IMPROVING FACIAL ATTRACTIVENESS PREDICTION VIA CO-ATTENTION LEARNING

Shengjie Shi[1] *Fei Gao*[1,2,*] *, Member, IEEE* *Xuantong Meng*[1] *Xingxin Xu*[1] *Jingjie Zhu*[1]

[1] Key Laboratory of Complex Systems Modeling and Simulation,
School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China.
[2] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China.

ABSTRACT

Facial attractiveness prediction has drawn considerable attention from image processing community. Despite the substantial progress achieved by existing works, various challenges remain. One is the lack of accurate representation for facial composition, which is essential for attractiveness evaluation. In this paper, we propose to use pixel-wise labelling masks as the meta information of facial composition, and input them into a network for learning high-level semantic representations. The other challenge is to define to what degree different local properties contribute to facial attractiveness. To tackle this challenge, we employ a co-attention learning mechanism to concurrently characterize the significance of different regions and that of distinct facial components. We conduct experiments on the SCUT-FBP5500 and CelebA datasets. Results show that our co-attention learning mechanism significantly improves the facial attractiveness prediction accuracy. Besides, our method consistently produces appealing results and outperforms previous advanced approaches.

***Index Terms*—** Attention, convolutional neural network, facial attractiveness, image quality assessment, face parsing

1. INTRODUCTION

Facial attractiveness plays a significant role in our daily life. For example, people are used to share facial images, such as selfies, through social networks. Naturally, they hope themselves expressed attractive in images. In addition, facial attractiveness has great influences on social acceptance, labor employment, and personal relationships [1]. This leads to a valuable topic - facial attractiveness prediction - in the image processing community [2–4]. Facial attractiveness prediction facilitates the development of applications such as automatic face beautification [5], automatic face makeup [6], and beauty-based face retrieval [7].

In early studies, researchers pay great efforts to design various features based on heuristics rules (e.g., averageness, symmetry, and facial geometry, golden ration) [2, 8–10]. However, these rules and hand-crafted features lack universality [11]. Recently, a number of deep learning based methods are proposed [11, 12]. To name a few, Xu et.al [12] propose a psychologically inspired convolutional neural network (PI-CNN), in which facial detail, lighting and color are sequentially used to train the network. Liang et.al [13] propose a new benchmark dataset and evaluate a number of classic networks on it. Most recently, Fan et.al [11] combine features extracted from ResNet [14] and hand-crafted geometric features to present a facial image, and use a label distribution learning (LDL) method to predict the attractiveness distribution.

Despite the substantial progress achieved by existing methods, various challenges remain. The first key challenge is the lack of accurate representation for facial composition, while facial composition is essential for attractiveness judgement [15]. Existing methods typically use hand-crafted features at landmarks for representing facial composition, which lack universality [11] and show limited performance [13]. To tackle this challenge, we propose to use pixel-wise labelling masks as the meta information of facial composition, and input them into a network for learning high-level semantic representations. To perform attractiveness prediction, we integrate it with another network, which learns attractiveness-aware representation from a facial image.

Another challenge is to define to what degree different regions contribute to facial attractiveness. Attractiveness does not express uniformly across the whole image. In order to evaluate facial attractiveness, instead of the entire spatial domain, we should focus on regions in which attractiveness naturally shows up. To tackle this challenge, we employ a co-attention learning mechanism to automatically and concurrently measure contributions of different regions, and those of distinct facial components.

We conduct experiments on the SCUT-FBP5500 and CelebA datasets. Experimental results show that our co-attention learning mechanism significantly improves the facial attractiveness prediction accuracy. Besides, our method consistently outperforms previous advanced approaches.

Our main contributions can be summarized as follows:

Corresponding Author: gaofei@hdu.edu.cn (Fei Gao).

This work was supported in part by the National Natural Science Foundation of China under Grants 61601158, 61622205, 61836002, 61702145, 61602136, and 61702143, and the Zhejiang Provincial Science Foundation under Grant LQ16F030004.

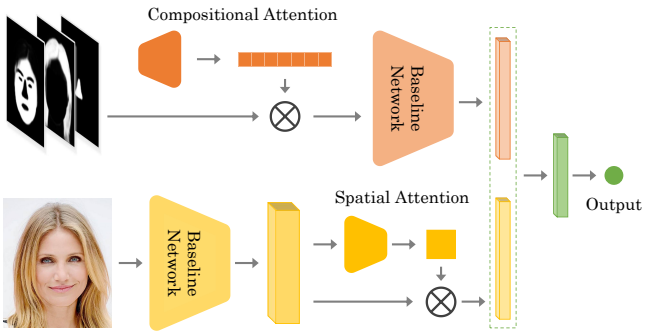


Fig. 1. Framework of the proposed method.

