# Beauty3DFaceNet: Deep geometry and texture fusion for 3D facial attractiveness prediction

Qinjie Xiao[a], You Wu[a], Dinghong Wang[b], Yong-Liang Yang[c], Xiaogang Jin[a,*]

[a] State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, China
[b] New York University, United States
[c] University of Bath, United Kingdom

A B S T R A C T

We present Beauty3DFaceNet, the first deep convolutional neural network to predict attractiveness on 3D faces with both geometry and texture information. The proposed network can learn discriminative and complementary 2D and 3D facial features, allowing accurate attractiveness prediction for 3D faces. The main component of our network is a fusion module that fuses geometric features and texture features. We further employ a novel sampling strategy for our network based on a prior of facial landmarks, which improves the performance of learning aesthetic features from a face point cloud. Comparing to previous work, our approach takes full advantage of 3D geometry and 2D texture and does not rely on handcrafted features based on highly accurate facial characteristics such as feature points. To facilitate 3D facial attractiveness research, we also construct the first 3D face dataset ShadowFace3D, which contains 6,000 high-quality 3D faces with attractiveness labeled by human annotators. Extensive quantitative and qualitative evaluations show that Beauty3DFaceNet achieves a significant correlation with the average human ratings. This validates that a deep learning network can effectively learn and predict 3D facial attractiveness.

## 1. Introduction

Psychological studies have demonstrated that the human face plays a significant role in conveying emotion and making a first impression. Hence three-dimensional facial attractiveness (defined as the aesthetic level of portraits to human raters [1]) prediction is extremely important for many applications [2], such as cosmetic surgery [3], face design [4], and entertainment. Researches show that the attractiveness of face is determined by frontal portrait, the profile view, and the combination of them [5]. In addition, cosmetic science research reveals that cosmetics can improve facial attractiveness without changing the underlying 3D facial geometry [6]. These results imply that both face geometry and texture contribute to its attractiveness.

Many works have been proposed for face attractiveness prediction, including CNN-based methods using frontal portrait images [7,8], learning-based methods using the landmarks of frontal and profile images [9], and methods using aesthetic criteria based on a sparse set of facial landmarks that are carefully specified [5], just to name a few. However, existing methods cannot handle 3D face data with much more abundant information (point cloud + texture image) and various cosmetic design details, which are particularly useful for cosmetic surgery. A naive solution is to employ PointNet++ [10] to predict attractiveness by learning features in sampled point sets. We note that human eyes have different perceptual sensitivity to different face regions [11]. Although PointNet++ can effectively distinguish geometric shapes, it has intrinsic limitations in handling 3D faces because 3D faces differ mainly in details rather than the overall shape. For one thing, the relatively sparse face point cloud contains valuable geometric information but lacks rich texture information that is vital for assessing facial attractiveness. For another thing, the dense texture image is only a 2D mapping from a 3D face without the underlying 3D geometry. Then how to construct a new discriminative 3D face representation to encode both point cloud and texture image, and is sensitive as human perception, such that various 3D cosmetic facial designs can be assessed beyond a single face view becomes a new challenge. Moreover, for attractiveness assessment research, there is a lack of 3D face dataset with annotations reflecting public aesthetic crite-

ria. Ideally, the dataset should include a wide variety of cosmetic designs by professional cosmetic surgery designers to facilitate real applications.

To this end, we propose Beauty3DFaceNet, a prior-inspired convolutional neural network that computes 3D face attractiveness automatically. The proposed Beauty3DFaceNet comprises a newly developed 3DFacePointNet++ module, a ResNet module, and a fusion module. It constructs a discriminative 3D face representation by seamlessly fusing the geometric and texture features extracted by 3DFacePointNet++ and ResNet modules, respectively. The two types of features complement each other to encode the geometry and texture of a 3D face. As a result, the proposed 3D face representations can distinguish not only geometry but also texture variations. In particular, the features of a 3D face with various cosmetic designs, some of which are hard to be observed from the front or profile views, can be faithfully extracted by the 3DFacePointNet++ module.

Our 3DFacePointNet++ utilizes facial landmark priors to simulate the perceptual sensitivity of human eyes to receive more information from sensitive regions of human faces. This can be viewed as a region of the interest (ROI)-based sampling using facial prior. Even though facial contour functions a vital role during facial attractiveness assessment, it is low-frequency, and requires less 3D points to describe facial contour compared to facial features. Different from PointNet++, we sample uneven point clouds instead of uniform point clouds. Besides, we employ a $k$-NN algorithm to group the neighbors of sample points with different densities adaptively. This strategy improves the performance and robustness of Beauty3DFaceNet. Moreover, it only requires facial landmarks with much lower accuracy compared to [5]. The landmarks are taken as input rather than being detected in Beauty3DFaceNet as it is not the core of our network. This allows us to use any suitable facial landmark detector without modifying the structure of Beauty3DFaceNet.

We further present ShadowFace3D (SF3D), a 3D face dataset with 6,000 high-quality 3D faces collected from people who care more about their facial attractiveness in beauty salons and plastic surgery hospitals. The dataset also contains attractiveness scores labeled by multiple human annotators to fulfill the public aesthetic criterion. Note that some of the 3D faces in SF3D are carefully designed on several parts such as eyes, nose, jaw, and cheeks, which improve facial attractiveness from several aspects that are hard to be observed from a single view.

We evaluate the proposed 3D facial attractiveness assessment method based on the unique ShadowFace3D dataset. The ablation study is carefully designed to validate the effectiveness of fusing geometric and texture features, along with the prior-based sampling strategy. The comparison with current work demonstrates the state-of-the-art performance of Beauty3DFaceNet.

Overall our work makes the following major contributions:

- We propose the first deep learning network, called Beauty3DFaceNet, for 3D facial attractiveness assessment. It integrates facial geometry, facial texture, and facial prior and computes a more convincing 3D facial attractiveness score like human raters.
- We create a 3D facial attractiveness dataset ShadowFace3D, which contains sufficient information such as the point clouds, texture images, and texture mappings of 3D faces as well as their public aesthetic criteria. Specifically, it contains the original and the lifted face pair of the same person, designed by professional designers of plastic surgeons. It is the first 3D face database for attractiveness assessment.
- We present 3DFacePointNet++, a new network based on facial landmark priors to simulate the perceptual sensitivity of human eyes, which improves the performance of the Beauty3DFaceNet.

