Label Distribution-Based Facial Attractiveness Computation by Deep Residual Learning

Yang-Yu Fan, Shu Liu, Bo Li, Zhe Guo, Ashok Samal, Jun Wan, *Member, IEEE*, and Stan Z. Li, *Fellow, IEEE*

***Abstract*—Two key challenges lie in the facial attractiveness computation research: the lack of discriminative face representations, and the scarcity of sufficient and complete training data. Motivated by recent promising work in face recognition using deep neural networks to learn effective features, the first challenge is expected to be addressed from a deep learning point of view. A very deep residual network is utilized to enable automatic learning of hierarchical aesthetics representation. The inspiration to deal with the second challenge comes from the natural representation of the training data, where each training face can be associated with a label (score) distribution given by human raters rather than a single label (average score). This paper, therefore, recasts facial attractiveness computation as a label distribution learning problem. Integrating these two ideas, an end-to-end attractiveness learning framework is established. We also perform feature-level fusion by incorporating the low-level geometric features to further improve the computational performance. Extensive experiments are conducted on a standard benchmark, the SCUT-FBP dataset, where our approach shows significant advantages over the other state-of-the-art work.**

***Index Terms*—Facial attractiveness computation, deep residual network, label distribution, feature fusion, SCUT-FBP.**

I. Introduction

THE attractiveness of a face influences many social endeavors. Psychological research indicates that many advantages and privileges come with having a beautiful face, such as social acceptance, professional advancement, and personal relationships [1], [2]. With the popularization of digital cameras, facial images have become pervasive in every aspect of human life, thus leading to an emerging topic—automatic computation of facial attractiveness—in the computer vision community. This facilitates the development of real-world applications such as unconstrained facial beauty assessment [3], beauty-based face retrieval [4], facial image retouching [5], [6], and face makeover recommendation [7], [8]. Unlike the task of photo quality assessment which considers lighting effects and photographic composition to evaluate the aesthetics of portrait photos [9], [10], the facial attractiveness computation research concentrates solely on the attractiveness of faces and attempts to decode it automatically and reliably.

Although substantial progress has been achieved in face attractiveness computation research, various challenges remain. The first outstanding challenge is the lack of an accurate face representation that captures the salient elements of attractiveness. Many heuristics rules have been developed over the years, therein inspiring researchers to manually design various features in most early studies. The features can be geometric, as well as appearance-based descriptors either at local or holistic scales [11]. However, there remain no conclusive findings on to what degree different types of features contribute to facial attractiveness. The effectiveness of the face representation and its characterized features are the main reason for performance bottleneck in attractiveness computation.

The face representations proposed in the existing literature can be divided into feature-based, holistic, and hybrid representation [12], [13]. In the feature-based approach, researchers derive a diverse set of features, including geometric, texture, color and other structural features [14]–[18], from a face as its aesthetics-aware predictors. Since features are always artificially handcrafted and selected for a particular experimental dataset, the interpretation of the extracted features is intuitive from a human perspective; the downside is that intensive manual work is inevitable, and the extracted features tend to lack universality. Moreover, some discriminative features may not be detected because of human bias.

The holistic approach is proposed to overcome the limitations of handcrafted features. Such approach extracts features from the whole face and examines the spatial interrelationships between these features. A typical representation of a face in the holistic approach can be a long concatenated vector of the raw facial image or the low-dimensional version after applying dimensionality reduction such as Eigenfaces [18], [19] and manifolds [20]. Recently, up-to-date deep learning methods, especially convolutional neural networks (CNNs) and

Manuscript received February 24, 2017; revised July 14, 2017, September 27, 2017, and November 13, 2017; accepted November 23, 2017. Date of publication December 6, 2017; date of current version July 17, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61402371, Grant 61461025, Grant 61702462, and Grant 61502491, in part by the Science and Technology Innovation Engineering Plan in Shaanxi Province of China under Grant 2013SZS15-K02, and in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2017JM6008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hatice Gunes. *(Corresponding author: Shu Liu.)*

Y.-Y. Fan, S. Liu, B. Li, and Z. Guo are with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: fan_yangyu@nwpu.edu.cn; liushu0922@mail.nwpu.edu.cn; libo.npu@gmail.com; guozhe@nwpu.edu.cn).

A. Samal is with the Department of Computer Science and Engineering, University of Nebraska–Lincoln, Lincoln, NE 68588 USA (e-mail: samal@cse.unl.edu).

J. Wan and S. Z. Li are with the Center for Biometrics and Security Research and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jun.wan@nlpr.ia.ac.cn; szli@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2780762

their variations, have been applied to automatically learn a hierarchical and higher-level face representation for facial attractiveness computation tasks. Even under the shallow and plain architectures, such as one-layer auto-encoder [21], two-layer convolutional restricted Boltzmann machine (CRBM) [22], two-layer convolutional principal component analysis filter-based network (PCANet) [4], [11], and six-layer CNN [23], [24], superior performance has been achieved compared to traditional handcrafted and holistic features.

The hybrid representation integrates the above two representations in some form. Since different types of features usually contain different aesthetic information about the face, their fusion is expected to achieve better performance than either feature set [4], [18], [19], [21], [25]. Because face attractiveness is a complex concept with as of yet no universally accepted representation, we are motivated to determine the effective visual characteristics by an appropriate hybrid representation.

The second key challenge in predicting face attractiveness is the scarcity of "sufficient and complete" training data. A large dataset ("sufficient") with diverse faces ("complete") is critical at the outset of the development of facial attractiveness computational models. Faces at all levels of attractiveness, especially extremely beautiful faces, should be included to provide different resolutions to facial attractiveness [12]. The publicly available databases for attractiveness computation typically contain no more than a few thousands faces, most of which are of average attractiveness; a very limited number of faces are considered to be very attractive. A recently-proposed benchmark, the SCUT-FBP dataset [23], has collected a higher percentage of beautiful faces than the general population; however, the size of the dataset (500 facial images) remains too small to develop a powerful learning algorithm.

In order to reinforce the attractiveness learning process and improve the generalization ability of the computational model on the available small datasets, additional knowledge need to be considered. Our idea is inspired by the natural representation of collected attractiveness scores. Normally, the quantitative analysis of facial attractiveness is formulated as a classification or regression problem. A typical method to derive an approximation of true attractiveness labels (scores) is to survey a diverse group of human raters who assign individual scores to a set of faces. The average score of all the raters for each face is then defined as its ground-truth label in the subsequent classification or regression task [12], [13]. However, the average score is not always a good indicator of universally accepted preference, especially for controversial faces. In contrast, the score distribution across different raters provides more aesthetics-degree information of a face than a single score. Fig. 1 illustrates this contrast between the average attractiveness score versus the score distribution (using a 5-point scale) of one face. It can be observed that there is a range of scores for a single face, and a distribution more accurately represents the full extent of the attractiveness of the face. In this sense, a straightforward idea is to employ a novel machine learning paradigm—label distribution learning (LDL) [26]—since the score distribution can be viewed as a natural representation of a label distribution. LDL provides us an alternate view of the attractiveness learning process, where each face will be mapped to a label distribution instead of a single label. Therefore, we recast facial attractiveness computation as an LDL problem rather than a traditional single-label learning task. The LDL is able to deal with insufficient and incomplete training data, since each face is expected to contribute to the learning of a number of attractiveness levels.

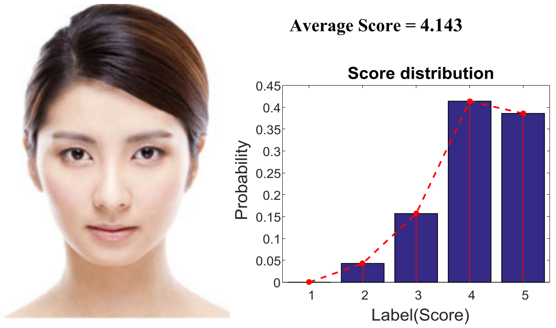


Fig. 1. The average attractiveness score and score distribution of one face in the SCUT-FBP dataset [23].

In this work, we intend to explore deeper facial aesthetics features by learning features from raw images directly through deeper architecture. A very deep convolutional residual network (ResNet) [27], which takes RGB raw images as inputs and automatically learns an effective face representation, is utilized. Inspired by the nature of the score distribution associated with each face instance, the facial attractiveness computation is formulated as an LDL problem. We integrate the idea of label distribution learning and deep feature learning into an end-to-end attractiveness learning framework. We are also interested in leveraging the deep high-level and low-level geometric features, and exploring the role of feature fusion to further improve the attractiveness prediction. A newly constructed benchmark SCUT-FBP dataset is used to demonstrate the superiority of our approach. In summary, the main contributions of this work are as follows.