- To our best knowledge, this is the first usage of pixel-wise labelling masks in the facial attractiveness prediction community;
- We propose a co-attention learning mechanism to automatically measure to what degree facial components contribute to attractiveness prediction, which gains significant performance improvement; and
- Our method distinctly outperforms previous state-of-the-art approaches on benchmark datasets.

2. PROPOSED METHOD

Our framework comprises two branched networks with a facial image and its pixel-wise labelling masks as input, respectively. Besides, we employ a co-attention learning mechanism to characterize significant regions that facial attractiveness naturally shows on. Figure 1 shows the pipeline of our method. Details will be introduced next.

2.1. Network Architecture

Facial attractiveness has a wide range of mobile applications, it is thus essential to employ a lightweight network. In this paper, we adopt MobileNetV2 [16] as our network prototype, because it is specifically tailored for mobile environments and has shown appealing performance over various image processing tasks. Our network architecture is shown in Table 1. We refer to [16] for details of MobileNetV2.

In the composition branch, we add a compositional attention module in front of the network. In the image branch, we replace the avgpool layer by a spatial attention layer. Each branch outputs a 1280-dimensional feature vector. Finally, we concatenate the outputs of both branch into a 2560-dimensional vector, and input it into a fully-connected (FC) layer to predict a facial attractiveness label.

2.2. Face Parsing

Given a face photo, we first decompose it into 7 components, i.e. two eyes, two eyebrows, nose, mouth, facial skin, hair, and background. Specially, we employ the face parsing method proposed by Liu et al. [17], and obtain 7 pixel-wise face labelling masks, denoted by $\{\mathbf{M}^{(i)}\}_{i=1}^{7}, \mathbf{M}^{(i)} \in \mathbb{R}^{m \times n}$, where m and n are the height and width of the input image. An element in $\mathbf{M}^{(i)}$ denotes the probability the corresponding pixel belongs to the i-th component, predicted by the model.

We note that an existing face parsing model [17] is adopted here, as this paper is mainly to explore how to use facial composition information for improving performance of facial attractiveness prediction. We expect that a novel advanced face parsing model will be complementary to our approach, but it is beyond the scope of this paper.

Table 1. Network architecture. Each line describes a sequence of 1 or more identical layers, repeated n times. All layers in the same sequence have the same number c of output channels. (This table follows [16])

MobileNetV2 (baseline network)			
Input	Layer	c	n
$224^2 \times 3$	Conv	32	1
$112^2 \times 32$	bottleneck	16	1
$112^2 \times 16$	bottleneck	24	2
$56^2 \times 24$	bottleneck	32	3
$28^2 \times 32$	bottleneck	64	4
$14^2 \times 64$	bottleneck	96	3
$14^2 \times 96$	bottleneck	160	3
$7^2 \times 160$	bottleneck	320	1
$7^2 \times 320$	Conv 1×1	1280	1
$7^2 \times 1280$	avgpool 7×7	-	1
spatial attention module			
Input	Layer	c	n
$7^2 \times 1280$	Conv	1280	1
$7^2 \times 1280$	Tanh	–	1
$7^2 \times 1280$	Conv	1	1
$7^2 \times 1$	Softmax	–	1
compositional attention module			
Input	Layer	c	n
1×1	FC	7	1
1×7	Softmax	–	1

2.3. Spatial Attention

The architecture of the spatial attention module is illustrated in Table 1. We add the spatial attention module after the last convolution layer in MobileNetV2. For facility, we denote its output by $\mathbf{X} = \{\mathbf{X}_{i,j}\}_{i,j=1}^{7} \in \mathbb{R}^{7 \times 7 \times 1280}$; $\mathbf{X}_{i,j} \in \mathbb{R}^{1 \times 1 \times 1280}$ is the feature vector at location (i, j).

Our spatial attention module comprises two convolution layers, which are followed by a Tanh and a Softmax activation layer, respectively. The output of the spatial attention module is a 7×7 map. Elements in the attention map denotes the significance of corresponding local properties for facial

attractiveness prediction.

