of our adaptive convolution to previous method. Finally, we compare our networks with the state-of-the-art methods for FBP, which shows the effectiveness of our approach.

The contributions of this paper are summarized as follows:

- We propose attribute-aware convolutional neural network (AaNet) to improve the performance for FBP, which uses facial attributes to adaptively update convolutional weights by parameter generators.
- For the cases without explicit attribute information, we propose Pseudo AaNet (P-AaNet) to update the filters based on the input face itself, where *pseudo attribute distiller* extracts the context embedding of each face to determine convolutional weights.
- Extensive experiments were conducted on on two FBP datasets for AaNet and P-AaNet, and the results illustrate the superiority of our method to previous schemes like conventional convolution or attention mechanisms.

2 Related Work

Facial beauty prediction. Based on the classic pattern recognition process, FBP has countered early success by the combination of hand-crafted features with shallow predictors. The hand-crafted features include the geometric features (e.g., geometric ratios and landmark distances) [Aarabi *et al.*, 2001; Zhang *et al.*, 2011; Chen and Zhang, 2014] and textural features (e.g., LBP-/Gabor-/SIFT-like features) [Zhang *et al.*, 2016; Ren and Geng, 2017]. However, these hand-crafted features are low-level features that are difficult to obtain discriminative facial representation. Recent years, with the rapid development of deep learning, more and more researchers have been using CNN to access facial beauty automatically. Because of the hierarchical nonlinear transformation, the CNN-based FBP models [Gray *et al.*, 2010; Xie *et al.*, 2015; Xu *et al.*, 2017; Liang *et al.*, 2017] have been proved to be superior to the previous traditional methods.

Dynamic convolution. Dynamic convolution is a mechanism in CNNs where filter parameters are generated dynamically by meta-networks instead of being learnt directly. There are several works related to dynamic convolution. One research [Bertinetto *et al.*, 2016] proposed a one-shot learner to predict the parameters of a pupil network from a single exemplar. A subsequent work [Ha *et al.*, 2016] proposed the utilization of a hypernetwork to generate filters for convolutional network and non-shared weights for recurrent network. The hypernetwork allows weight-sharing within a layer and across layers of the main network, with the aim to reduce abundant parameters in CNN. Most relevant to our study is the dynamic filter network [De Brabandere *et al.*, 2016] designed for image prediction tasks (e.g., predicting the next frame image from the previous one), which involves temporal information computation. However, our work focuses on analyzing and solving the pattern variation problem brought by facial attributes. This is an important topic in facial aes-

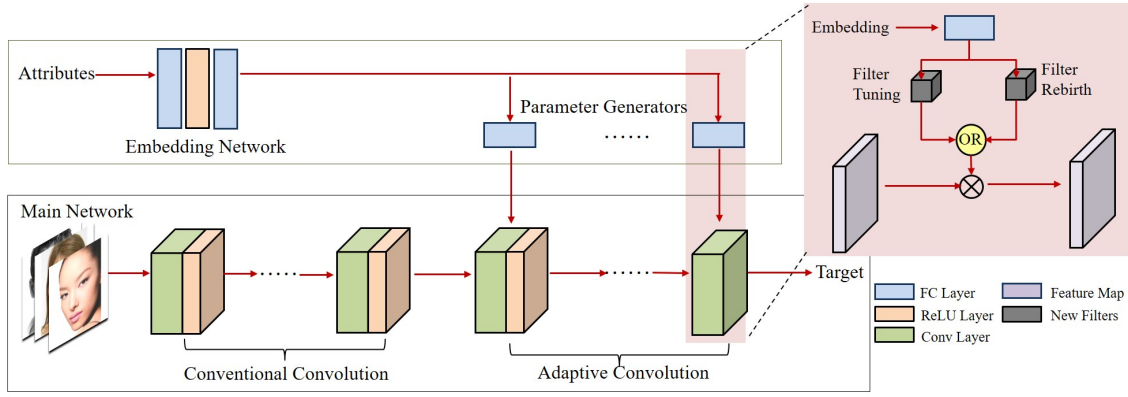


Figure 3: Architecture of attribute-aware convolutional neural network (AaNet). It consists of a main network, an embedding network and parameter generators. The adaptive convolutional layers in the main network update filter weights by the parameter generators via taking the output of the embedding network as input. There are two kinds of filter updating manners: filter tuning or filter rebirth.

thetic computation. To the best of our knowledge, it is the first time we develop guidelines and frameworks in this field.