1) By investigating the natural representation of attractiveness score collection, we suggest a novel paradigm for facial attractiveness computation, i.e., associating each training instance with a label distribution and recasting the task as a label distribution learning problem.
2) The pre-trained ResNet is exploited and adaptively transferred to our task. An end-to-end attractiveness learning framework is proposed accordingly. In comparison to the related work that used shallow architectures (typically several layers), this is the first to leverage the power of the very deep network to predict facial attractiveness.
3) Our framework incorporates low-level geometric features, which are more discriminative than those heuristic features commonly used in the existing methods, to further boost the attractiveness prediction performance. The experimental results have demonstrated the effectiveness of our method for the task of facial attractiveness computation.

The remainder of this paper is organized as follows. In Section II, we review the existing work related to facial attractiveness computation as well as LDL and its applications. The details about the aesthetics-aware feature extraction and LDL model are proposed in Section III. The experimental settings, performance results, and the related analysis are presented in Section IV. Finally, Section V is devoted to the main conclusions and future work.

II. Related Work

Previous facial attractiveness computation research focused on exploring handcrafted features such as geometric, texture, color, as well as holistic descriptors. Geometric features are the most commonly used aesthetics-aware predictors in face attractiveness analysis in terms of the facial landmark position, distances between the landmarks, and ratios of the distances [16]–[19], [25]. Typical geometric features are artificially designed based on heuristics and universal rules of beauty such as Golden ratios, the Facial Fifths and Thirds, and the Symmetry Theory [15], [17], [28]. Widely investigated texture features are Gabor filter responses, local binary patterns (LBPs) and skin smoothness indicators [14], [19]. Active appearance model (AAM) parameters, which encode both facial shape and texture information, are also extracted as potential features [4]. Color features include color symmetry, hue, saturation and value (HSV) coordinates, and color distribution [14], [18]. Holistic descriptors, like Eigenface [18], [19], [25], manifold [20], and face shape model [29], contain the information of the whole face.

Research on handcrafted and holistic features has led to some early success; however, these features are low-level features extracted from images. With state-of-art deep learning methods, it has now become possible to automatically learn higher levels of image representations, which have been reported to achieve promising performance for many vision applications. Krizhevsky *et al.* [30] trained a deep CNN on the ImageNet database [31] and achieved substantial increase in performance in classification tasks. Since then, various classical architectures of deep CNNs have been proposed, when transfer learning becomes popular. The learned features from pre-trained CNN models can be efficiently transferred to new visual classification or recognition problems with remarkable results [32], [33].

The effectiveness of CNNs and variations has been cursorily explored in facial attractiveness computation. Gray *et al.* [34] were the first to develop a hierarchical feed-forward model, which can be seen as a form of CNN, to extract multi-layer appearance features for attractiveness prediction. Wang *et al.* [21] trained pairs of one-layer auto-encoders from low-level features. Even under such shallow networks, the learned features were proven to be more attractiveness-aware than the traditional low-level features. Gan *et al.* [22] utilized deep self-taught learning and CRBM to learn human-like predictors, which outperformed the Eigenfaces, Gabor features, and LBP. In the work [4], the face representations extracted from two-stage PCANet were validated for attractiveness prediction task, and the performance was further enhanced by an optimal fusion with LBP and AAM features. Xu *et al.* [24] constructed a six-layer CNN model with the detail layer image to investigate the most discriminative facial attractiveness attributes. Very recently, the 16-layer pre-trained VGGNet [35] was used to extract powerful deep features for attractiveness preference prediction [36]. To some extent, the deeper architectures could extract more discriminative features. Motivated by the very recent residual network trained with hundreds of layers [27], our work aims to transfer the pre-trained ResNet to facial attractiveness computation task with currently available but small datasets, and further boost the performance by incorporating the low-level geometric features.

Conventional facial attractiveness computation is formulated as a classification or regression problem. After the feature extraction, the next step is to develop an effective classifier [16], [20] or regressor [21], [28] that takes features and average scores as the inputs. Many regression-based methods have been adapted to learn a mapping between multi-dimensional features and scores [22], [23], [25]. The key aspects of this kind of single-label learning framework lie in either the improved features or more sophisticated regression algorithms.

The LDL is a novel machine learning paradigm recently proposed for facial age estimation [37]. It was found that age is an ambiguous variable and faces with adjacent ages are strongly correlated; in other words, one age might be used to learn its neighbors. These observations form the basis of LDL theory, the main idea of which is to utilize the adjacent ages when learning a particular one. Therefore, a proper label distribution, representing the extent to which each label describes the overall instance, has been generated for each face. Customized LDL algorithms such as IIS-LLD and CPNN have also been proposed [37]. The problem of crowd counting shares the same characteristics as facial age and has successfully adopted the LDL framework to achieve outstanding performance [38].

Facial attractiveness appears similar to facial age, i.e., changes in attractiveness are a relatively slow and smooth process, and faces at adjacent beauty levels are highly correlated. More importantly, the score distribution of each face, which is naturally viewed as a representation of label distribution, can be derived directly from the rating procedure whereby multiple labeling sources assign scores to faces. Thus, the existing LDL paradigm is an ideal match for the task of attractiveness computation. A similar application like pre-release movie rating prediction is given in [39], where another LDL-based algorithm named LDSVR was proposed.

It is worth mentioning that although the label distribution aggregates attractiveness scores from multiple raters and thus captures the inter-rater variability, this paper is aimed at predicting the universal facial attractiveness preference shared by the group of observers rather than the personalized preference of a particular individual. Personalized facial attractiveness computation is a potentially more challenging task because it takes previous ratings from a target person or other people with similar preferences to develop an individualized model. Such models are useful when person-specific attraction to faces is required, for example, in online dating services for recommending partners. Several representative works can be found in [19] and [36]; however, the prediction accuracy reported is inferior to universal models.

III. The Proposed Attractiveness Computational Model

The overview of our framework is illustrated in Fig. 2. Each face instance is annotated with a label distribution collected

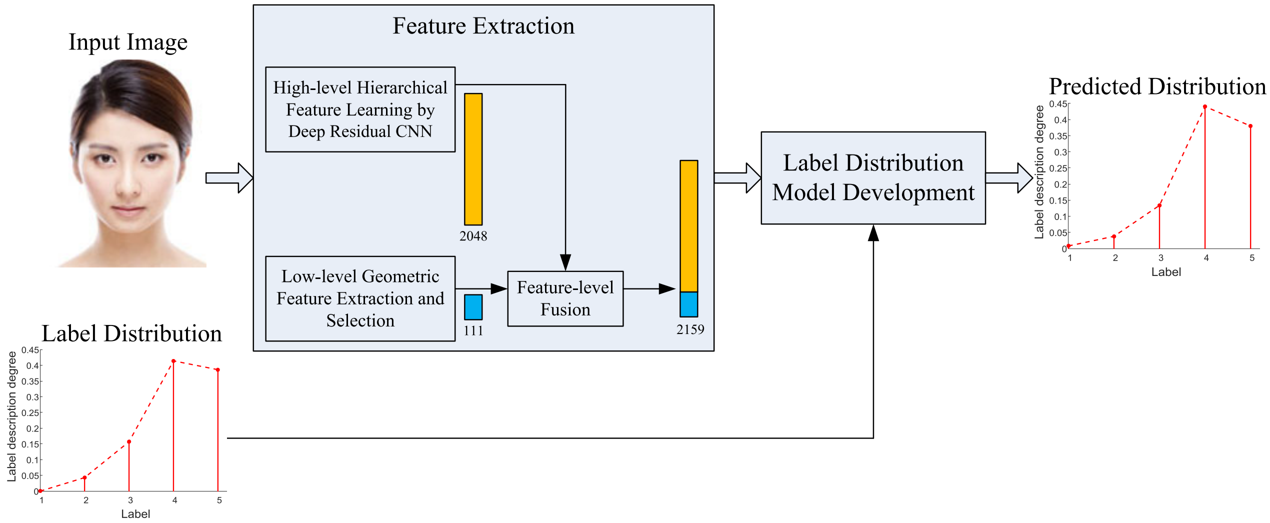


Fig. 2. The proposed LDL framework of facial attractiveness computation.

from multiple human raters, rather than the average-score label. A hybrid face representation is extracted, including the high-level hierarchical features from the deep residual network and low-level geometric features. Feature-level fusion is then performed to form a concatenated vector of these two types of features. A neural-network-based label distribution model is adapted to learn a mapping between the feature vector and the label distribution. Once the model is developed, it can be used for predicting the attractiveness-label distribution of an instance.