Let $\mathbf{A}^{(s)} = \{a_{i,j}^{(s)}\}_{i,j=1}^{7}$ denotes the learned spatial attention. $\mathbf{A}^{(s)}$ is used to integrate local activation vectors by:

$$\mathbf{x}_a = \sum_{i=1}^{7}\sum_{j=1}^{7} a_{i,j}^{(s)} \mathbf{X}_{i,j}. \quad (1)$$

$\mathbf{x}_a \in \mathbb{R}^{1\times 1280}$ is concatenated with the output of the composition branch, and then used for attractiveness prediction.

2.4. Compositional Attention

We further explore a channel-wise attention for compositional labelling masks. Specially, it assigns a weighting factor for each composition. Our compositional attention module comprises a fully-connected layer followed by a Softmax layer. The input of this module is a constant, 1. Thus the elements of the fully-connected layer are exactly the weighting factors for the compositional masks. The Softmax layer here ensures the sum of the attention vector being 1.

We denote the compositional attention vector by:

$$\mathbf{a}^{(c)} = (a_1^{(c)}, a_2^{(c)}, ..., a_7^{(c)}), \text{with } \sum_{i=1}^{7} a_i^{(c)} = 1. \quad (2)$$

$a_i^{(c)}$ measures the correlation between the i^{th} component and facial attractiveness. Afterwards, $\mathbf{a}^{(c)}$ is used to integrate the pix-wise labelling masks by:

$$\mathbf{M}_a = \sum_{i=1}^{7} a_i^{(c)} \mathbf{M}^{(i)}. \quad (3)$$

$\mathbf{M}_a$ is input into a network for learning high-level representation of facial composition, and finally used in attractiveness prediction.

2.5. Objective

There are typically two types of attractiveness labels: one is binary, denoting whether a face is attractive or not [18]; the other is a score, denoting to which extent a face is attractive. We formulate the former as a binary classification problem, and use Binary Cross-Entropy (BCE) loss in the learning process. We formulate score prediction as a regression task and use the L_2 loss for training the network.

3. EXPERIMENTS

We evaluate our method and conduct a number of ablation studies on the largest two benchmark datasets, i.e. the SCUT-FBP5500 dataset [13] and Celeb Faces Attributes Dataset (CelebA). Details will be introduced next.

3.1. Settings

Datasets. The SCUT-FBP5500 dataset [13] contains 5500 frontal, unoccluded faces aged from 15 to 60. Each image is released with a attractiveness score. We train and test our model for score prediction using 5-fold cross validation. The average accuracy of all the 5 folds is reported.

CelebA [18] is a large-scale face attributes dataset with more than $200K$ celebrity images, each with 40 attribute annotations. In this paper, we focus on *binary attractiveness classification*. Following standard settings, we use 80% images for training and the rest 20% for testing.

Implementation details. We use `Pytorch` to implement our networks. We first train each branched network with the randomly initialized weights, with an initialized learning rate of 0.001. Afterwards, we apply an initial learning rate of 0.0001 to fine-tune the integrated model in an end-to-end manner. In all the experiments, we use Adam optimizer to train the network for 100 epochs with a batch size of 16, the weight decay of $1e-5$, and momentum of $(0.9, 0.999)$. We gradually decrease the learning rate by 0.1 per 20 epochs.

Criteria. We adopt four widely used indices to evaluate attractiveness score prediction performance on the SCUT-FBP5500 dataset, i.e. the *Pearson linear correlation coefficient* (PLCC), *Spearman's rank-order correlation coefficient* (SRCC), *Mean Absolute Error* (MAE), and *Root Mean Squared Error* (RMSE) between predicted attractiveness scores by using a model and ground-truth scores reported by human. Greater PLCC and SRCC values, while lower MAE and RMSE values, indicate higher score prediction precision. Besides, we use the classification accuracy as the criterion for binary attractiveness classification models on CelebA.

3.2. Ablation Analysis

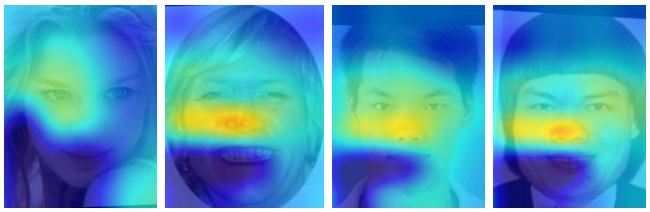
We conduct ablation study on both CelebA and SCUT-FBP5500 datasets. Specially, we evaluate the following five variants of our model: (i) *image*: predict attractiveness merely based on a facial image; (ii) *image+spat.att.*: add the spatial attention module to (i); (iii) *masks*: predict attractiveness merely based on the pixel-wise labelling masks; (iv) *masks+comp.att.*: add the compositional attention module to (iii); (v) *full*: final version of our method.

