TYPE Original Research
PUBLISHED 15 January 2026
DOI 10.3389/fcomp.2025.1692523

Check for updates

OPEN ACCESS

EDITED BY
Elisa Ricci,
University of Trento, Italy

REVIEWED BY
Jiahao Bao,
Shanghai Jiao Tong University, China
Djamel Eddine Boukhari,
Scientific and Technical Research Center on Arid Regions (CRSTRA), Algeria

*CORRESPONDENCE
Dong Huang
✉ huangdong1007785@outlook.com

RECEIVED 26 August 2025
REVISED 27 October 2025
ACCEPTED 08 December 2025
PUBLISHED 15 January 2026

CITATION
Wang K, Li Y, Huang D, Feng J and Feng X (2026) DeepGeoFusion: personalized facial beauty prediction through geometric-visual fusion. *Front. Comput. Sci.* 7:1692523. doi: 10.3389/fcomp.2025.1692523

DeepGeoFusion: personalized facial beauty prediction through geometric-visual fusion

Kunwei Wang[1], Yanzhi Li[2], Dong Huang[3]*, Junmei Feng[4] and Xiaoyi Feng[1]

[1]School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China, [2]Shaanxi Chang'an Computing Technology Co., Ltd., Xi'an, Shaanxi, China, [3]Department of Biomedical Engineering, Air Force Medical University, Xi'an, China, [4]Guangzhou Institute of Technology, Xidian University, Guangzhou, China

Introduction: Personalized facial beauty prediction is a critical advancement beyond population-level models with transformative applications in aesthetic surgery planning and user-centric recommendation systems, while contemporary methods face limitations in modeling aesthetically sensitive facial regions, fusing heterogeneous geometric and visual features, and reducing extensive annotation dependency for personalization.
Methods: We propose DeepGeoFusion, a novel framework that synergizes Vision Mamba-extracted global visual features with anatomically constrained facial graphs (constructed from 86 landmarks via Delaunay triangulation), using the Graph Node Attention Projection Fusion (GNAPF) block for cross-modal alignment and a lightweight adaptation mechanism to generate personalized preference vectors from 10 seed images via confidence-gated optimization.
Results: Extensive experiments on SCUT-FBP5500 demonstrate statistically significant improvements in personalized prediction accuracy and robust performance across genders and ethnicities compared to state-of-the-art methods.
Discussion: DeepGeoFusion effectively addresses key limitations of existing methods by integrating complementary geometric and visual features, enabling efficient personalization with minimal annotation and highlighting practical value for aesthetic-related applications requiring personalized assessments.

KEYWORDS

face beauty prediction, personalized beauty prediction, geometric feature, feature fusion, graph attention

1 Introduction

Automated facial beauty prediction (FBP) lies at the intersection of computer vision and computational aesthetics, emerging as a critical component in next-generation intelligent systems. Its applications span a wide range of domains, including personalized content curation in social recommendation engines, objective surgical outcome evaluation in medical cosmetology, and emotion recognition in human-computer interaction. This growing significance is reflected in market projections, with Grand View Research estimating that the global beauty technology sector will grow from $66.17 billion in 2024 to $172.99 billion by 2030, driven by a 17.9% compound annual growth rate (CAGR). FBP is a foundational technology that enables the development of personalized solutions across these industries.

The methodological evolution of FBP can be divided into three phases, each addressing key limitations of prior approaches. Early studies primarily relied on geometric feature engineering, using anthropometric measurements like the golden ratio and facial symmetry indices to build evaluation models (Eisenthal et al., 2006; Zhang et al., 2011; Gunes and Piccardi, 2006; Zhang et al., 2017; Peng et al., 2023). With the advent of deep learning, the field shifted to data-driven approaches, leveraging convolutional neural networks (CNNs) to automatically extract aesthetic features directly from raw pixel data (Gan et al., 2014; Xie et al., 2015; Liang et al., 2018; Gray et al., 2010; Bougourzi et al., 2022; Gan et al., 2024; Moridani et al., 2023; Dornaika et al., 2020; Zhai et al., 2020; Li et al., 2018; Chen et al., 2018). More recently, hybrid approaches have emerged, combining geometric constraints with deep learning architectures to improve the physiological relevance and accuracy of predictions (Wang et al., 2022; Xiao et al., 2021; Peng et al., 2024; Sun et al., 2024a,b; Gan et al., 2023).

While these advancements have enabled robust population-level predictions, achieving correlations greater than 0.9 on aggregated datasets, they overlook a crucial aspect: individual differences in aesthetic perception. Psychological studies have shown that personal preferences play a significant role in beauty ratings, often surpassing other factors such as cultural influences and image quality. This highlights a significant gap in current models and points to the need for a paradigm shift–toward personalized beauty prediction frameworks (Whitehill and Movellan, 2008; Lin et al., 2023; Lebedeva et al., 2021, 2023) that account for the subjective nature of beauty and can adapt to individual preferences across various contexts.