Attention. Attention can be considered a self-adaptive mechanism that assigns most of the computing resources towards the most informative part of the input signal. With such a property, this mechanism has made significant success in various fields, from the sequence-based models such as machine translation [Vaswani *et al.*, 2017] and lip reading [Chung *et al.*, 2017], to the vision tasks like image classification [Wang *et al.*, 2017]. The squeeze and excitation (SE) module is a representative attention method that models channel-wise relationships between the output neurons of each convolutional layer [Hu *et al.*, 2018]. In this paper, we take this self-adaptive mechanism as a reference to further confirm the effectiveness of our attribute-aware convolution.

3 Attribute-Aware CNN

3.1 Framework

Using facial attributes as prior knowledge, the neural networks should focus on learning attribute-related features to predict the beauty score adaptively. Therefore, we propose an *Attribute-aware CNN (AaNet)* that uses the related-attributes label as extra input to modulate the filters to adapt to variations caused by different attributes. The AaNet architecture is shown in Fig.3. It contains the following main components: a main network, an embedding network and parameter generators. The main network is designed for facial beauty prediction, whereas the weights of its adaptive convolutional layers are produced by the parameter generators. More details of AaNet are provided as below.

Parameter Generator

The parameter generating methods can be divided into two manners (filter tuning and filter rebirth) according to whether a common part of the parameters is shared across inputs with different attributes or not.

Filter tuning generates residuals to add to the original filters of the adaptive convolutional layers in the main network,

with the aim to modulate the filters towards the attributes-specific ones. Given attributes I_{attr} as input, the filter tuning of parameter generator can be formulated as:

$$W_n^l = W_o^l + \mathcal{G}^l(\mathcal{E}(I_{attr})) \quad (1)$$

where W_o^l and W_n^l separately denote the original parameters and the updated new parameters in the l_{th} adaptive convolutional layer, and $\mathcal{G}^l(\mathcal{E}(I_{attr}))$ denotes their residuals term. Here $\mathcal{G}^l(\cdot)$ is the l_{th} parameter generator and $\mathcal{E}(\cdot)$ is an embedding network shared by all parameter generators. In this paper, the embedding network is designed as a multi-layer perception (MLP) network to produce a 1-dim embedding based on attributes information, while each parameter generator is implemented as fully-connected (FC) layer.

Filter rebirth discards the original filters and directly generates brand-new filters for each adaptive convolutional layer per feedforward. The process of filter rebirth can be formulated as:

$$W_n^l = \mathcal{G}^l(\mathcal{E}(I_{attr})) \quad (2)$$

Compared to filter tuning, filter rebirth can decouple facial beauty with related attributes more thoroughly in high dimensional nonlinear spaces, reducing the influences caused by attributes and leading to easier optimization of the network, which is further discussed in the experiments.

Adaptive Convolutional Layer

Compared with the fixed filters of conventional convolution, the filters of the adaptive convolutional layers are updated dynamically by the corresponding parameter generators mentioned above. Given the feature maps F_{in}^l as input, the l_{th} adaptive convolutional layer convolves on the input to generate an output F_{out}^l :

$$F_{out}^l = F_{in}^l \otimes W_n^l \quad (3)$$

Main Network

We stack L_c conventional convolutional layers parameterized with $(W_c^1, \dots, W_c^{L_c})$ in the shallow layers to extract basic features and stack L_n adaptive convolutional layers in the deeper layers to extract adaptive semantic features for FBP (Batch-Norm layers [Ioffe and Szegedy, 2015], ReLU layers and