## 2. Related work

**2D image-based approach.** Early facial attractiveness assessment relies on handcrafted features (e.g., geometry, texture, color) and holistic descriptors. Geometric features are mainly based on facial landmark positions. For instance, the distances between landmarks and their ratios [12–16] are mostly used. Other geometric properties are pre-defined based on heuristics and classical rules of beauty, such as golden ratios, the facial fifths and thirds, and the symmetry theory [12,15,16]. Texture features, such as Gabor filter responses, local binary patterns (LBPs), and skin smoothness indicators [14,17], are widely investigated. Besides, the active appearance model (AAM) parameters, which encode both facial shape and texture information, are employed in [18]. Color features, including color symmetry, hue/saturation/value (HSV) coordinates, and color distribution [13,14], are also effective in practice. Holistic descriptors, such as Eigenface [13,17], face manifold [19], and face shape model [20], are proved to be useful due to containing 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 explicitly defined and extracted from images. With the development of deep learning, it is possible to learn higher-level facial feature representations and apply them for facial attractiveness assessment. Gray et al. [21] first propose a CNN-like model to extract high-level features for facial attractiveness prediction. Various deep learning methods are applied for the facial attractiveness prediction task, such as self-taught learning [22], psychologically inspired convolutional neural network (PI-CNN) [23], feature combination [18], label distribution learning (LDL) [24], multi-task learning [7], etc. Deep learning methods outperform traditional approaches due to the ability to effectively learning high-level features. Inspired by this, our scheme includes both image and point cloud CNN modules, taking advantage of trained high-level features in both 2D and 3D.

**3D shape-based approach.** Compared to 2D image-based approaches, the research of 3D facial attractiveness prediction is much less explored. Recently, face reshaping methods [25,26] are presented to generate more shapely 3D faces. However, they did not evaluate facial attractiveness. O'Toole et al. [27] present a PCA-based model to analyze the effect of averageness of facial attractiveness using 2D texture and 3D shape. Kim et al. [28] use symmetric deformation to enhance facial attractiveness. Based on a set of 3D face landmarks, Liao et al. [5] develop a scoring system through a set of rules, such as local and global symmetrization, frontal facial proportion via neoclassical canons and golden ratios, and facial angular profile proportion. Xu et al. [29] predict personality trait based on a 2.5D face feature model. Note that Liu et al. [9,30] introduce a landmark-based data-driven approach for multi-view (frontal and profile view) facial images. Most of these methods use sparse information (facial landmarks) and ignore texture and dense geometric information, which is vital for facial attractiveness prediction.

**3D CNN for point clouds.** CNN-based classification methods have gained popularity due to their ability to learn mid-level and high-level features. As point cloud is a particularly important type of geometric representation captured by 3D scanners, there exist various works focusing on adapting 3D CNN for point clouds. Qi et al. [31] utilize max-pooling as the symmetric function to solve unordered point clouds and generate global features for point cloud learning. Qi et al. [10] further present PointNet++ to encode local features. Li et al. [32] present pointCNN that estimates a $\mathcal{X}-$transform to deal with irregular and unordered properties of point clouds. Wu et al. [33] propose pointconv that uses dynamic

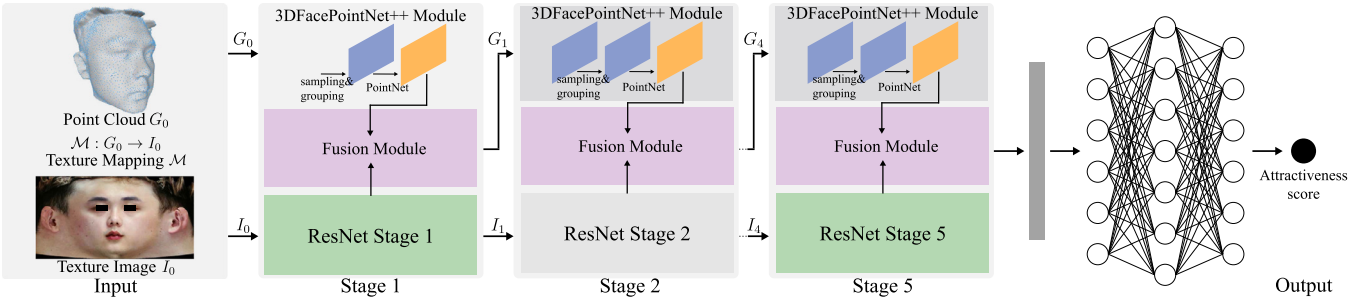



**Fig. 1.** The structure of the proposed Beauty3DFaceNet for 3D facial attractiveness prediction. It contains 5 stages for learning geometric and textural features followed by a fully connected layer to predict attractiveness score. Each stage consists of a 3DFacePointNet++ module with a novel sampling strategy based on facial landmarks, a ResNet module, and a fusion module. The input 3D face $(\boldsymbol{G}_0, \boldsymbol{I}_0, \mathcal{M}: \boldsymbol{G}_0 \rightarrow \boldsymbol{I}_0)$ comprises a point cloud $\boldsymbol{G}_0$, a texture image $\boldsymbol{I}_0$, and a texture mapp $\mathcal{M}: \boldsymbol{G}_0 \rightarrow \boldsymbol{I}_0$. Stage 1 processes $\boldsymbol{G}_0$ and $\boldsymbol{I}_0$ using the 3DFacePointNet++ module and the ResNet module, respectively. The generated geometric features and texture features are seamlessly fused using the fusion module based on the texture map, resulting in the intermediate geometric and textural features $(\boldsymbol{G}_1, \boldsymbol{I}_1, \mathcal{M}: \boldsymbol{G}_1 \rightarrow \boldsymbol{I}_1)$ for the next stage to be processed similarly. After stage 5, Beauty3DFaceNet generates a high-level face feature representation fed into a fully connected layer and finally outputs an attractiveness score.

filters to handle point clouds. These methods can only deal with point clouds that are too sparse for our facial attractiveness analysis. Xu et al. [34] fuse the last level of features of PointNet++ and ResNet for 3D object detection. Nevertheless, their method does not leverage the inherent connection between point cloud and texture image. To overcome the above limitations, a 2D and 3D fusion module is designed in Beauty3DFaceNet to seamlessly fuse point cloud and texture image based on the texture mapping in-between. Moreover, inspired by [35], we use facial landmarks as prior to guide point cloud sampling, such that our network can handle input data with sufficient information efficiently and robustly.

## 3. Deep 3D facial attractiveness prediction

In this section, we elaborate the details of our deep-learning-based 3D face attractiveness prediction network - Beauty3DFaceNet, which predicts facial attractiveness by fusing the features extracted from the 3D face point cloud and its corresponding texture image. As shown in Fig. 1, Beauty3DFaceNet consists of 5 stages for learning geometric and textural features. Each stage has three constituent modules: a 3DFacePointNet++ module that extracts point cloud features, a ResNet module that extracts texture features, and a fusion module that seamlessly combines point cloud features and texture features. We first describe Beauty3DFaceNet in detail in Section 3.1. Then we present how to perform effective point cloud sampling based on facial landmark prior in Section 3.2. Finally, we introduce the new 3D face dataset - ShadowFace3D with annotated facial attractiveness scores in Section 3.3.