A. *Label Distribution of Attractiveness*

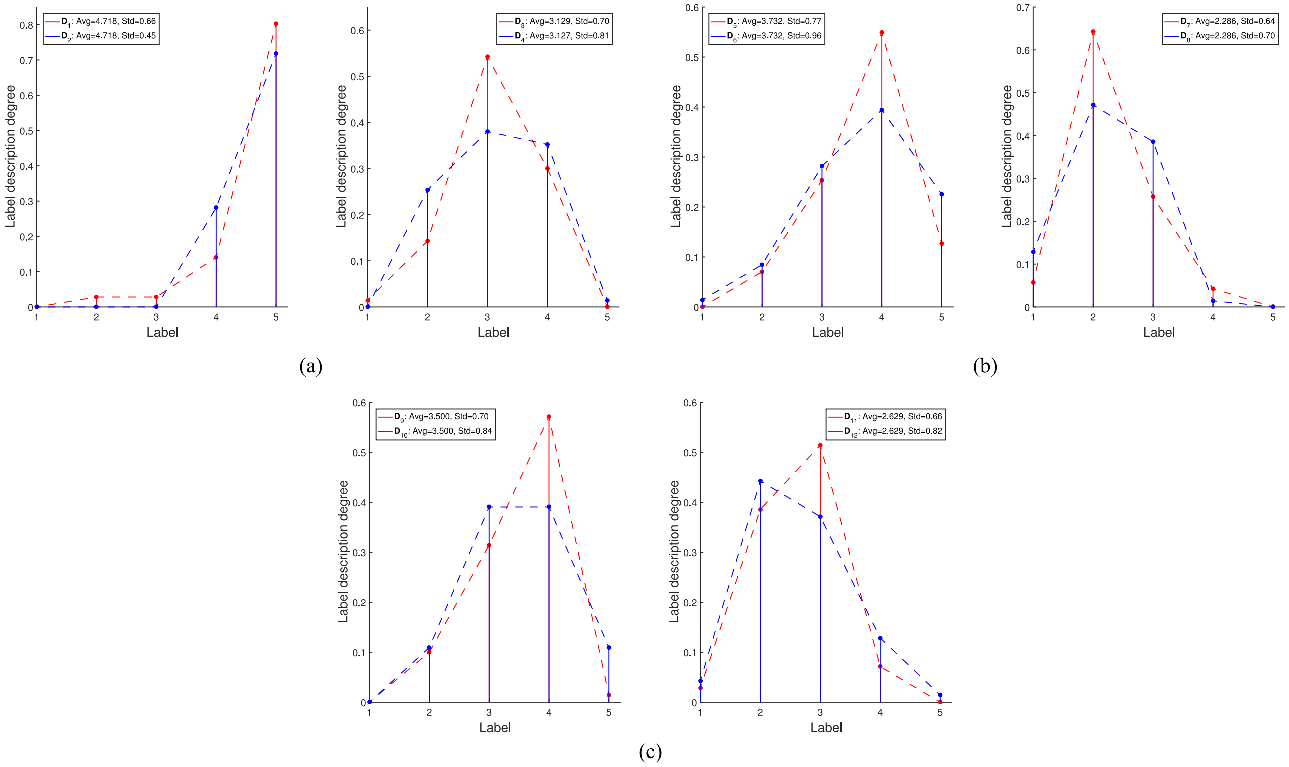
As discussed in Section I, the score distribution collected from human raters can be viewed as a natural representation of a label distribution. Following the definitions in [26], we regard the attractiveness scores offered for the rating procedure as the complete set of labels $\mathbf{y} = \{y_1, y_2, \ldots, y_c\}$, and the probability distribution of rating scores as the label distribution $\mathbf{D} = \{d_{\mathbf{x}}^{y_1}, d_{\mathbf{x}}^{y_2}, \ldots, d_{\mathbf{x}}^{y_c}\}$ where $d_{\mathbf{x}}^{y_j}$ denotes the description degree (distribution) of the label y_j to the instance $\mathbf{x}$. The description degree represents the extent to which each label describes the overall instance, under the constraint $\sum_{j=1}^{c} d_{\mathbf{x}}^{y_j} = 1$ meaning that the label set $\mathbf{y}$ fully describes the instance.

The average score of all the raters is the most popular ground-truth label in current attractiveness model. In contrast, one novelty of this paper lies in the fact that we associate each face instance with a label distribution, which provides more aesthetics-degree information of the face. Fig. 3 shows several examples of different label distributions with five labels for faces with similar average score. In Case (a), two instances with almost the same average score are described by different sets of labels, e.g., in Fig. 3(a) - left, instance $\mathbf{x}_1$ (shown in red) is described by the label set $\{y_2, y_3, y_4, y_5\}$ when $\mathbf{x}_2$ (shown in blue) is described by only the labels y_4 and y_5. In Case (b), two instances are described by the same label set, and the general patterns of the two distributions appear similar, which means that their peaks appear at the same label. However, some label description degrees vary considerably, representing different variances of two sets of rating scores. For example, in Fig. 3(b) - left, both distributions $\mathbf{D}_5$ (shown in red) and $\mathbf{D}_6$ (shown in blue) show some skew to the left, and both peaks appear at label y_4; however, $\{d_{\mathbf{x}_5}^{y_4}, d_{\mathbf{x}_5}^{y_5}\} = \{0.549, 0.127\}$ and $\{d_{\mathbf{x}_6}^{y_4}, d_{\mathbf{x}_6}^{y_5}\} = \{0.394, 0.225\}$ vary most among the description degrees inducing different standard deviations. In Case (c), although two instances are described by the same label set, the general patterns of the two distributions are dissimilar, which means that their peaks occur at different labels. As shown in Fig. 3(c), the distribution $\mathbf{D}_9$ shows some left skew, while $\mathbf{D}_{10}$ is bimodal; the peak of $\mathbf{D}_{11}$ is at y_3, while the peak of $\mathbf{D}_{12}$ is at y_2.

Fig. 3 reinforces the significance of the label distribution over the conventional average score. The average score is not always a fair representation of crowd attractiveness preference since the averaging process weakens the influence of relative extremely lower and higher scores. The label distribution, on the other hand, not only incorporates the general tendency of the crowd preference but also reveals the variability within it. In other words, the label distribution introduces additional aesthetics-degree knowledge of faces, and therefore is applied as the ground truth to our attractiveness modeling.

B. *Feature Extraction*

1) High-Level Face Representation Learning: In this work, we aim to leverage the power of the very recently developed deep residual network [27] to automatically learn a high-level aesthetics face representation for predicting attractiveness. The original ResNet was specially designed for the ImageNet classification challenge. The architecture utilized here is an adapted version of the ResNet, i.e., all layers are kept except for the last fully-connected layer. The network takes fixed-size inputs (224 × 224 × 3), however, it is not appropriate to resize the facial images directly since their size and aspect ratios will affect the visual aesthetics of the faces. Instead, we first extract the face region from each original image, making it possible to concentrate solely on the attractiveness of the face itself. We then pad



(a)

(b)

(c)

Fig. 3. Three cases of different label distributions for the instances with similar average scores: (a) distributions with different label sets; (b) distributions with the same label set and similar general pattern but varying description degrees; and (c) distributions with the same label set but different general patterns.

the border pixels with zeros to the shorter side of the image and resize it to generate a normalized square input for our network.

The details of the ResNet architecture can be found in [27]. Basically, it contains a convolutional layer, a max-pooling layer, and four convolutional "bottleneck" blocks followed by an average-pooling layer. Batch normalization is performed after each convolution layer. The depth of the architecture depends on the number of four blocks. For the 50-layer ResNet (ResNet-50) and 101-layer ResNet (ResNet-101), the number of blocks is $\{2, 3, 5, 2\}$ and $\{2, 3, 22, 2\}$, respectively.

It is challenging to train an effective ResNet on a small experimental dataset due to high likelihood of severe over-fitting. To mitigate this problem, the ResNets pre-trained on the ImageNet dataset [31] are exploited as an initialization in our task.[1] Several augmentation techniques including the standard color augmentation [30], rotation, and contrast enhancement are used to generate diverse training samples. We then fine-tune the ResNets using the augmented data.