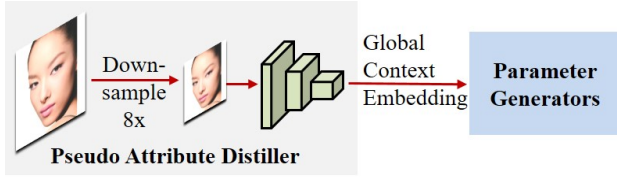


Figure 4: Different from AaNet, the parameter generators of P-AaNet take self-guided global context embedding as input instead of external attribute information.

pooling layers are omitted here for the sake of simplification). Therefore, given a facial image I_{img} as input, the main network outputs a beauty score:

$$y = \mathcal{F}(I_{img}; W = \{W_c^1, \dots, W_c^{L_c}, W_n^1, \dots, W_n^{L_n}\}), \quad (4)$$

Using gradient back-propagation, the weights of the parameter generators can be learnt jointly with the main network in an end-to-end manner. In the inference stage, the parameter generators keep fixed weights, whereas, the main network updates the weights according to the given attributes.

Further Discussion

Guided by the attributes information, the AaNet can modulate convolutional parameters adaptively for different attributes. To some extent, it is equivalent to an ensemble of several conventional CNNs separately trained on each attribute pattern. However, it is much beyond the simple ensemble CNN, owing to its weight-sharing across different attributes. The parameter generators can summarize the common knowledge and reserve the differences across attributes. As a consequence, our AaNet is much more flexible in adapting to each attribute pattern for FBP, with much less parameters than the ensemble CNN.

3.2 Pseudo Attribute-Aware CNN

As mentioned above, the AaNet strongly depends on using external attribute information as a conditional input to guide the adaptive filter learning. Unfortunately, there exist many datasets that are ill-suited to AaNet because of the lack of exact attribute information. To solve this problem, we propose a variant of the AaNet, namely *Pseudo Attribute-aware CNN (P-AaNet)*, without using any external prior knowledge. We exploit the information of the input image through a lightweight network (pseudo attribute distiller) which produces attribute-like knowledge to guide the parameter generators producing self-adaptive filters, as shown in Fig.4.

Pseudo Attribute Distiller

To imitate the mechanism of AaNet under the situation without prior knowledge, we design a lightweight network to extract the global context embedding from each image. The embedding is extremely low-dimensional, and thus, can be considered as a pseudo attribute (e.g. 1-dim is enough to summarize the context information of the input image, leading to no more parameters increase in P-AaNet with filter rebirth than the conventional CNN, which will be further discussed in experiments). In this way, the l_{th} parameter generator can

be formulated as:

$$W_n^l = \lambda W_o^l + \mathcal{G}^l(\mathcal{D}(I_{img})) \quad (5)$$

where λ is a fixed boolean variable which controls the switch between filter tuning and filter rebirth. $\mathcal{D}(\cdot)$ denotes the pseudo attribute distiller consisting of a down-sampling layer and three convolutional layers followed by BatchNorm and ReLU activation.

Actually, P-AaNet can be considered as a special case of AaNet, with each input image regarded as a specific attribute that solves the variation across images. Hence, the network can be adapted to each input image without the need for prior annotated attributes. This special property of P-AaNet makes it easier to extend to other scenarios.

4 Experiments

We conduct extensive experiments to evaluate the performance and explore the properties of AaNet and P-AaNet mainly on the SCUT-FBP5500 benchmark dataset. Several detailed ablation studies on this benchmark are carried out to justify the effectiveness of our proposed networks. To further stress the superiority, we also compare our method with related methods on the SCUT-FBP5500 and SCUT-FBP datasets, and the results show that our method achieves state-of-the-art performance on these benchmarks.

4.1 Experimental Benchmarks and Settings

All the experiments are carried out on Caffe [Jia *et al.*, 2014] with a NVIDIA Geforce GTX Titan X GPU. To ensure the effectiveness, five-folds cross validation is performed. The average results of the validations are reported below.