### 3.1. Beauty3DFaceNet structure

Given a 3D face $\boldsymbol{F} = \{\boldsymbol{G}_0, \boldsymbol{I}_0, \mathcal{M}\}$ consists of face geometry $\boldsymbol{G}_0$, face texture $\boldsymbol{I}_0$, and the texture map in-between $\mathcal{M}: \boldsymbol{G}_0 \rightarrow \boldsymbol{I}_0$, our goal is to leverage deep neural networks to learn representative features from the 3D face, which can be used for accurate facial attractiveness prediction. Since digitized human face contains both geometry and texture, how to extract and complement features in 3D and 2D effectively becomes the fundamental problem.

Here we assume face geometry $\boldsymbol{G}_0$ is represented as a 3D point clouds, which comprises a set of unorganized 3D points. Such a representation is general as it does not require point connectivity within each face and point correspondences between faces. Besides, existing point-based CNN such as PointNet++ [10] can be applied here. Although face texture $\boldsymbol{I}_0$ is simply a 2D image and 2D CNN can be directly employed, special consideration needs to be taken to make the learned 2D (texture) features complementary to 3D (geometry) features, allowing consolidated attractiveness prediction. Thanks to the texture map $\mathcal{M}$ between 2D texture and 3D geometry when reconstructing 3D digital faces, we can easily fuse 2D and 3D features corresponding to the same location on a 3D face.

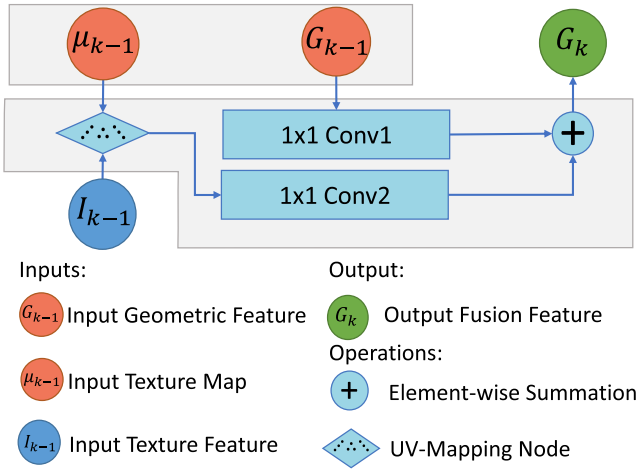



**Fig. 2.** Flowchart that illustrates the process of the fusion module. After feature learning with 3DFacePointNet++ (which generates a texture map $\boldsymbol{\mu}_{k-1}$ and a geometric feature $\boldsymbol{G}_{k-1}$) and ResNet (which generates a texture feature $\boldsymbol{I}_{k-1}$), a feature fusion process takes place. Taking $\boldsymbol{\mu}_{k-1}$ and $\boldsymbol{I}_{k-1}$ as input, the corresponding texture feature is extracted with the help of the UV-Mapping Node, and fused with the geometric feature $\boldsymbol{G}_{k-1}$ to generate the new geometric feature $\boldsymbol{G}_k$ for the next feature learning stage. In the UV-Mapping Node, each point in the down-sampled point cloud has UV coordinates that encode the correspondence between it and its location in the input texture image, and thus it can be directly mapped into the down-sampled texture features. Specially, we utilize two different $1 \times 1$ convolution layers to normalize the geometric feature and the corresponding texture feature, then fuse them using element-wise summation.

Based on the above analysis, we design the proposed Beauty3DFaceNet with five feature learning stages and a fully connected layer for facial attractiveness prediction (see Fig. 1). Each feature learning stage in Beauty3DFaceNet has three modules including a 3DFacePointNet++ module, a ResNet module, and a fusion module (Fig. 2). Taking stage 1 as an example, the input is composed by 3D face features $\boldsymbol{F} = \{\boldsymbol{G_0}, \boldsymbol{I_0}, \mathcal{M}\}$. More specifically, $\boldsymbol{G_0}$ is the current geometric feature. It contains a point cloud and the corresponding learned 3D features at individual points (for stage 1, it is just the input face point cloud). Due to the close correlation, we use one variable to simplify the notation. $\boldsymbol{I_0}$ is the current textural feature (for stage 1 it is just the input face texture) fed into the ResNet module. $\mathcal{M}: \boldsymbol{G_0} \rightarrow \boldsymbol{I_0}$ is the texture map between

point cloud and texture image. The 3DFacePointNet++ module outputs the intermediate geometric feature composed of the downsampled point cloud (as well as the corresponding texture map) and the corresponding 3D features. The ResNet module outputs the intermediate texture feature $\boldsymbol{I}_1$. The fusion module (Fig. 2) uses the texture map $\mathcal{M}$ to extract the corresponding texture feature of the intermediate geometric feature and fuses these two types of features to generate new geometric features $\boldsymbol{G_1}$. Eventually stage 1 extracts a higher level 3D face feature $\boldsymbol{F_1} = \{\boldsymbol{G_1}, \boldsymbol{I_1}, \mathcal{M}\}$ for the next feature learning stage. The next four feature learning stages (stage 2-5) are performed in the same way.

For predicting facial attractiveness, the loss function of Beauty3DFaceNet is defined as:

$$\begin{aligned} s &= Beauty3DFaceNet(\boldsymbol{F}), \\ L &= \sum ||s^* - s||^2, \end{aligned} \tag{1}$$

where $\boldsymbol{F}$ is an example 3D face in the dataset, $s$ is the predicted facial attractiveness score of $\boldsymbol{F}$ by Beauty3DFaceNet, and $s^*$ is the corresponding ground-truth score.

### 3.2. 3DFacePointNet++

Different from the original PointNet++, our 3DFacePointNet++ module processes the input point cloud using a facial prior-based sampling strategy. PointNet++ uses farthest point sampling (FPS) that is more suitable for complete and uniform sampling of unevenly distributed points. However, for judging face attractiveness, more attention tends to be paid on the inner part of the face. FPS simply ignores the region of interests and may not sample enough discriminative points with rich information for attractiveness prediction. To solve this problem, we propose a novel sampling strategy by using facial landmarks as a prior, such that points closer to the landmarks have a higher probability of being sampled. We employ $K$ nearest neighbor search ($k$-NN), an alternative range query method to adaptively group the neighbors of sample points with different densities.