2) Low-Level Geometric Feature Extraction: One of the unique aspects of this work is how the low-level geometric features are selected. In previous work, researchers took a pre-defined set of features based on heuristics and universal rules of beauty (e.g., Golden ratios, and the Facial Fifths and Thirds) as the aesthetics-aware predictors. These handcrafted features are generally derived from cognitive findings in social sciences. Such rule-driven feature extraction tends to lack universality because diverse features are used in different studies. In contrast, we use a data-driven approach [40] to determine which subset of features is important for beauty perception from an exhaustive pool of candidates. This is a significantly different perspective on attractiveness computation as all the features are treated equally without bias towards previous claim of their relevance to attractiveness.

The geometric features are derived in three steps: facial landmark detection, candidate feature extraction, and incremental feature selection. Based on a set of detected facial landmarks, three types of geometric features are extracted (i.e., ratios, angles, and vertical inclinations), of which the most discriminative subset is then incrementally selected as our final attractiveness-aware geometric predictors.

Based on our previous work in [41], 82 facial landmarks are automatically annotated. Procrustes analysis [42] is then applied to normalize each face into a pre-shaped space. Candidate features including ratios, angles and vertical inclinations are induced by the landmark set. From the 82 facial points, 3321 line segments connecting any two points (C_{82}^{2}, where C is the combination function) can be obtained, which induce 3321 vertical inclinations. Angle features are formed by any two lines; thus, a total of 5,512,860 angles (C_{3321}^{2}) can be obtained. Similarly, 11,025,720 ratios (P_{3321}^{2}, where P is the permutation function) are derived from 3321 facial distances. Note that the permutation function is used instead of combination, since the arrangement of two distances in a ratio as the antecedent or consequent may have different effects on attractiveness perception. All possible ratios, angles, and inclinations constitute an

[1] The pre-trained models ResNet-50 and ResNet-101 can be downloaded from https://github.com/KaimingHe/deep-residual-networks.

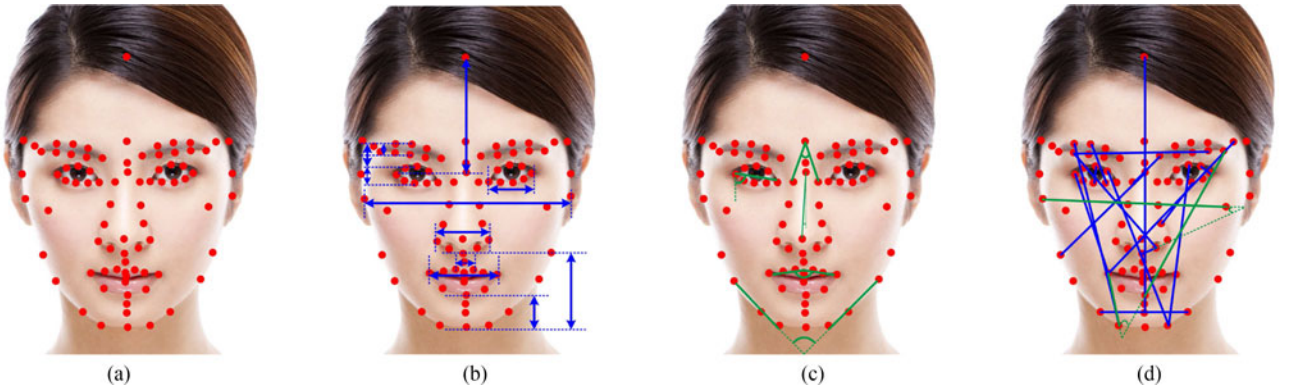


Fig. 4. (a) The 82 facial landmarks used in [41]; (b) examples of candidate ratios (shown in the form of distances); (c) examples of candidate angles and vertical inclinations; and (d) the 10 most discriminative geometric features.

Algorithm 1: Incremental Feature Selection (IFS)

```
Input: The original feature set $\mathbf{S}_0$, the number of features in each incremental set $n_i$, and the threshold number of discriminative features $n_t$.
Output: A subset of discriminative features $\mathbf{S}$.
repeat
    Initialize $\mathbf{S} = \emptyset$;
    Partition $\mathbf{S}_0$ into $\frac{size(\mathbf{S}_0)}{n_i}$ incremental sets;
    for each incremental set do
        Run stepwise regression to select most discriminative features;
        Add selected discriminative features to current $\mathbf{S}$;
    end for
    $\mathbf{S}_0 \leftarrow \mathbf{S}$;
until $size(\mathbf{S}_0) \leq n_t$
return S
```

exhaustive pool of candidate features. The illustration of the facial landmark distribution and examples of candidate features are given in Fig. 4(a)–(c).

While most candidate features are unlikely significant for attractiveness determination and their intrinsic dimension should not be high, the stepwise regression [43] is applied to select the most important features from the original set of candidates. The main idea of the stepwise regression can be found in [40]. In practice, selecting the discriminative ones from such large number of features in a one-step procedure is computationally intractable. Therefore, in this paper, we partition the original feature set into smaller subsets and run the stepwise regression in batch on the subsets to incrementally select the most discriminative features. This process may be repeated if the size of the selected subset remains too large to be input into the final LDL model. The algorithm is named Incremental Feature Selection (IFS) and summarized in Algorithm 1.

We first perform the IFS algorithm on the original sets of ratios, angles, and vertical inclinations separately, and then pool the selected features together and run a second-stage IFS. As a result, a combined subset of 38 most discriminative features (19 ratios, 17 angles, and 2 inclinations) is identified as our final geometric predictors. The top ten features ranked by their significance are visualized in Fig. 4(d). Although the interpretation of these measures is not intuitive, some general insights into the nature of the discriminative structural information can be obtained. The selected features are predominantly associated with the facial regions of eyebrows, eyes, nose, and chin, while the mouth, cheeks, and upper face contour fail to show clear correlation with attractive traits.

C. Label Distribution Model Development

In this work, facial attractiveness computation is formulated as an LDL problem (shown in Fig. 2). The extracted features and the label distribution of instances are fed into a label distribution model, which aims to learn a mapping between them. According to the discussion in [26], there are three strategies in designing LDL algorithms: two indirect strategies including problem transformation and algorithm adaption, and one direct strategy of specialized algorithms. The specialized algorithms are expected to be best suited to the LDL problem because they attempt to perform the optimization for LDL directly. After preliminary experimentation with existing LDL algorithms, it is found that label distribution support vector regressor (LDSVR) [39] performs the best.

In our framework, we design a simple neural network to deal with label distributions, which can also be trained together with the above feature learning. The network consists of a fully-connected layer with the number of neurons equal to the total number of labels, and a softmax layer followed by a loss layer. Given a training set $\{(\mathbf{x}_i, \mathbf{D}_i), 1 \leq i \leq n\}$, where $\mathbf{D}_i = \{d_{\mathbf{x}_i}^{y_1}, d_{\mathbf{x}_i}^{y_2}, \ldots, d_{\mathbf{x}_i}^{y_c}\}$ is the label distribution associated with the instance (feature representation) $\mathbf{x}_i$, the label distribution learning process is to train a set of network parameters $\boldsymbol{\theta}$ to generate a probability distribution $p(\mathbf{y}|\mathbf{x}_i; \boldsymbol{\theta})$ for the label set $\mathbf{y}$, which is similar to $\mathbf{D}_i$. By feeding $\mathbf{x}_i$ into the c-neuron fully-connected layer, a distribution of c components is produced. To ensure that the components satisfy the constraints of the description degree mentioned in Section III-A (i.e., $d_{\mathbf{x}_i}^{y_j} \in [0, 1]$ and $\sum_{j=1}^{c} d_{\mathbf{x}_i}^{y_j} = 1$), an additional normalization process is performed by the softmax layer:

$$p(y_j|\mathbf{x}_i; \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}_j^{\mathrm{T}} \mathbf{x}_i)}{\sum_{j=1}^{c} \exp(\boldsymbol{\theta}_j^{\mathrm{T}} \mathbf{x}_i)} \tag{1}$$

Finally, the Euclidean distance and Kullback-Leibler (KL) divergence are defined as the loss function measuring the similarity between the ground-truth distribution $\mathbf{D}_i$ and the predicted distribution $p(\mathbf{y}|\mathbf{x}_i;\boldsymbol{\theta})$. The objective of the label distribution model is to minimize either of the following overall losses:

$$L_{Eu} = \sum_{i=1}^{n}\sum_{j=1}^{c}(d_{\mathbf{x}_i}^{y_j} - p(y_j|\mathbf{x}_i;\boldsymbol{\theta}))^2$$

$$L_{KL} = \sum_{i=1}^{n}\sum_{j=1}^{c}\left(d_{\mathbf{x}_i}^{y_j}\ln\frac{d_{\mathbf{x}_i}^{y_j}}{p(y_j|\mathbf{x}_i;\boldsymbol{\theta})}\right) \tag{2}$$

It is worth noting that the feature learning module and our label distribution model can be regarded as an integrated network for attractiveness learning. To optimize (2), the fine-tuning for ResNet and training of the label distribution model parameters $\boldsymbol{\theta}$ are implemented together using the stochastic gradient descent (SGD) method with the backpropagation algorithm. Thus, the entire attractiveness learning framework is end-to-end.