Datasets and evaluation protocols. Most of our experiments are conducted on the SCUT-FBP5500 dataset [Liang *et al.*, 2018], which contains 5500 facial images with diverse attributes (e.g., male/female, Asian/Caucasian) and diverse labels (e.g., facial landmarks, beauty scores). The attributes and facial landmarks can be employed as prior knowledge to guide adaptive filter learning of AaNet. Additionally, we also conduct experiments on the SCUT-FBP dataset [Xie *et al.*, 2015] which contains 500 facial images sampled from Asian female subject. Due to the lack of attribute labels, only P-AaNet is evaluated on this benchmark. The beauty scores of these datasets range between [1,5], which indicates that FBP could be formulated as a regression problem. Hence, the Pearson correlation (PC), mean absolute error (MAE) and root mean squared error (RMSE) are utilized to evaluate the regression performance of our method; a high PC, small MAE and RMSE indicate better performance.

Implementation details. All the facial images (350×350) are resized to 256×256 firstly. Then a 224×224 crop and horizontal flipping are performed randomly, followed by per-pixel rescale to [0,1] and mean value subtraction. In the following experiments, AaNet takes facial images and their corresponding attributes as inputs, while the others use the facial images alone. For AlexNet and its related networks, they are trained by using mini-batch Stochastic Gradient Descent (SGD) with a batch size of 32, a momentum of 0.9, and a weight decay of $5e-4$. We use a specific learning policy that

the learning rate is increased from 0 to a peak value of 0.01 in a warm-up schedule of $2K$ iterations, and then decreased to 0, linearly, in $18K$ iterations. Note that, for P-AaNet, the learning rate of pseudo attribute distiller is 0.1 times as that of the main network in order to ensure stable training. For ResNet-18 and its extension networks, we set the peak value of learning rate and weight decay as 0.1 and $1e-4$.

4.2 Ablation Studies

In order to investigate the property of each components of our networks more clearly, a simple but representative CNN, AlexNet, is chosen to implement our ablation studies. It is taken as a 7-layer CNN in this paper (note that FC layers are considered as convolutional layers). We use its variant as the baseline, where the local response normalization is removed and each convolutional layer is equipped with BatchNorm to facilitate the convergence. In our experiments, this modified AlexNet is set as the default backbone network of AaNet and P-AaNet with filter rebirth as the default filter updating manner, unless specifically stated.

However, not each convolutional layer in the main network is suitable to be attribute-aware. We conduct a series of experiments on AaNet and P-AaNet that increases the number of the adaptive convolutional layers from higher layer to shallow layer gradually, and empirically find that the main network achieves the highest performance equipped with the first layer as a fixed convolutional layer and the remaining layers as adaptive convolutional layers. In the following, we conduct extensive ablation studies to explore the properties of AaNet and P-AaNet with this setting.

Different prior knowledge for AaNet. To explore the effect of different prior knowledge on FBP, we use the embedding of binarized gender and race labels to guide the adaptive filter learning in AaNet, and also adopt the geometric ratios as prior knowledge for comparison. Specifically, the geometric ratios are determined from facial landmark distances according to the heuristic rules [Chen and Zhang, 2014]. Furthermore, for fair comparison, we introduce another traditional integration manner which directly concatenates geometric ratio features with the RGB channels of raw images to feed into AlexNet for FBP. For the sake of dimension consistency during concatenation, we decode the geometric ratios into a high-dimensional feature map with the same spatial size as the raw image through a deconvolutional neural network. We refer to it as C-AlexNet for short. The comparison results on the SCUT-FBP5500 dataset are shown in Table 1, from which we can draw the following conclusions: 1) AaNet outperforms C-AlexNet in terms of different auxiliary inputs, which indicates that adaptive convolution can better exploit the prior knowledge than direct concatenation; 2) The performances of both C-AlexNet and AaNet with gender and race as prior knowledge are superior to that with geometric ratios, which suggests that gender and race can provide more informative cues for FBP than geometric ratios.