Given a point $\boldsymbol{p}$ in the point cloud, and a landmark set $\boldsymbol{M}$, we define the minimal Euclidean distance between $\boldsymbol{p}$ and $\boldsymbol{m} \in \boldsymbol{M}$ as the distance from $\boldsymbol{p}$ to $\boldsymbol{M}$. The distance is further normalized as:

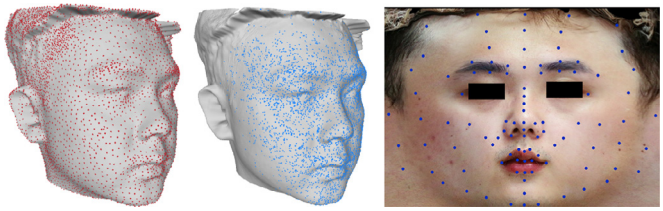
$$D(\boldsymbol{M}, \boldsymbol{p}) = \frac{1}{DP} \min_{\boldsymbol{m} \in \boldsymbol{M}} ||\boldsymbol{m} - \boldsymbol{p}||, \tag{2}$$

where $DP$ is the pupil distance between left and right eyes. Based on the distance defined above, we introduce an empirical prior for point sampling preference as follows:

$$P_{prior}(\boldsymbol{M}, \boldsymbol{p}) \propto e^{\frac{1}{\lambda^2} \frac{D(\boldsymbol{M}, \boldsymbol{p})^2}{2\sigma^2}}, \tag{3}$$

where $\sigma^2$ is the variance of $D(\boldsymbol{M}, \boldsymbol{p})$, and the parameter $\lambda$ determines the sampling density variation with respect to facial landmarks. We find that using facial prior greatly improves the performance and robustness of attractiveness prediction (see Section 4.3). Fig. 3 (a) and (b) demonstrate the differences between sampling results with and without the proposed prior.

Note that instead of relying on high-quality facial landmarks to extract aesthetics-aware features for facial attractiveness analysis [1,5], we only use facial landmarks to adaptively sample face point cloud. The network reliably learns the high-level features of a 3D face, which dramatically reduces the facial landmarks' quality requirement. Thus the features extracted by Beauty3DFaceNet are compatible with various face alignment algorithms [35,36]. Fig. 3 (c) shows the set of facial landmarks used to represent the face structure such as facial features, forehead, jaw, and cheeks, etc.



(a) FPS (b) FPBS (c) The landmarks used for FPBS

**Fig. 3.** Examples of (a) farthest point sampling (FPS), (b) facial prior-based sampling (FPBS), and the facial landmarks used for prior-based sampling (c). This landmark set contains facial landmarks located on eyes, eyebrows, nose, mouth, jaw, etc., and landmarks on face contour.

### 3.3. ShadowFace3D dataset

We construct a 3D face dataset called ShadowFace3D with 6,000 3D faces of Asian males and females age from 18 to 45 with academic usage agreement. Each face contains a 3D point clouds (4096 points), a 3D texture image, the texture map in-between, and an annotated attractiveness score. Twenty raters score each 3D face by dragging a slider from 1 to 5, where a value of 1 represents the least attractive, and 5 represents the most attractive. We choose the average rating score as the attractiveness score for each 3D face. We use Bellus3D [37] to scan faces indoors based on the built-in light control of the scanner. The 3D faces are collected from people who care more about their facial attractiveness in beauty salons and plastic surgery hospitals. We also collect a certain number of original 3D faces and corresponding lifted faces designed by cosmetic doctors and plastic surgeons for the application of cosmetic surgery. There are 500 pairs of such 3D faces in our dataset. The plastic surgeons usually adjust the nose, jaw, cheeks to make a face more shapely. Such changes are hard to be observed from a single view.

## 4. Experiments

In this section, we present the experimental results to demonstrate that the proposed Beauty3DFaceNet can effectively utilize the 3D face data and learn discriminative features, leading to more accurate attractiveness prediction results. We first describe how we prepare data based on the ShadowFace3D (SF3D) dataset for various facial attractiveness evaluations. Then we demonstrate the implementation details and the performance of our method. After that, we conduct ablation studies to evaluate the effectiveness of the design choices of the proposed Beauty3DFaceNet, including feature learning modules and the novel facial prior based sampling strategy. Finally, we compare Beauty3DFaceNet with other state-of-the-art methods to show the improvement in facial attractiveness prediction.

### 4.1. Data preparation

To extensively evaluate our work, we prepare experimental data based on the SF3D dataset as follows. We split the 6,000 3D faces in SF3D into a training set (4,200 + (300 + 300)), a validation set (400 + (100 + 100)) and a test set (400 + (100 + 100)), which are employed for training, ablation studies, and comparisons. Here, "(300 + 300)" means 300 faces and their lifted ones. For conducting ablation studies with different settings, we construct variant sets of experimental data based on SF3D, including only point clouds (PC), only texture images (TI), point clouds with point colors (PC+RGB), both point clouds and texture images without texture maps (PC+TI), and data with all information (PC+TI+TM).

**Table 1**
Beauty3DFaceNet architecture.

| Stage | Input size | Point cloud and image convolution | Intermediate output size | Texture map | Fusion operation | |
|---|---|---|---|---|---|---|
| | | | | | Convolution layer | Element-wise summation |
| 1 | 4096 × 5 | SA(2048, 0.10, [ 64, 64, 128]) | 2048 × (5 + 128) | | 1 × 128 × 128 | 2048 × (5 + 128) |
| | 224 × 224 × 3 | ResNet Conv1 | 112 × 112 × 64 | 2048 × 64 | 1 × 64 × 128 | |
| 2 | 2048 × (5 + 128) | SA(1024, 0.15, [128, 128, 256]) | 1024 × (5 + 256) | | 1 × 256 × 256 | 1024 × (5 + 256) |
| | 112 × 112 × 64 | ResNet Conv2 | 56 × 56 × 256 | 1024 × 256 | 1 × 256 × 256 | |
| 3 | 1024 × (5 + 256) | SA(512, 0.20, [256, 256, 512]) | 512 × (5 +512) | | 1 × 512 × 512 | 512 × (5 + 512) |
| | 56 × 56 × 256 | ResNet Conv3 | 28 × 28 × 512 | 512 × 512 | 1 × 512 × 512 | |
| 4 | 512 × (5 + 512) | SA(128, 0.40, [512, 512, 1024]) | 128 × (5 +1024) | | 1 × 1024 × 1024 | 128 × (5 +1024) |
| | 28 × 28 × 512 | ResNet Conv4 | 14 × 14 × 1024 | 128 × 1024 | 1 × 1024 × 1024 | |
| 5 | 128 | SA([1024, 1024,2048]) | 1 × 2048 | | 1 × 2048 × 2048 | 1 × 2048 |
| | 14 × 14 × 1024 | ResNet Conv5 and average pool | 1 × 1 × 2048 | | 1 × 2048 × 2048 | |
| | 1 × 2048 | FC(512,0.5) → FC(256, 0.5) → FC(1) | | | | Attractiveness Score |

### 4.2. Implementation details

*Network architecture*. Table 1 details the network architecture of Beauty3DFaceNet. For clarity, we use the following notations to present our out network architecture. SA($k, r, [l_1, ...l_n]$) is a set abstraction (SA) layer of 3DFacePointNet++ (a variant PointNet++ using our sampling prior), which samples $k$ local regions with ball radius $r$ using PointNet of $n$ fully connected layers with width $l_i (i = 1, ..., n)$. FC($m, p$) represents a fully connected layer (width $m$) followed by a dropout layer (drop ratio $p$). Please refer to PointNet++ [10] for more details. We utilize ResNet50 [38] (which is fixed as constant in Beauty3DFaceNet) pre-trained on ImageNet [39] to extract texture features. We use "ResNet Conv1,..., ResNetConv5" to represent 5 convolutional stages of ResNet. More details can be found in [38]. Specifically, each geometric feature's first five dimensions comprise the 3D position $(x, y, z)$ and texture coordinates $(u, v)$ that is only used for set abstraction and texture mapping, respectively.