IV. Experiments and Results

In this section, we report the results of our experiments on a benchmark dataset to demonstrate the effectiveness of our facial attractiveness computational model. The dataset, implementation details, and evaluation measures are first introduced. The experiments have five main purposes: 1) to verify the power of ResNet in learning high-level face representations for facial attractiveness prediction; 2) to demonstrate the effectiveness of label distributions in attractiveness learning; 3) to demonstrate the effectiveness of feature-level fusion; 4) to compare our model with the state of the art; and 5) to present the qualitative analysis of our model and evaluation results on other datasets.

A. Dataset and Implementation Details

The SCUT-FBP dataset[2] [23] is used in the experiments to evaluate the performance of our approach. It contains high-resolution facial images of 500 Asian females, of similar age, with almost neutral expression. The faces are in a frontal-on position, with minimal accessories and occlusion. We first preprocess the original images before the following experiments by extracting the face region where the neck and other body parts have almost been removed. This makes it possible to concentrate solely on the attractiveness of the faces themselves.

Each face was rated by approximately 70 human raters on a 5-point scale, where 1 means strong disagreement about the face being beautiful (i.e., being the least attractive) and 5 means strong agreement about the face being beautiful (i.e., being the most attractive). All the ratings have been verified as reasonable and appropriate, and they were ultimately recorded in the dataset, which enable us to calculate the score distribution and the average score for each image.

Following the data partition setting in [24], 400 images are randomly selected as the training set, and the remaining 100 images as the test set. Several augmentation techniques are used to enlarge the size of the training data to 8,000. The steps of the feature extraction and LDL model development are applied to the training set only. The trained model is then evaluated on the non-overlapped test set.

The experiments are conducted on the deep learning platform of Caffe [44]. For high-level feature learning, we initialize network weights with pre-trained models. The low-level geometric feature extraction is a separate process and is performed in advance. The deep features and geometric features are then concatenated via a "concat" layer and fed into our label distribution model. In the end, the feature extraction module and label distribution model are trained together by the SGD method with the back propagation algorithm. We fix the batch size to 5 and the weight decay to 0.0001. The initial learning rate is 0.001 and decreases by 0.1 every 4,000 iterations. The maximum number of iterations is set to 12,000. As mentioned in Section III-C, both the Euclidean distance and KL divergence are used in the loss function, where experiments with the Euclidean distance achieve slightly better results. Therefore, all the following results are obtained under this setting.

B. Evaluation Measures

Given a test image in the test set $\{(\mathbf{x}_i', \mathbf{D}_i'), 1 \leq i \leq m\}$, the normalized input is fed into the feature extraction module to compute the feature representation $\mathbf{x}_i'$, and the predicted label distribution is generated by $p(\mathbf{y}|\mathbf{x}_i';\boldsymbol{\theta})$. As suggested in [26], six metrics can be applied to measure the average distance/similarity between the ground-truth and the predicted distributions. The metrics include four distance measures (smaller values are better), i.e., Chebyshev distance, Clark distance, Canberra metric, and KL divergence, and two similarity measures (larger values are better), i.e., Cosine coefficient and Intersection similarity. The formulas of the six metrics are given below:

$$Chebyshev = \frac{1}{m}\sum_{i=1}^{m}\max_{j}\left|d_{\mathbf{x}_i'}^{y_j} - p(y_j|\mathbf{x}_i';\boldsymbol{\theta})\right|$$

$$Clark = \frac{1}{m}\sum_{i=1}^{m}\sqrt{\sum_{j=1}^{c}\frac{(d_{\mathbf{x}_i'}^{y_j} - p(y_j|\mathbf{x}_i';\boldsymbol{\theta}))^2}{(d_{\mathbf{x}_i'}^{y_j} + p(y_j|\mathbf{x}_i';\boldsymbol{\theta}))^2}}$$

$$Canberra = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{c}\frac{\left|d_{\mathbf{x}_i'}^{y_j} - p(y_j|\mathbf{x}_i';\boldsymbol{\theta})\right|}{d_{\mathbf{x}_i'}^{y_j} + p(y_j|\mathbf{x}_i';\boldsymbol{\theta})}$$

$$KL = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{c}d_{\mathbf{x}_i'}^{y_j}\ln\frac{d_{\mathbf{x}_i'}^{y_j}}{p(y_j|\mathbf{x}_i';\boldsymbol{\theta})}$$

$$Cosine = \frac{1}{m}\sum_{i=1}^{m}\frac{\sum_{j=1}^{c}d_{\mathbf{x}_i'}^{y_j}p(y_j|\mathbf{x}_i';\boldsymbol{\theta})}{\sqrt{\sum_{j=1}^{c}(d_{\mathbf{x}_i'}^{y_j})^2}\sqrt{\sum_{j=1}^{c}(p(y_j|\mathbf{x}_i';\boldsymbol{\theta}))^2}}$$

$$Intersection = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{c}\min_{j}\left(d_{\mathbf{x}_i'}^{y_j}, p(y_j|\mathbf{x}_i';\boldsymbol{\theta})\right) \tag{3}$$

Since the comparable work is aimed at exact attractiveness computation, the predicted score can be estimated by the weighted mean of the predicted label distribution. Following the majority of previous work, three evaluation metrics, namely Pearson correlation (PC), mean absolute error (MAE) and root mean squared error (RMSE), are utilized to gauge the perfor-

[2] [Online]. Available: http://www.hcii-lab.net/data/SCUT-FBP/

TABLE I
COMPARISON OF DIFFERENT NETWORKS WITH SINGLE-LABEL REGRESSION IN TERMS OF PC, MAE, AND RMSE

Network	PC	MAE	RMSE
ResNet-50	**0.8858**	**0.2871**	**0.3643**
ResNet-101	0.8763	0.2919	0.3748
VGGNet-19	0.8510	0.3233	0.4296

mance.

$$PC = \frac{\sum_{i=1}^{m}(\hat{l}_i' - \bar{\hat{l}}')(l_i' - \bar{l}')}{\sqrt{\sum_{i=1}^{m}(\hat{l}_i' - \bar{\hat{l}}')^2}\sqrt{\sum_{i=1}^{m}(l_i' - \bar{l}')^2}}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|l_i' - \hat{l}_i'|$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(l_i' - \hat{l}_i')^2} \qquad (4)$$

where l_i' and $\hat{l}_i' = \sum_{j=1}^{c} y_j p(y_j|\mathbf{x}_i'; \boldsymbol{\theta})$ are the ground truth (average score) and the predicted score for the i-th test image, respectively, and $\bar{l}'$ and $\bar{\hat{l}}'$ are the mean values of the ground truth and predicted scores over all test images. Higher values of PC and smaller values of MAE and RMSE indicate more accurate predictions.

C. Comparison Among Different Deep Architectures

To start with, we investigate the role of deep architectures in facial attractiveness computation in the context of traditional single-label regression. The extracted features and average-score labels of instances are fed into a standard regression model, which aims to learn a mapping between them. The attractiveness computation, in other words, is first formulated as a standard regression problem in order to evaluate the representative features learned from different deep architectures.

Various classical architectures of deep CNNs have been proposed for the ImageNet classification challenge. Transferring the pre-trained models to other vision-related problems has become increasingly common in recent years. Indicated by a previous work [30] that deep CNN architectures may critically affect the performance, we fine-tune three candidate models including ResNet-50, ResNet-101 and VGGNet-19 [35] on the augmented training data. The performance of the three networks is compared in Table I. It is notable that the baseline result in [24] achieved with raw images is a correlation of 0.83, rather than the highest correlation of 0.88 achieved by several inputting channels. As can be seen, all three networks achieve superior performance to the baseline due to the much deeper architectures than the six-layer CNN in [24]. The 50-layer ResNet performs the best, being slightly better than its 101-layer counterpart, and VGGNet performs the worst. To reveal the possible reasons, Fig. 5 presents the behaviors of the three networks during the training procedure. Without the residual structure, VGGNet may not be effectively trained due to the degradation problem described in [27], thus, the training error of VGGNet is the highest during most of the training period. Compared to ResNet-50, the slightly inferior performance of 101-layer ResNet could most likely be caused by its over-sophisticated architecture being applied to such an insufficient and imbalanced training dataset.