Different sizes of global embedding for P-AaNet. For P-AaNet, the dimension of global context embedding influences the parameter number of the whole network easily, which has further impact on the practical storage. In this section, we in-

Network	Attributes	PC	MAE	RMSE
AlexNet	-	0.8534	0.2768	0.357
C-AlexNet	race, gender	0.8589	0.2721	0.3509
C-AlexNet	geometric ratios	0.8536	0.2737	0.3557
AaNet	race, gender	0.8842	0.243	0.3196
AaNet	geometric ratios	0.8774	0.252	0.3297

Table 1: Comparison among AlexNet, C-AlexNet and AaNet with different prior knowledge on SCUT-FBP5500 dataset.

Network	Size	PC	MAE	RMSE	Params.
AlexNet	-	0.8534	0.2768	0.3571	7.051M
P-AaNet	1×1	0.8758	0.2512	0.3304	7.058M
P-AaNet	2×2	0.8778	0.2526	0.3271	28.10M
P-AaNet	3×3	0.8752	0.2556	0.3308	63.19M

Table 2: Comparison of P-AaNet with different size of global context embeddings on SCUT-FBP5500 dataset.

Network	Operation	PC	MAE	RMSE
AlexNet	-	0.8534	0.2768	0.3571
SE-AlexNet	channel attention	0.8607	0.2688	0.3486
AaNet	filter rebirth	0.8842	0.243	0.3196
P-AaNet	filter rebirth	0.8758	0.2512	0.3304
AaNet	filter tuning	0.8728	0.2601	0.3342
P-AaNe	filter tuning	0.8666	0.2645	0.3408

Table 3: Comparison among AlexNet, SE-AlexNet, AaNet and P-AaNet implemented with different filter updating manners on SCUT-FBP5500 dataset.

Network	Backbone	PC	MAE	RMSE
AlexNet	AlexNet	0.8534	0.2768	0.3571
AaNet	AlexNet	0.8842	0.243	0.3196
P-AaNet	AlexNet	0.8758	0.2512	0.3304
ResNet	ResNet-18	0.89	0.2419	0.3166
AaNet	ResNet-18	0.9055	0.2236	0.2954
P-AaNet	ResNet-18	0.8965	0.2285	0.3035

Table 4: Comparison among AaNet, P-AaNet implemented with different backbone networks (AlexNet, ResNet-18) on SCUT-FBP5500 dataset.

investigate the effect of different embedding size, by changing the outputs of the pseudo attributes distiller from 1×1 to 2×2 and 3×3 . The comparison results are shown in Table 2 where we can see that P-AaNets of any embedding size always outperform the baseline, which validates the effectiveness of the self-adaptive convolution. Moreover, the performance of P-AaNet increases with the larger size of global context embedding until the size reaches 3×3 . However, the slight effect boost is at the expense of a rapid growth of parameter count. To make a trade-off between efficiency and performance, we use 1×1 as our default setting, where the parameter count is maintained nearly the same as the baseline.

Methods on SCUT-FBP dataset	PC	MAE	RMSE
Hybrid handcrafted features + Gaussian regress [Xie <i>et al.</i> , 2015]	0.6482	0.3931	0.5149
6-layers CNN [Xie <i>et al.</i> , 2015]	0.8187	-	-
Region aware scattering CNN-based features + SVR [Liang <i>et al.</i> , 2017]	0.83	-	-
LBP/HOG/Gabor features + Structured label distribution learning (LDL) [Ren and Geng, 2017]	-	0.3015	0.4076
6-layers Psychological inspired CNN [Xu <i>et al.</i> , 2017]	0.854	-	-
ResNeXt-50 based R ² -ResNeXt [Lin <i>et al.</i> , 2018]	0.8957	0.2416	0.3046
ResNet-18 based P-AaNet (ours)	0.9103	0.2224	0.2816

Methods on SCUT-FBP5500 dataset	PC	MAE	RMSE
ResNeXt-50 [Liang <i>et al.</i> , 2018]	0.8997	0.2291	0.3017
ResNet-18 based AaNet (ours)	0.9055	0.2236	0.2954
ResNet-18 based P-AaNet (ours)	0.8965	0.2285	0.3035

Table 5: Comparison with other state-of-the-art approaches in terms of PC, MAE and RMSE.