*Performance*. The size of our model is 176MB. The forward time of our model is 228ms (with batch size 16 using PyTorch 1.2 on NVIDIA GTX 1080Ti).

### 4.3. Ablation studies

Here we present two ablation studies to validate the proposed Beauty3DFaceNet, including 1) the effectiveness of the feature learning, fusion modules, and 3DFacePointNet++; and 2) the robustness of Beauty3DFaceNet to the quality of facial landmarks and the number of points in the input point clouds.

#### 4.3.1. Feature learning and fusing modules

**Baselines.** To accurately evaluate the three modules (e.g., 3DFacePointNet++, ResNet, and fusion modules) in our Beauty3DFaceNet, we create five baselines: 1) the original PointNet++ [10] applied on facial attractiveness prediction (FAP); 2) the 3DFacePointNet++ applied on FAP; 3) the regression network based on ResNet-18 [38]; 4) the simplified version of Beauty3DFaceNet (noted as Beauty3DFaceNet-S), which only fuses the last CNN features of ResNet-18 and 3DFacePointNet++; and 5) the Beauty3DFaceNet using the original PointNet++ (noted as Beauty3DFaceNet-O).

**Evaluation.** Table 2 shows the prediction results with different settings using corresponding experimental data, which demonstrates the effectiveness of the 3DFacePointNet++ module, ResNet module, and fusion module of the proposed Beauty3DFaceNet respectively.

*The 3DFacePointNet++ module*. As shown in Table 2, the mean absolute error (MAE) of Beauty3DFaceNet is 0.170, which is much lower than that of ResNet-18 (0.260) by 0.09. The 3DFacePointNet++ shows slight improvement on PC compared to PointNet++ (0.306 vs. 0.316), and further improves Beauty3DFaceNet compared to Beauty3DFaceNet-O (0.17 vs. 0.196). Moreover, Beauty3DFaceNet-S also leads to a better result of 0.243 than ResNet-18. This validates the advantage of our 3DFacePointNet++ module for its 3D feature learning ability.

**Table 2**
Comparison of mean absolute error (MAE) for various models trained on different experimental datasets.

| Method | Dataset | MAE |
|---|---|---|
| PointNet++ | PC | 0.316 |
| PointNet++ | PC+RGB | 0.306 |
| 3DFacePointNet++ | PC | 0.305 |
| 3DFacePointNet++ | PC+RGB | 0.295 |
| ResNet-18 | TI | 0.260 |
| Beauty3DFaceNet-S | PC+TI | 0.243 |
| Beauty3DFaceNet-O | PC+TI+TM | 0.196 |
| Beauty3DFaceNet | PC+TI+TM | **0.170** |

**Table 3**
Comparison of MAE for Beauty3DFaceNet with a different number of fusion modules.

| Fusion depth | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| MAE | 0.170 | 0.181 | 0.194 | 0.219 | 0.243 |

*The ResNet module*. We compare 3DFacePointNet++ on PC and PC+RGB, and Beauty3DFaceNet-S on PC+TI. It can be seen that Beauty3DFaceNet-S results in a much lower MAE (0.243) compared to 3DFacePointNet++ on PC+RGB (0.295). It also outperforms PointNet++ on PC by reducing MAE from 0.316 to 0.243. The results show that the ResNet module can greatly improve attractiveness prediction performance. Besides, the dense texture image is more helpful than sparse point color information.

*The fusion module*. The MAE of Beauty3DFaceNet on PC+TI+TM is much lower (0.170) than Beauty3DFaceNet-S on PC+TI (0.243). This demonstrates that the fusion module can seamlessly fuse the geometric and texture information to generate a more discriminative high-dimensional facial feature representation.

*The fusion depth*. We further test Beauty3DFaceNet that uses a different number of fusion modules (from 1 stage to 5 stages). Table 3 shows the corresponding MAE losses, which are significantly reduced with the number of fusion modules, increased.

#### 4.3.2. The robustness of Beauty3DFaceNet

*The quality of facial landmarks*. For the 3DFacePointNet++ module in the Beauty3DFaceNet, facial landmarks are employed for perceptual sensitivity sampling, which improves our results. Moreover,

**Table 4**
The MAE of the Beauty3DFaceNet with respect to different Gaussian noise added on facial landmarks.

| Gaussian noise | None | $N(2, 9)$ | $N(5, 9)$ |
|---|---|---|---|
| MAE | 0.170 | 0.172 | 0.171 |

**Table 5**
Comparison of MAE between the original PointNet++ [10], the 3DFacePointNet++, and the proposed Beauty3DFaceNet with different number of points of the input point cloud.

| Number of Input Points | 1024 | 2048 | 4096 |
|---|---|---|---|
| PointNet++ [10] | 0.316 | 0.316 | 0.317 |
| 3DFacePointNet++ | 0.310 | 0.302 | **0.295** |
| Beauty3DFaceNet | 0.196 | 0.182 | **0.170** |

**Table 6**
Comparison between our method and Fan et al. [24] with respect to MAE, Pearson correlation coefficient (PCC) and root-mean-square error (RMSE).

| Methods | Metrics | | | |
|---|---|---|---|---|
| | MAE | PCC | RMSE | DL |
| Fan et al.[24] | 0.181 | 0.832 | 0.226 | 0.08 |
| Ours | 0.170 | 0.849 | 0.223 | 0.18 |

Beauty3DFaceNet is not sensitive to the quality of facial landmarks. Table 4 shows that the MAE of the Beauty3DFaceNet is almost constant when adding Gaussian noises to the facial landmarks, which validates the robustness of our method against the quality of landmarks.

*The point number*. Here we evaluate the influence of the point number for 3D facial attractiveness prediction. Note that the original PointNet++ [10] is used for 3D object classification on ModelNet [40], where the objects (e.g., chairs, cars, etc.) are diverse in terms of the overall geometric shape. Thus a sparse point cloud would be enough to learn features and classify objects. However, 3D faces are much less discriminative, and our 3D learning module is required to learn features at a more detailed level. Thus the level of detail of the 3D face geometry is crucial for 3D facial attractiveness prediction. To validate this, we compare the 3DFacePointNet++ with the original PointNet++ with different numbers of input points on PC. We also test the performance of Beauty3DFaceNet, which fuses 3D and 2D features on PC+TI+TM. Table 5 shows that the original PointNet++ is not sensitive to the number of input points, while the MAE of our 3DFacePointNet++ decreases as the number of input points increases, and the MAE of Beauty3DFaceNet reduces from 0.196 to 0.170. This proves that our 3DFacePointNet++ captures more geometric details with an increasing number of input points.