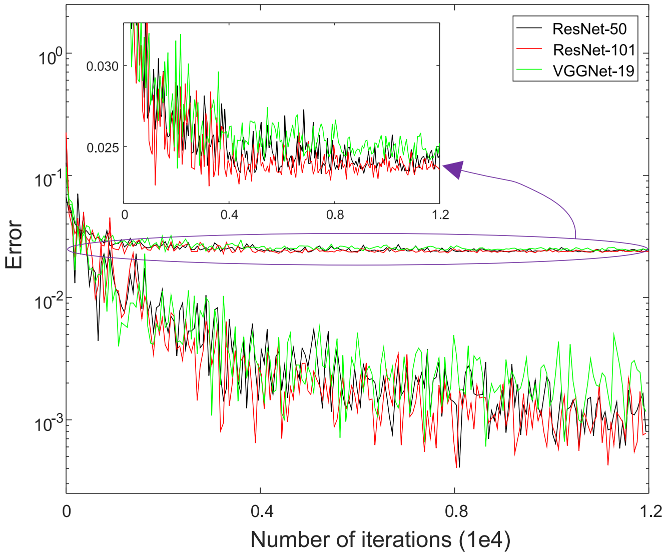


Fig. 5. Training behavior of the three networks. Thin curves denote the training error, and bold curves denote the testing error.

D. Effect of Label Distributions

We then identify the role of label distributions in facial attractiveness computation. Instead of the average score, the label distributions associated with instances are fed into a label distribution model. We also conduct preliminary experiments with existing LDL algorithms, including IIS-LLD [37], CPNN [37], LDSVR [39], and find that LDSVR performs the best. Table II compares the performance of three deep networks with our neural-network-based LDL algorithm and LDSVR, measured by six label-distribution-based distance/similarity metrics and three single-label-based accuracy metrics.

The first observation from Table II is that the ranks of the three networks are consistent with Table I. ResNet-50 performs the best with respect to all nine measures, ResNet-101 is slightly worse, and VGGNet is the worst. Hence, we use ResNet-50 to extract deep features in the following experiments.

The second observation is that by utilizing label distributions of instances and introducing LDL algorithms, the overall accuracy of the three networks is increased, in contrast to the single-label-based results shown in Table I, where the highest increase of 4% is achieved with our LDL algorithm. This reinforces the advantage of label distribution in that it provides more aesthetics-degree information than the conventional average score. Because of the LDL algorithm, each instance is expected to contribute to the learning of a number of attractiveness labels. Thus, the whole framework is able to deal with insufficient and incomplete training data, while still achieving an outstanding performance.

TABLE II
COMPARISON OF DIFFERENT NETWORKS WITH LDL IN TERMS OF NINE METRICS

Network		Chebyshev	Clark	Canberra	KL	Cosine	Intersection	PC	MAE	RMSE
ResNet-50	Our LDL	**0.1312**	1.1669	**1.9369**	**0.1046**	**0.9493**	**0.8516**	**0.9172**	**0.2174**	**0.2997**
	LDSVR	0.1603	**1.1626**	1.9637	0.1547	0.9297	0.8178	0.8977	0.2621	0.3453
ResNet-101	Our LDL	0.1371	1.1720	1.9534	0.1080	0.9482	0.8468	0.9156	0.2273	0.3024
	LDSVR	0.1635	1.1662	1.9719	0.1592	0.9269	0.8141	0.8947	0.2671	0.3596
VGGNet-19	Our LDL	0.1649	1.2024	2.0686	0.1563	0.9294	0.8077	0.8847	0.2883	0.3595
	LDSVR	0.1685	1.1705	1.9969	0.1707	0.9211	0.8081	0.8763	0.3052	0.3903

TABLE III
PERFORMANCE OF INDIVIDUAL FEATURES AND FEATURE-LEVEL FUSION

Feature	Chebyshev	Clark	Canberra	KL	Cosine	Intersection	PC	MAE	RMSE
Deep features	0.1312	1.1669	1.9369	0.1046	0.9493	0.8516	0.9172	0.2174	0.2997
Geometric features	0.1864	1.2072	2.1039	0.2300	0.8987	0.7822	0.7905	0.3771	0.5113
Fusion features	**0.1276**	**1.1384**	**1.8671**	**0.0977**	**0.9542**	**0.8559**	**0.9301**	**0.2127**	**0.2781**

TABLE IV
COMPARISON WITH THE STATE OF THE ART

Method	Feature	Modeling	PC	MAE	RMSE
Xie *et al.* [23]*	Handcrafted distances, Gabor	Single-label regression	0.6482	0.3931	0.5149
Liu *et al.* [41]	Data-driven ratios	Single-label regression	0.6938	0.4441	0.5768
Wang *et al.* [21]	One-layer autoencoder	Single-label regression	0.6406	0.9316	1.1233
Xu *et al.* [24]	Six-layer CNN	Single-label regression	0.8800	–	–
Chen *et al.* [4]	LBP, AAM, PCANet	Single-label regression	0.8369	0.3450	0.4511
Ours	ResNet, data-driven geometric features	LDL	**0.9301**	**0.2127**	**0.2781**

* Xie *et al.* [23] also obtained a correlation of 0.8187 with deep CNN features, which was further improved to 0.88 in [24].

The third observation is that our LDL model outperforms LDSVR on eight of the nine measures. This may be because our model is trained together with the feature learning module, whereas LDSVR has more parameters to learn, and such a small dataset may lead to overfitting. LDSVR appears to be slightly better with respect to the "Clark distance" because it could generate fewer near-zero outputs, making the metric more robust.

E. Effect of Feature Fusion

In our facial attractiveness computation framework, we are also interested in leveraging the power of a hybrid face representation including the high-level deep CNN features and low-level geometric features, to further boost the attractiveness computation accuracy. To this end, we train our LDL model with deep features from ResNet-50 and the extracted geometric features from IFS algorithm, respectively. Finally, a concatenated vector of these two types of features is fed into our LDL model. Table III presents the performance of individual features as well as their fusion.

It is not surprising that the deep features are rather discriminative and achieve significantly better results than geometric features. However, these two types of features emphasize different aspects of face representations. The deep features are good at extracting high-level abstract features based on color, texture, light, etc., whereas geometric features provide lower-level geometry information that ResNet might find difficult to learn. The fusion of them, therefore, can lead to more accurate and aesthetics-aware measures, which further advances the Pearson correlation to 0.93.

F. Comparison With the State of the Art

In this subsection, we compare our method against several recent works on the same SCUT-FBP dataset. The comparison with state of the art is given in Table IV, which shows that our method achieves a significant improvement over the others. There are two main reasons for this: more aesthetics-aware face representations and the utilization of label distributions. The first two methods in Table IV adopted low-level geometric and texture features like distances, ratios, and Gabor, while the next two methods learned multiple levels of face representations through shallow architectures, i.e., a few layers of deep networks. The fusion of these types of features is expected to derive a more aesthetics-aware representation. A representative work that constructed an optimal combination of traditional appearance features including LBP and AAM parameters as well as middle-level PCANet features was presented in [4]. The feature set, however, seems incomplete since all the included features are appearance-based descriptors, and the local geom-