Method	PC	MAE	RMSE
AlexNet	0.8534	0.2768	0.3571
AaNet	0.8842	0.243	0.3196
P-AaNet	0.8758	0.2512	0.3304

AlexNet + LDL	0.8687	0.2571	0.3356
AaNet + LDL	0.8881	0.2425	0.3148
P-AaNet + LDL	0.8804	0.2498	0.3254

Table 6: Comparison among AlexNet, AaNet, and P-AaNet combined with LDL on SCUT-FBP5500 dataset.

Different filter updating manners. We explore the effect of the two different filter updating manners (filter tuning and filter rebirth) on AaNet and P-AaNet respectively. For further validation, we also implement AlexNet equipped with SE module (SE-AlexNet) as another baseline, which is a very efficient self-adaptive mechanism based on channel-wise attention. As shown in Table 3, we can see that: 1) The adaptive convolution performs better than the SE module; 2) Under the same filter updating conditions, AaNet outperforms P-AaNet because of the introduction of prior knowledge; 3) For both AaNet and P-AaNet, the performance of filter rebirth is superior to that of filter tuning. We infer that filter rebirth can decouple facial beauty with attributes more thoroughly, leading to easier optimization. Therefore, we use it as the default filter updating method in this paper.

Different backbone networks. To stress the effectiveness of our method on different network architectures, we implement AaNet and P-AaNet with another popular architecture family, namely, Residual Neural Networks [He *et al.*, 2016], among which ResNet-18 is taken as a representative for our following experiments. As shown in Table 4, the proposed adaptive convolution also works well on ResNet-18, and the ResNet-18 based AaNet provides the best performance.

4.3 Comparison with State-of-the-Art Methods

We also compare our method with other state-of-the-art approaches on the SCUT-FBP and SCUT-FBP5500 datasets. Since we have obtained the best result with ResNet-18 based AaNet in the filter rebirth manner, this setting is maintained

in the following comparisons. However, due to the absence of extra attribute labels, only P-AaNet is evaluated on the SCUT-FBP benchmark. As shown in Table 5, our P-AaNet and AaNet achieve state-of-the-art performances in terms of various metrics. It is worth noting that we do not use ImageNet [Deng *et al.*, 2009] pretrained model in any of our experiments, unlike most of previous FBP approaches [Xu *et al.*, 2017; Lin *et al.*, 2018; Liang *et al.*, 2018]. Moreover, our ResNet-18 based P-AaNet can obtain comparable even superior performance to ResNeXt-50 with less than half of parameters, which exactly validates the superiority of P-AaNet.

Additionally, we reproduce another representative FBP method on SCUT-FBP5500, namely label distribution learning (LDL) [Ren and Geng, 2017], which is equipped with AlexNet, AlexNet-based AaNet and P-AaNet severally for further comparisons. The results on Table 6 indicate that our method is orthogonal to other approaches which can be combined to make a further performance boost.

5 Conclusion

In this paper, we propose an adaptive convolution framework for FBP, which includes: 1) an attribute-aware network to take full advantage of attributes as prior knowledge; 2) a pseudo attribute-aware network to utilize image context information to generate attribute-like knowledge under the cases without attribute label. Extensive ablation studies confirm the effectiveness of the two networks. More than facial beauty prediction, we also draw some psychological conclusions about facial beauty from the experiments. For instance, we discover that gender and race may have stronger impact on facial attractiveness than geometric ratios. This is an interesting direction for the future explorations.

Acknowledgements

This research is supported in part by the National Key Research and Development Program of China (No. 2016YFB1001405), GD-NSF (No. 2017A030312006), NSFC (No.: 61472144, 61502176), GZSTP (No. 201607010227), China Postdoctoral Science Fund (No. 2019M652896), Fundamental Research Funds for the Central Universities.