### 4.4. Comparison with the state-of-the-art

In this subsection, we compare our work with Fan et al. [24], the state-of-the-art method for 2D facial attractiveness prediction. Beauty3DFaceNet is tested on PC+TI+TM data, while the competing method is tested on the frontal view of the 3D faces. Note that the 3D faces are rendered using a mesh representation to achieve high-quality face images. The results are shown in Table 6. It can be seen that our method performs better than Fan et al.with respect to MAE (0.170 vs. 0.181), Pearson correlation coefficient (PCC, 0.849 vs. 0.832), and root-mean-square error (RMSE, 0.223 vs. 0.226). This benefits from the much more abundant geometric information of point clouds compared with only images, and the discriminative 3D face features learned by Beauty3DFaceNet.

To further validate the discriminability of aesthetic-aware feature representation of Beauty3DFaceNet and demonstrate the necessity of our approach, we also compare the predicted attractiveness based on 100 pairs of faces $\{(F_i^o, F_i^d), (1 \le i \le 100)\}$ in the test set of the SF3D dataset, where $F_i^o$ is the original face and $F_i^d$ is the lifted face. We measure the discriminability of a model noted as *Net* using the following metric:

$$DL(Net) = \frac{1}{N}\sum_{i=0}^{i=N} |Net(\boldsymbol{F}_i^o) - Net(\boldsymbol{F}_i^d)|, \tag{4}$$

where $Net(\boldsymbol{F})$ is the predicted attractiveness for given face $\boldsymbol{F}$, and $DL(Net)$ is the discriminative value of *Net*.

The comparison results are given in the last column of Table 6, which shows that our method is much more discriminative (0.18 vs. 0.08) than the state-of-the-art 2D method [24]. Fig. 4 shows four pairs of examples of the original 3D face and the lifted face with profile proportion enhancement. Each face is rendered in multiple views in order to exhibit the profile difference better. Note that each pair is highly similar in the frontal view but quite different in profiles. As we consider view-independent geometric information instead of the frontal view only, our method is more effective when dealing with the lifted 3D faces. Taking the girl in the first row of Fig. 4 as an example, with the enhanced profiles, our prediction score increases from 3.82 to 4.02 while the score of Fan et al.'s method only increases from 3.71 to 3.80. Considering the lifted effect, our approach provides a more reasonable estimation. Fig. 5 shows several example test faces in the database along with the ground truth and predicted attractiveness using the

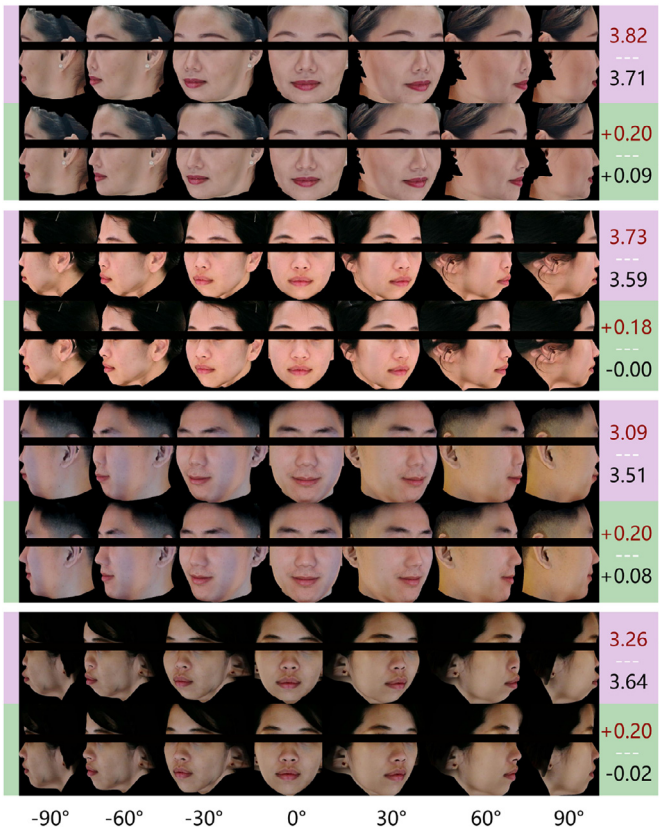



**Fig. 4.** Exemplar pairs of the original (pink) and the lifted (green) 3D faces. Each face is rendered in multiple views (-90°, -60°, -30°, 0°, 30°, 60°, 90°). In the pink bar, we show the attractiveness scores of the original faces estimated by our method (red) and Fan et al.'s method (black), respectively. In the green bar, we show the attractiveness discriminability of the lifted faces estimated by our method (red) and Fan et al.'s method (black), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