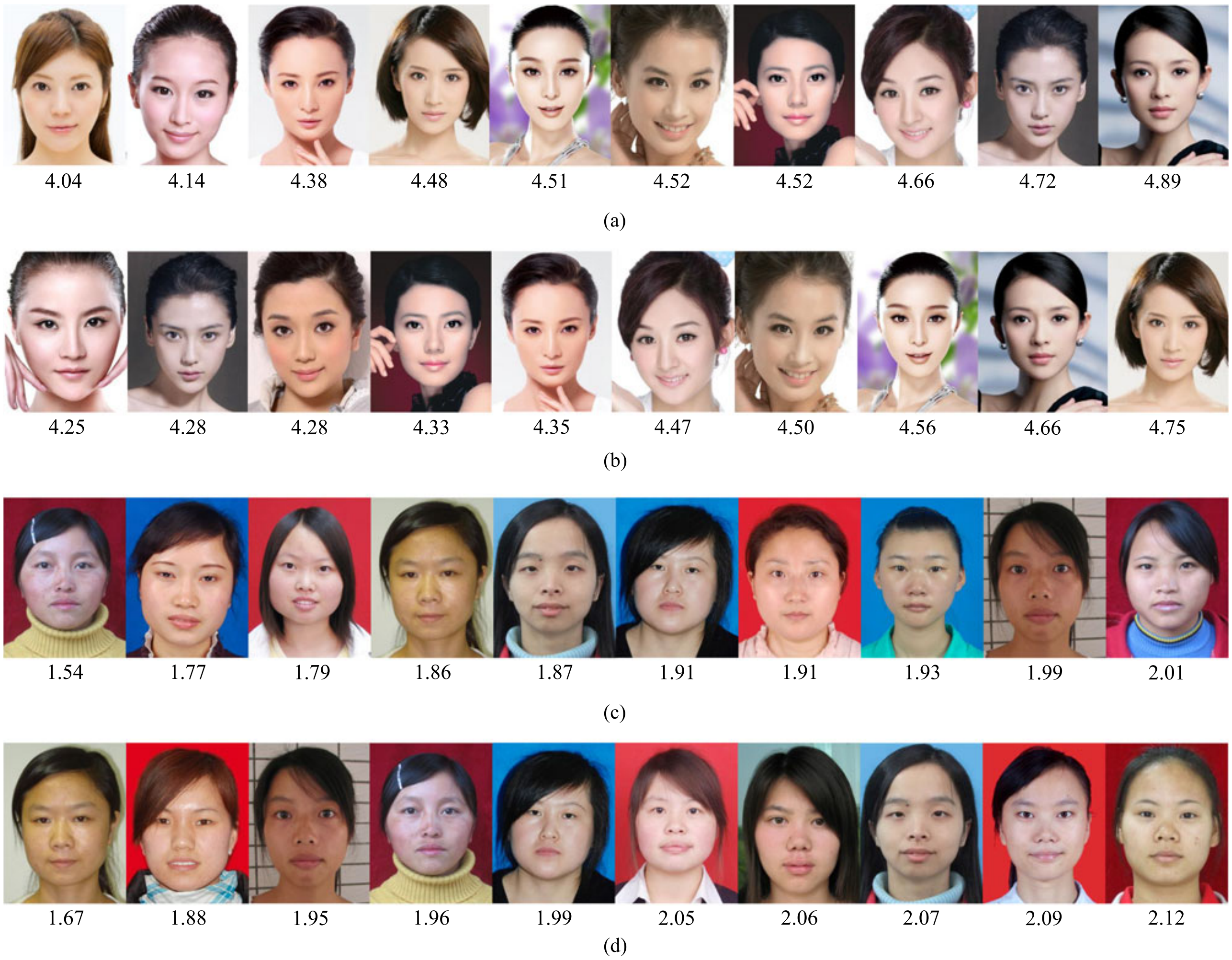


Fig. 6. The top ten test images rated by (a) human raters and (b) our method, and the bottom ten test images rated by (c) human raters and (d) our method.

etry of the face was not considered. In contrast, we leverage the power of a very deep residual network to learn higher-level features in conjunction with some low-level geometric features. All five comparable works were designed for exact attractiveness computation, with the average score as the ground truth, based on the training data that is not sufficient and complete enough to develop a powerful regression algorithm. Our work, instead, takes the advantage of the label distribution, making each training sample contribute to the learning of multiple attractiveness levels rather than simply its own class.

TABLE V
COMPARISON RESULTS ON THE DATASET IN [40]

Method	PC	MAE	RMSE
Liu *et al.* [40]	0.7420	0.5681	0.7300
Ours	**0.8115**	**0.4915**	**0.6220**

G. Further Discussion

To give an intuitive visualization of our final computational model, we present the top ten and bottom ten test images based on the ratings of human raters and our method in Fig. 6. As shown, the selected images by our model are highly correlated with the ones selected by human raters, especially in the case of the top ten images where eight of them are the same rated by human and the prediction. This qualitatively demonstrates that our model is able to generate human-like results. Some possible attractiveness cues may be observed from the comparison between the more and less attractive faces. Generally, faces ranked the highest have smoother and lighter skin, slimmer face with larger eyes, and better harmony in facial organs than those with the lowest scores.

We also evaluate our method on two other datasets provided by Liu *et al.* [40] and Mu [25]. Liu *et al.*'s dataset contains frontal images of 360 male and female faces with 48 attractiveness scores from 1–10. Mu's dataset is a collection of 250 facial images of both genders, labeled on a 7-point scale by 32 human raters. The hair above the forehead and other body parts below the neck have almost been removed in both datasets. We follow their data partition setting, and calculate the label distribution for each image based on the records of multiple raters. Similar augmentation techniques are also applied to enlarge the training data. Tables V and VI tabulate the comparison on these two

TABLE VI
COMPARISON RESULTS ON THE DATASET IN [25]

Method	PC	MAE	RMSE
Mu [25]	–	0.5200	–
Ours	**0.7878**	**0.4488**	**0.5746**

datasets. Apparently, our method also shows its effectiveness and advantages on different datasets.

V. CONCLUSION

This paper presents a novel computational model for facial attractiveness, which is formulated as a label distribution learning problem rather than conventional single-label (average score) supervised learning. In order to extract aesthetic-related face representations, a very deep residual network is utilized and adaptively transferred to this task. This is the first work to predict facial attractiveness by leveraging the power of such a deep architecture. After each instance is associated with a label (score) distribution, a neural-network-based LDL algorithm is adapted to learn a mapping between the extracted features and the label distribution. An end-to-end attractiveness learning framework is developed accordingly. Some discriminative geometric features selected by an incremental feature selection algorithm are also incorporated and fused with the deep features to boost the computational performance. Experimental results demonstrate that our model achieves significantly better performance than state-of-the-art methods. Possible attractiveness cues are also discovered from the comparison between the highly attractive and less attractive faces.

As stated before, most of the research in face attractiveness computation has so far focused on prediction models that rely on the averaged opinion score for a face. This of course can be directly derived from the label distribution by calculating the weighted mean value. It should be notable that with our predicted distribution, many possible results can be derived based on application scenarios. For a face image uploaded by a user, our model can provide its most likely attractiveness score by assigning the maximum description degree label, as well as its attractiveness confidence intervals by statistical analysis on the distribution. We can also estimate an opinion toll about how attractive this face is by presenting the label distribution since it is a direct representation of the public's opinion. In summary, the attractiveness label distribution is useful in a variety of applications that can benefit from a greater level of detail in understanding the attractiveness of a face as perceived by a diverse population.

Although using LDL can address insufficient and incomplete training data, the small size of datasets remains an outstanding problem in facial attractiveness computation research. We believe that with the future availability of large-scale and diverse (multiple ethnicities, ages, etc.) benchmarks, more effective deep networks and a diverse set of low-level features (geometric, appearance, etc.) can be considered to further enhance the attractiveness prediction. The investigation of applying new machine learning paradigms for attractiveness modeling (cost-sensitive learning, deep learning, etc.) would also be an interesting future topic.

REFERENCES

[1] A. Calder, G. Rhodes, M. Johnson, and J. Haxby, *The Oxford Handbook of Face Perception*. Oxford, U.K.: Oxford Univ. Press, 2011.

[2] M. Bashour, "History and current concepts in the analysis of facial attractiveness," *Plastic Reconstructive Surg.*, vol. 118, no. 3, pp. 741–756, 2006.

[3] H. Yan, "Cost-sensitive ordinal regression for fully automatic facial beauty assessment," *Neurocomputing*, vol. 129, pp. 334–342, 2014.

[4] F. Chen, X. Xiao, and D. Zhang, "Data-driven facial beauty analysis: Prediction, retrieval and manipulation," *IEEE Trans. Affective Comput.*, vol. PP, no. 99, p. 1, doi: 10.1109/TAFFC.2016.2599534.

[5] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski, "Data-driven enhancement of facial attractiveness," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 55–63, 2008.

[6] S. Melacci, L. Sarti, M. Maggini, and M. Gori, "A template-based approach to automatic face enhancement," *Pattern Anal. Appl.*, vol. 13, no. 3, pp. 289–300, 2010.

[7] D. Guo and T. Sim, "Digital face makeup by example," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 73–79.

[8] L. Liu *et al.*, "Wow! You are so beautiful today!" *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 11, no. 1s, pp. 1–22, 2014.

[9] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, Dec. 2013.

[10] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, Nov. 2015.

[11] D. Zhang, F. Chen, and Y. Xu, *Computer Models for Facial Beauty Analysis*, vol. 4. Basel, Switzerland: Springer, 2016.