As shown in Table 2, the model while merely using the pixel-wise labelling masks as input, perform on par with that using a facial image on CelebA. This demonstrate that facial composition is essential for attractiveness evaluation.

Notably, the spatial attention module consistently improves both the attractiveness classification accuracy and score prediction precision. Besides, the compositional attention module gains about 3 points improvement in score prediction precision on the SCUT-FBP5500 dataset. As expected, our full model achieves the best performance on both datasets. These improvements demonstrate the effectiveness of our co-attention learning mechanism.

Table 2. Results of ablation study.

Model Variants	CelebA Acc.(%)	SCUT-FBP5500 PLCC	SCUT-FBP5500 SRCC
image	83.4	0.920	0.909
image+spat.att.	84.4	0.925	0.914
masks	84.1	0.806	0.785
masks+comp.att.	84.1	0.835	0.813
full	**85.2**	**0.926**	**0.916**



(3.80, 4.00) (2.13, 1.93) (2.25, 2.50) (1.65, 1.35)

Fig. 2. Visualization of spatial attention maps. All the images are selected from the SCUT-FBP5500 dataset. The numbers below each image are the corresponding (*ground-truth score*, *predicted score by using our model*).

3.3. Attention Analysis

To better understand what has been learned by attention modules, we visualize the spatial attention weights for different images. Some sample images with attention are shown in Figure 2. We find that the center regions, especially the eyes and nose, express greater weights than the other locations. In addition, the predicted scores by using our model approach the ground-truth ones, and are consistent with human perception.

Interestingly, the eyes and nose statistically correspond to higher compositional attention weights as well. In addition, the average compositional weight of all the facial components is about 0.09, while that of the background is 0.08. In other words, facial components are more important for attractiveness prediction than the background. These properties of learned attention weights improve the interpretability of our attractiveness prediction model.

3.4. Performance on the SCUT-FBP5500 dataset

We first conduct attractiveness score prediction experiment on the SCUT-FBP5500 dataset. To our best knowledge, only Liang *et.al* [13] have evaluated several classic methods on this dataset. Here, we compare with the following ones [13]: 1) *LBP+GR*: Local binary pattern features + Gaussian Regression; 2) *Gabor+SVR*: Use 64-keypoints to obtain a 2560-dimensional Gabor feature vector + Support Vector Regression; and 3) three advanced deep networks, including AlexNet [19], ResNet-18 [14], and ResNeXt-50 [20].

Table 3 shows that our method significantly outperforms the best performance reported in [13]. We note that AlexNet, ResNet-18, and ResNeXt-50 might achieve better performance while using a large-scale labelled data. Nevertheless, with the limited labelled data on the SCUT-FBP5500 dataset, our model shows significant superiority in both efficiency and effectiveness over them. Our model is thus more suitable for applications in constraint computation environments.

Table 3. Performance on the SCUT-FBP5500 dataset.

Method	PLCC	SRCC	MAE	RMSE
LBP+GR [13]	0.674	–	0.391	0.509
Gabor+SVR [13]	0.807	–	0.401	0.518
AlexNet [13]	0.863	–	0.265	0.348
ResNet-18 [13]	0.890	–	0.242	0.317
ResNeXt-50 [13]	0.900	–	0.229	0.302
Ours	**0.926**	**0.916**	**0.202**	**0.266**

Table 4. Performance on the CelebA dataset.

Method	Publication	Acc.(%)
PANDA [22]	CVPR'14	81.0
Liu *et.al* [18]	ICCV'15	81.0
MOON [23]	ECCV'16	81.7
Ding *et.al* [24]	Arxiv'17	82.9
Cao *et.al* [21]	CVPR'18	84.4
Ours	ICASSP'19	**85.6**

3.5. Performance on the CelebA dataset

We further evaluate our method on the CelebA dataset. To our best knowledge, no existing facial attractiveness prediction methods have been evaluated on this dataset. We here compare our method to a number of advanced facial attribute classification methods. These methods typically use multi-task learning technique, which tends to outperform single-task learning based method [18]. Thus, this comparison is disadvantageous to our method. As shown in Table 4, our method gains 1.2 classification accuracy improvement over the previous state-of-the-art [21], and outperform most existing methods by a large margin.