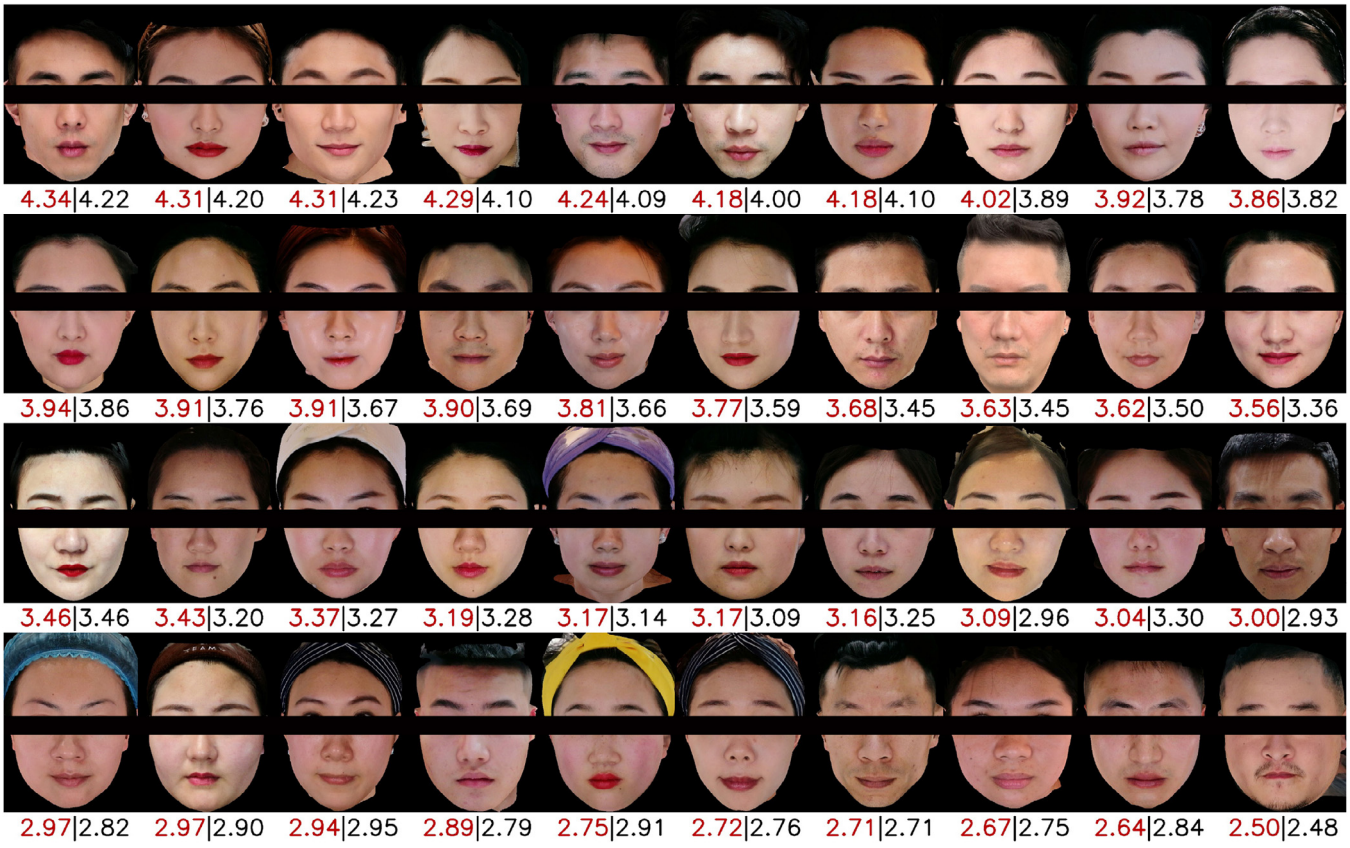



**Fig. 5.** The facial attractiveness prediction results tested on the SF3D dataset. The ground truth scores are in red, while the predicted scores by Beauty3DFaceNet are in black. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Beauty3DFaceNet. Two-tailed paired t-test reveals that there is no significant difference between the ground truth and our predicted results.

## 5. Discussion and limitation

In the previous section, by demonstrating the superior performance of the proposed Beauty3DFaceNet in ablation studies and comparisons with the state-of-the-art, we have shown its competence in fusing both the 3D point clouds and 2D texture to predict facial attractiveness. Since our method focuses on 3D facial attractiveness prediction, it can be employed to assess the attractiveness of synthesized 3D faces for gan-based 3D face design [41–46]. Specifically, our method can be potentially utilized to find the hyperplane of attractiveness in the latent space of TBGAN [47], to help enhance the synthesis quality. The newly developed facial attractiveness prediction network together with the 3D face dataset can benefit several applications where facial features and attractiveness play an essential role, such as face design and cosmetic surgery.

Our approach has some limitations. First, our dataset may not be general enough since it is collected from beauty salons and plastic surgery hospitals and does not cover all ages, such as children, as they are too young to be considered for cosmetic surgery. Second, we did not classify faces according to gender, age, region, etc., and all faces in the dataset are processed uniformly. We believe that the attractiveness prediction will be more accurate if the dataset is classified into fine-grained groups. The same as previous works on facial attractiveness assessment (such as [24]), the topic, methodology, data collection, and data publication may raise several ethical concerns even though we have achieved the academic usage agreement of ShadowFace3D dataset. We prepare a de-identification version by removing eyes from portrait images for publication to protect data privacy.

## 6. Conclusion and future work

In this paper, we present the first deep convolutional neural network, called Beauty3DFaceNet, for 3D facial attractiveness prediction. Through carefully designed feature learning and fusing modules, the proposed Beauty3DFaceNet can reliably learn and complement 3D and 2D features from face geometry and texture, resulting in more accurate facial attractiveness prediction. Moreover, we propose a novel facial prior-based sampling strategy to preserve important face features for attractiveness prediction while reducing the prediction cost in terms of the number of input points. We also present a new 3D face dataset, called ShadowFace3D. It contains 6,000 faces collected from beauty salons and cosmetic surgery hospitals and has attractiveness annotations reflecting public aesthetic criterion. We validate our network's effectiveness through extensive quantitative and qualitative evaluations, including carefully designed ablation studies and comparisons with the state-of-the-art.

For future work, we would like to investigate the explainability of our deep-learning-based network to interpret the facial attractiveness according to learned geometric and textural features. More general and cross-cultural aesthetics can be further explored by collecting more data of different regions. We will also utilize Beauty3DFaceNet and ShadowFace3D dataset for various facial attractiveness related applications, such as 3D face attractiveness enhancement, 3D face plastic surgery, etc.

## References


[1] Leyvand T, Cohen-Or D, Dror G, Lischinski D. Data-driven enhancement of facial attractiveness. In: ACM SIGGRAPH 2008 papers. New York, NY, USA: Association for Computing Machinery; 2008. ISBN 9781450301121 .

[2] Perrett DI, May KA, Yoshikawa S. Facial shape and judgements of female attractiveness. Nature 1994;368(6468):239–42. https://doi.org/10.1038/368239a0

[3] Bottino A, Simone M, Laurentini A, Sforza C. A new 3D tool for planning plastic surgery. IEEE Trans Biomed Eng 2012;59:3439–49. doi:10.1109/TBME.2012.2217496 .

[4] Diamant N., Zadok D., Baskin C., Schwartz E., Bronstein A.M. Beholder-GAN: generation and beautification of facial images with conditioning on their beauty level. arXiv preprint2019;:arXiv:1902.02593.

[5] Liao Q, Jin X, Zeng W. Enhancing the symmetry and proportion of 3D face geometry. IEEE Trans Vis ComputGraph 2012;18(10):1704–16.

[6] Mulhern R, Fieldman G, Hussey T, Lévêque J-L, Pineau P. Do cosmetics enhance female caucasian facial attractiveness? Int J Cosmetic Sci 2003;25(4):199–205.

[7] Gao L, Li W, Huang Z, Huang D. Automatic facial attractiveness prediction by deep multi-task learning. In: International Conference on Pattern Recognition (ICPR); 2018. p. 3592–7. doi: 10.1109/ICPR.2018.8545033 .

[8] Zhai Y, Cao H, Deng W, Gan J, Piuri V, Zeng J. BeautyNet: joint multiscale CNN and transfer learning method for unconstrained facial beauty prediction. Comput Intell Neurosci 2019 .

[9] Liu S, Fan Y, Guo Z, Samal A, Ali A. A landmark-based data-driven approach on 2.5D facial attractiveness computation. Neurocomputing 2017a;238:168–78. doi: 10.1016/j.neucom.2017.01.050 .

[10] Qi CR, Yi L, Su H, Guibas LJ. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems(NIPS); 2017a. p. 5099–108 .

[11] Itier RJ, Van Roon P, Alain C. Species sensitivity of early face and eye processing. NeuroImage 2011;54(1):705–13.