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

[13] A. Laurentini and A. Bottino, "Computer analysis of face beauty: A survey," *Comput. Vis. Image Understanding*, vol. 125, pp. 184–199, 2014.

[14] A. Kagian, G. Dror, T. Leyvand, D. Cohen-Or, and E. Ruppin, "A humanlike predictor of facial attractiveness," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 649–656.

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

[16] P. Aarabi, D. Hughes, K. Mohajer, and M. Emami, "The automatic measurement of facial beauty," in *Proc. IEEE Int. Conf. Syst. Man, Cybern.*, 2001, vol. 4, pp. 2644–2647.

[17] H. Gunes and M. Piccardi, "Assessing facial beauty through proportion analysis by image processing and supervised learning," *Int. J. Hum.-Comput. Stud.*, vol. 64, no. 12, pp. 1184–1199, 2006.

[18] Y. Eisenthal, G. Dror, and E. Ruppin, "Facial attractiveness: beauty and the machine," *Neural Comput.*, vol. 18, no. 1, pp. 119–142, 2006.

[19] J. Whitehill and J. R. Movellan, "Personalized facial attractiveness prediction," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2008, pp. 1–7.

[20] A. Bottino and A. Laurentini, "The intrinsic dimensionality of attractiveness: A study in face profiles," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. New York, NY, USA: Springer, 2012, pp. 59–66.

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

[22] J. Gan, L. Li, Y. Zhai, and Y. Liu, "Deep self-taught learning for facial beauty prediction," *Neurocomputing*, vol. 144, pp. 295–303, 2014.

[23] D. Xie, L. Liang, L. Jin, J. Xu, and M. Li, "SCUT-FBP: A benchmark dataset for facial beauty perception," in *Proc. IEEE Int. Conf. Syst. Man, Cybern.*, 2015, pp. 1821–1826.

[24] J. Xu, L. Jin, L. Liang, Z. Feng, and D. Xie, "A new humanlike facial attractiveness predictor with cascaded fine-tuning deep learning model," arXiv:1511.02465, 2015.

[25] Y. Mu, "Computational facial attractiveness prediction by aesthetics-aware features," *Neurocomputing*, vol. 99, pp. 59–64, 2013.

[26] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[28] 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.

[29] B. C. Davis and S. Lazebnik, “Analysis of human attractiveness using manifold kernel regression,” in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 109–112.

[30] 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.

[31] J. Deng *et al.*, “Imagenet: A large-scale hierarchical image database,” in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

[32] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[33] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1717–1724.

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

[35] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Represent.*, 2015.

[36] R. Rothe, R. Timofte, and L. Van Gool, “Some like it hot—Visual guidance for preference prediction,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5553–5561.

[37] X. Geng, C. Yin, and Z. H. Zhou, “Facial age estimation by learning from label distributions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.

[38] Z. Zhang, M. Wang, and X. Geng, “Crowd counting in public video surveillance by label distribution learning,” *Neurocomputing*, vol. 166, pp. 151–163, 2015.

[39] X. Geng and P. Hou, “Pre-release prediction of crowd opinion on movies by label distribution learning,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 3511–3517.

[40] S. Liu, Y.-Y. Fan, Z. Guo, A. Samal, and A. Ali, “A landmark-based data-driven approach on 2.5d facial attractiveness computation,” *Neurocomputing*, vol. 238, pp. 168–178, 2017.

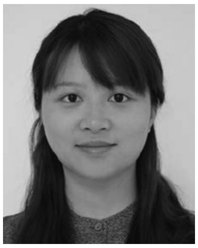
[41] S. Liu, Y. Fan, Z. Guo, and A. Samal, “2.5D facial attractiveness computation based on data-driven geometric ratios,” in *Intelligence Science and Big Data Engineering. Image and Video Data Engineering*. New York, NY, USA: Springer, 2015, pp. 564–573.

[42] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis*. Chichester, U.K.: Wiley, 1998.

[43] M. Efroymson, “Multiple regression analysis,” *Math. Methods Digit. Comput.*, vol. 1, pp. 191–203, 1960.

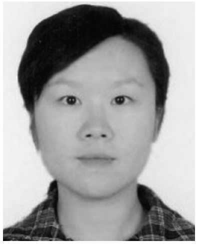
[44] Y. Jia *et al.*, “Caffe: Convolutional architecture for fast feature embedding,” in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

Yang-Yu Fan received the M.S. degree in electromechanical engineering from the Shaanxi University of Science and Technology, Xi’an, China, and the Ph.D. degree in acoustics signal processing from Northwestern Polytechnical University, Xi’an, China, in 1992 and 1999, respectively. He is currently a Professor with the School of Electronics and Information, Northwestern Polytechnical University, Xi’an, China. He has authored or coauthored numerous papers that appeared in various publications, including *Neurocomputing*, *Signal Processing*, *Image and Vision Computing*, IEEE Transactions on Antennas and Propagation, *Multimedia Tools and Applications*, etc. His research interests include image processing, pattern recognition, and virtual reality.



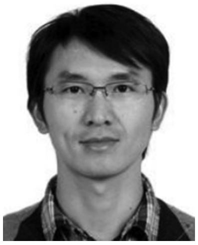
Shu Liu received the B.S. degree in electronics and information engineering from Northwestern Polytechnical University, Xi’an, China, in 2011. She is currently working toward the Ph.D. degree at the School of Electronics and Information, Northwestern Polytechnical University. From 2013–2015, she was a Visiting Research Scholar with the Department of Computer Science and Engineering, University of Nebraska–Lincoln, Lincoln, NE, USA. Her research interests include image processing, computer vision, machine learning, and their applications to facial attractiveness analysis.

Bo Li received the B.S. degree in electronics and information engineering from Northwestern Polytechnical University, Xi’an, China, in 2011. He is currently working toward the Ph.D. degree at the School of Electronics and Information, Northwestern Polytechnical University. From 2013–2015, he was a Visiting Research Scholar with the School of Computer Science, University of Adelaide, Adelaide, SA, Australia. His research interests include deep learning and computer vision.

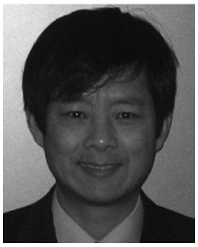


Zhe Guo received the M.S. and Ph.D. degrees in computer science from Northwestern Polytechnical University, Xi’an, China, in 2008 and 2012, respectively. She is currently a Lecturer with the School of Electronics and Information, Northwestern Polytechnical University. She has authored or coauthored more than 30 papers in international conferences and journals. Her research interests include pattern recognition, computer vision, and virtual reality.

Ashok Samal received the B.Tech. degree in computer science from the Indian Institute of Technology, Kanpur, Kanpur, India, and the Ph.D. degree from the University of Utah, Salt Lake City, Utah, USA, in 1983 and 1988, respectively. He is currently a Professor with the Department of Computer Science and Engineering at the University of Nebraska–Lincoln, Lincoln, NE, USA. He has authored or coauthored more than 100 papers in international conferences and journals. His research interests include image understanding, computer vision, data mining, and geospatial computation.



Jun Wan received the B.S. degree from the China University of Geosciences, Beijing, China, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2008 and 2015, respectively. Since January 2015, he has been an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, machine learning, especially for gesture and action recognition, and facial attribution analysis (i.e., age estimation, facial expression, gender, and race classification). He has published papers in top journals, such as JMLR, TPAMI, TIP, TCYB, and TOMM. He has served as a Reviewer on several top journals and conferences, such as JMLR, TPAMI, TIP, TMM, TSMC, PR, ICPR2016, CVPR2017, ICCV2017, and FG2017.



Stan Z. Li received the B.Eng. degree from Hunan University, Changsha, China, the M.Eng. degree from the National University of Defense Technology, Changsha, China, and the Ph.D. degree from the University of Surrey, Guildford, U.K. He is currently a Professor and the Director with the Center for Biometrics and Security Research, Institute of Automation, Chinese Academy of Sciences, Beijing, China. From 2000 to 2004, he was a Researcher with Microsoft Research Asia. Prior to that, he was an Associate Professor with Nanyang Technological University, Singapore. He has authored or coauthored more than 200 papers in international journals and conferences and edited eight books. His research interests include pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance. Prof. Li is a member of the IEEE Computer Society. He was an Associate Editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence and is the Acting Editor-in-Chief for the Encyclopedia of Biometrics. He was the Program Cochair for the International Conference on Biometrics 2007 and 2009 and has been involved in organizing other international conferences and workshops in the fields of his research interests. He was elevated to IEEE fellow for his contributions to the fields of face recognition, pattern recognition and computer vision.