References

- [Aarabi *et al.*, 2001] Parham Aarabi, Dominic Hughes, Keyvan Mohajer, and Majid Emami. The automatic measurement of facial beauty. In *IEEE Int. Conf. on Systems, Man, and Cybernetics*, volume 4, pages 2644–2647, 2001.
- [Bertinetto *et al.*, 2016] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *NIPS*, pages 523–531, 2016.
- [Chen and Zhang, 2014] Fangmei Chen and David Zhang. Evaluation of the putative ratio rules for facial beauty indexing. In *ICMB*, pages 181–188. IEEE, 2014.
- [Chung *et al.*, 2017] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *CVPR*, pages 3444–3453. IEEE, 2017.
- [De Brabandere *et al.*, 2016] Bert De Brabandere, Xu Jia, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *NIPS*, pages 667–675, 2016.
- [Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.
- [Gray *et al.*, 2010] Douglas Gray, Kai Yu, Wei Xu, and Yihong Gong. Predicting facial beauty without landmarks. In *ECCV*, pages 434–447, 2010.
- [Ha *et al.*, 2016] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.
- [Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015.
- [Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, pages 675–678, 2014.
- [Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.
- [Li *et al.*, 2015] Jianshu Li, Chao Xiong, Luoqi Liu, Xiangbo Shu, and Shuicheng Yan. Deep face beautification. In *ACM MM*, pages 793–794, 2015.
- [Liang *et al.*, 2017] Lingyu Liang, Duorui Xie, Lianwen Jin, Jie Xu, Mengru Li, and Luojun Lin. Region-aware scattering convolution networks for facial beauty prediction. In *ICIP*, pages 2861–2865, 2017.
- [Liang *et al.*, 2018] Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. *ICPR*, 2018.
- [Lin *et al.*, 2018] Luojun Lin, Lingyu Liang, and Lianwen Jin. R2-resnext: A resnext-based regression model with relative ranking for facial beauty prediction. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 85–90. IEEE, 2018.
- [Liu *et al.*, 2014] Luoqi Liu, Junliang Xing, Si Liu, Hui Xu, Xi Zhou, and Shuicheng Yan. Wow! you are so beautiful today! *ACM Trans., Multimedia Computing, Communications, and Applications (TOMM)*, 11(1s):20, 2014.
- [Perrett *et al.*, 1998] David I Perrett, Kieran J Lee, Ian Penton-Voak, D Rowland, Sakiko Yoshikawa, D Michael Burt, SP Henzi, Duncan L Castles, and Shigeru Akamatsu. Effects of sexual dimorphism on facial attractiveness. *Nature*, 394(6696):884, 1998.
- [Ren and Geng, 2017] Yi Ren and Xin Geng. Sense beauty by label distribution learning. In *IJCAI*, volume 17, pages 2648–2654, 2017.
- [Rothe *et al.*, 2016] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Some like it hot-visual guidance for preference prediction. In *CVPR*, pages 5553–5561, 2016.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.
- [Wang *et al.*, 2017] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017.
- [Xie *et al.*, 2015] Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li. Scut-fbp: A benchmark dataset for facial beauty perception. In *IEEE Int. Conf. on Systems, Man, and Cybernetics*, pages 1821–1826, 2015.
- [Xu *et al.*, 2017] Jie Xu, Lianwen Jin, Lingyu Liang, Ziyong Feng, Duorui Xie, and Huiyun Mao. Facial attractiveness prediction using psychologically inspired convolutional neural network (pi-cnn). In *ICASSP*, pages 1657–1661, 2017.
- [Zhang *et al.*, 2011] David Zhang, Qijun Zhao, and Fangmei Chen. Quantitative analysis of human facial beauty using geometric features. *Pattern Recognition*, 44(4):940–950, 2011.
- [Zhang *et al.*, 2016] David Zhang, Fangmei Chen, and Yong Xu. *Computer models for facial beauty analysis*. Springer, 2016.