[12] Gunes H, Piccardi M. Assessing facial beauty through proportion analysis by image processing and supervised learning. Int J Hum Comput Stud 2006;64(12):1184–99. doi: 10.1016/j.ijhcs.2006.07.004 .

[13] Eisenthal Y, Dror G, Ruppin E. Facial attractiveness: beauty and the machine. Neural Comput 2006;18(1):119–42.

[14] Kagian A, Dror G, Leyvand T, Cohen-Or D, Ruppin E. A humanlike predictor of facial attractiveness. In: Advances in Neural Information Processing Systems(NIPS); 2007. p. 649–56. ISBN 9780262195683 .

[15] Schmid K, Marx D, Samal A. Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. Pattern Recognit 2008;41(8):2710–17. doi: 10.1016/j.patcog.2007.11.022 .

[16] Fan J, Chau KP, Wan X, Zhai L, Lau E. Prediction of facial attractiveness from facial proportions. Pattern Recognit 2012;45(6):2326–34. doi:10.1016/j.patcog.2011.11.024 .

[17] Whitehill J, Movellan JR. Personalized facial attractiveness prediction. In: IEEE international conference on automatic face & gesture recognition; 2008. p. 1–7.

[18] Chen F, Xiao X, Zhang D. Data-driven facial beauty analysis: prediction, retrieval and manipulation. IEEE Trans Affect Comput 2018;9(2):205–16. doi:10.1109/TAFFC.2016.2599534 .

[19] Bottino A, Laurentini A. The intrinsic dimensionality of attractiveness: astudy in face profiles. Lect Notes Comput Sci 2012;7441:59–66. doi:10.1007/978- 3- 642- 33275- 3 _ 7 .

[20] Davis B, Lazebnik S. Analysis of human attractiveness using manifold kernel regression. In: International Conference on Image Processing (ICIP); 2008. p. 109–12. doi:10.1109/ICIP.2008.4711703.

[21] Gray D, Yu K, Xu W, Gong Y. Predicting facial beauty without landmarks. Lect Notes Comput Sci 2010;6316:434–47. doi:10.1007/978-3-642-15567-3_32.

[22] Gan J, Li L, Zhai Y, Liu Y. Deep self-taught learning for facial beauty prediction. Neurocomputing 2014;144:295–303. doi:10.1016/j.neucom.2014.05.028.

[23] Xu J, Jin L, Liang L, Feng Z, Xie D, Mao H. Facial attractiveness prediction using psychologically inspired convolutional neural network (PI-CNN). In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017. p. 1657–61 .

[24] Fan Y, Liu S, Li B, Guo Z, Samal A, Wan J, et al. Label distribution-based facial attractiveness computation by deep residual learning. IEEE Trans Multimed 2018;20(8):2196–208. doi: 10.1109/TMM.2017.2780763 .

[25] Zhao H, Jin X, Huang X, Chai M, Zhou K. Parametric reshaping of portrait images for weight-change. IEEE Comput Graph Appl 2018;38(1):77–90. doi:10.1109/MCG.2018.011461529 .

[26] Xiao Q, Tang X, Wu Y, Jin L, Yang Y-L, Jin X. Deep shapely portraits. In: Proceedings of the 28th ACM international conference on multimedia. New York, NY, USA: Association for Computing Machinery; 2020. p. 1800–8. ISBN 9781450379885 . https://doi.org/10.1145/3394171.3413873

[27] O'Toole AJ, Price T, Vetter T, Bartlett JC, Blanz V. 3D shape and 2D surface textures of human faces: the role of "averages" in attractiveness and age. Image Vis Comput 1999;18(1):9–19 .

[28] Kim J-S, Choi S-M. Symmetric deformation of 3D face scans using facial features and curvatures. Comput Anim Virtual Worlds 2009;20:289–300.

[29] Xu J, Tian W, Fan Y, Lin Y, Zhang C. Personality trait prediction based on 2.5D face feature model. In: Cloud computing and security; 2018a. p. 611–23. ISBN 978-3-030-0 0 021-9

[30] Liu S, Fan Y, Guo Z, Samal A. 2.5D facial attractiveness computation based on data-driven geometric ratios. In: Intelligence science and big data engineering. image and video data engineering; 2015. p. 564–73. ISBN 978-3-319-23989-7 .

[31] Qi CR, Su H, Mo K, Guibas LJ. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR); 2017b. p. 652–60 .

[32] Li Y, Bu R, Sun M, Wu W, Di X, Chen B. Pointcnn: Convolution on -transformed points. In: Proceedings of the 32nd international conference on neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.; 2018. p. 828–38.

[33] Wu W, Qi Z, Li F. Pointconv: deep convolutional networks on 3D point clouds. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR); 2019. p. 9621–30 .

[34] Xu D, Anguelov D, Jain A. PointFusion: deep sensor fusion for 3D bounding box estimation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR); 2018b. p. 244–53 .

[35] Wu W, Qian C, Yang S, Wang Q, Cai Y, Zhou Q. Look at boundary: a boundary-aware face alignment algorithm. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR); 2018. p. 2129–38 .

[36] Liu Y, Jourabloo A, Ren W, Liu X. Dense face alignment. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(ICCV); 2017b. p. 1619–28.

[37] Bellus3D. Fast & easy lifelike 3D face scanning. https://www.bellus3d.com; 2021.

[38] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR); 2016. p. 770–8 .

[39] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR). Ieee; 2009. p. 248–55 .

[40] Vishwanath KV, Gupta D, Vahdat A, Yocum K. Modelnet: towards a datacenter emulation environment. In: IEEE ninth international conference on peer–to-peer computing; 2009. p. 81–2 .

[41] Deng J, Cheng S, Xue N, Zhou Y, Zafeiriou S. Uv-gan: adversarial facial uv map completion for pose-invariant face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018. p. 7093–102.

[42] Zhou Y, Deng J, Kotsia I, Zafeiriou S. Dense 3d face decoding over 2500fps: joint texture & shape convolutional mesh decoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 1097–106 .

[43] Gecer B, Ploumpis S, Kotsia I, Zafeiriou S. Ganfit: generative adversarial network fitting for high fidelity 3d face reconstruction. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 1155–64 .

[44] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 4401–10 .

[45] Lin J., Zhang R., Ganz F., Han S., Zhu J.-Y. Anycost GANS for interactive image synthesis and editing. 2021. 2103.03243.

[46] Chen S.-Y., Su W., Gao L., Xia S., Fu H. Deep generation of face images from sketches. 2020. 2006.01047.

[47] Gecer B, Lattas A, Ploumpis S, Deng J, Papaioannou A, Moschoglou S, et al. Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. In: Vedaldi A, Bischof H, Brox T, Frahm J-M, editors. Computer vision – ECCV 2020. Cham: Springer International Publishing; 2020. p. 415–33 .