4. CONCLUSIONS

In this paper, we propose to employ facial parsing masks for learning accurate representation of facial composition. Besides, we propose a co-attention learning mechanism to improve facial attractiveness prediction. Experiments demonstrate the effectiveness of our proposed techniques. Despite this achievement, it is still challenging to precisely predict attractiveness for challenging data. This will be one of our future work. Besides, It is promising to further boost the performance by inferring from multiple scales [25]. Finally, it is meaningful to integrate facial attractiveness prediction models into automatic face beautification frameworks.

References

[1] R. Thornhill and S. W. Gangestad, "Facial attractiveness," *Trends in Cognitive Sciences*, vol. 3, no. 12, pp. 452–460, 1999.

[2] S. Liu, Y.-Y. Fan, A. Samal, and Z. Guo, "Advances in computational facial attractiveness methods," *Multimedia Tools and Applications*, vol. 75, no. 23, pp. 16633–16663, Dec 2016.

[3] M. Redi, N. Rasiwasia, G. Aggarwal, and A. Jaimes, "The beauty of capturing faces: Rating the quality of digital portraits," in *IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, May 2015, vol. 1, pp. 1–8.

[4] S. Wang, M. Shao, and Y. Fu, "Attractive or not?:beauty prediction with attractiveness-aware encoders and robust late fusion," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 805–808.

[5] L. Liang, L. Jin, and X. Li, "Facial skin beautification using adaptive region-aware masks," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2600–2612, 2017.

[6] X. Ou, S. Liu, X. Cao, and H. Ling, "Beauty emakeup: A deep makeup transfer system," in *ACM on Multimedia Conference*, 2016, pp. 701–702.

[7] D. Zhang, F. Chen, and Y. Xu, *Data-Driven Facial Beauty Analysis: Prediction, Retrieval and Manipulation*, pp. 217–234, Springer International Publishing, Cham, 2016.

[8] Y. Mu, "Computational facial attractiveness prediction by aesthetics-aware features," *Neurocomputing*, vol. 99, no. 1, pp. 59–64, 2013.

[9] K. Schmid, D. Marx, and A. Samal, "Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios," *Pattern Recognit.*, vol. 41, no. 8, pp. 2710–2717, 2008.

[10] D. Gray, K. Yu, W. Xu, and Y. Gong, "Predicting facial beauty without landmarks," in *Proc. Europ. Conf. Comput. Vis.*, 2010, pp. 434–447.

[11] Y. Y. Fan, S. Liu, B. Li, Z. Guo, A. Samal, J. Wan, and S. Z. Li, "Label distribution-based facial attractiveness computation by deep residual learning," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2196–2208, 2018.

[12] J. Xu, L. Jin, L. Liang, Z. Feng, D. Xie, and H. Mao, "Facial attractiveness prediction using psychologically inspired convolutional neural network (pi-cnn)," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 1657–1661.

[13] L. Liang, L. Lin, L. Jin, D. Xie, and M. Li, "SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction," 2018.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[15] J. Fan, K. P. Chau, X. Wan, L. Zhai, and E. Lau, "Prediction of facial attractiveness from facial proportions," *Pattern Recognit.*, vol. 45, no. 6, pp. 2326–2334, 2012.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," *arXiv:1801.04381v3*, 2018.

[17] S. Liu, Ji. Yang, C. Huang, and M. H. Yang, "Multi-objective convolutional learning for face labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3451–3459.

[18] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec 2015, pp. 3730–3738.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[20] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 5987–5995.

[21] J. Cao, Y. Li, and Z. Zhang, "Partially shared multi-task convolutional neural network with local constraint for face attribute learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 290–4299.

[22] N. Zhang, M. Paluri, M. Ranzato, and T. Darrell, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1637–1644.

[23] E. M. Rudd, M. Gnther, and T. E. Boult, "Moon: A mixed objective optimization network for the recognition of facial attributes," in *Proc. Europ. Conf. Comput. Vis.*, 2016, pp. 19–35.

[24] H. Ding, H. Zhou, S. K. Zhou, and R. Chellappa, "A deep cascade network for unaligned face attribute classification," *arXiv:1709.03851v2*, 2017.

[25] F. Gao, J. Yu, S. Zhu, Q. Huang, and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," *Pattern Recognit.*, vol. 81, pp. 432–442, 2018.