Contemporary personalized FBP research aims to model subjective user preferences. Lebedeva et al. (2023) propose a meta-learning framework that personalizes facial beauty assessment based on user data, focusing on semantic features extracted by deep convolutional networks (CNNs). Lin et al. (2023) present a high-order predictor for the same task, aiming at personalization through latent parameter generation. However, both methods face critical limitations. They rely solely on deep convolutional features (Huang et al., 2022) and neglect important facial geometry metrics, such as golden ratios and symmetry indices, which are known to influence aesthetic perceptions (Li et al., 2024). Additionally, while both approaches attempt personalization, they fail to effectively integrate user-specific aesthetic preferences, unlike the hierarchical feature fusion strategies successfully employed in related facial analysis tasks (Huang et al., 2022; Li et al., 2024; Zhang et al., 2016). Consequently, these models require extensive annotations and lack robustness across demographic groups. These drawbacks highlight the need for a more effective framework that combines both geometric and deep visual features for improved personalization.

To address the critical limitations of existing personalized FBP models, we propose a novel integrated architecture–DeepGeoFusion. Our proposed framework overcomes these challenges by seamlessly combining global visual features with geometrically constrained facial topological modeling, ensuring more accurate and personalized predictions of facial beauty. The integration of these two types of features allows DeepGeoFusion to offer an innovative and highly efficient solution to predicting facial attractiveness based on both visual semantics and geometric structure.

Our model is built on three key innovations:

(1) We enhance facial feature representation by combining global visual features extracted via Vision Mamba with a geometric topology constructed from 86 facial landmarks, structured through Delaunay triangulation. These heterogeneous features are integrated using a Graph Node Attention Projection Fusion (GNAPF) module, capturing both fine-grained visual textures and facial structure in a unified representation.

(2) We introduce an efficient personalization framework that learns user-specific aesthetic preferences from a small set of seed images. Through this process, our model generates personalized preference vectors that are dynamically adjusted using a confidence gating mechanism. This enables the model to tailor its predictions to the subjective beauty standards of individual users, enhancing the accuracy of beauty scoring based on personal tastes.

(3) Extensive evaluations across facial datasets with diverse genders and ethnicities show that our model not only achieves high accuracy in general beauty prediction but also retains robustness in personalized scoring, demonstrating strong generalizability and adaptability acros populations.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of the related literature; Section 3 details the proposed personalized evaluation framework; Section 4 presents experimental validation; and the final section provides a summary of the key contributions of this research and outlines potential directions for future work.

2 Related work

Based on the perspective of the rater (the aesthetic subject), research in facial beauty prediction (FBP) falls into two main streams (Ibrahem and Abdulazeez, 2025; Ibrahem et al., 2025): general prediction, which aims to learn population-level aesthetic consensus, and personalized prediction, which models individual aesthetic variations. This section provides a systematic review of the technical evolution and representative work within both categories.

2.1 General facial beauty prediction

The field of Facial Beauty Prediction (FBP) has undergone a significant paradigm shift, evolving from shallow models to deep learning approaches. Tracing its evolutionary trajectory, methodologies for general FBP can be broadly categorized into three primary approaches:

2.1.1 Geometric-driven methods

Early facial aesthetic assessment primarily relied on geometric features, such as interocular distance and facial proportions, as fundamental indicators of facial attractiveness (Juravle and Spence, 2024; Londono et al., 2024). Early research focused on handcrafted geometric features for model training. For instance, Gunes and Piccardi (2006) identified 16 key ratios across facial contours, eyes, brows, nose, lips, and chin, using them as input features for a C4.5 decision tree, achieving competitive results on a small dataset. Similarly, Zhang et al. (2011) developed a geometric framework that extracted 58 features from 83 facial landmarks, demonstrating that proximity to average facial geometry correlates with perceived beauty. Although geometric methods are intuitive and interpretable, they suffer from limitations such as labor-intensive feature engineering, shallow feature expressiveness, and high reliance on precise landmark localization and manual design.

2.1.2 Deep learning-based methods

Early advancements in deep learning for FBP were pioneered by Gray et al. (2010), who introduced a multilayer feedforward network, akin to modern CNNs, to assess female facial attractiveness. Their model achieved a high correlation using a dataset of 2,056 facial images. Bougourzi et al. (2022) further advanced this field by designing a dual-branch architecture that integrates ResNeXt-50 and Inception-v3 modules, outperforming traditional regression baselines. Deep learning models, by leveraging hierarchical representations from raw pixel data, capture complex patterns in texture, contours, and subtle interactions between facial attributes. While these methods provide state-of-the-art accuracy and scalability, their "black-box" nature limits interpretability compared to geometric or hybrid models.

2.1.3 Hybrid fusion methods

Xiao et al. (2021) proposed Beauty3DFaceNet, a deep convolutional neural network that integrates both 3D geometric features (from point clouds) and 2D texture features for facial beauty prediction. Their framework introduced a fusion module that enhanced feature learning by combining geometry and texture and included a landmark-guided sampling method to optimize aesthetic feature extraction. This approach was further extended with the introduction of the ShadowFace3D dataset. Similarly, Peng et al. (2024) developed a geometric prior guided hybrid deep neural network that systematically combines deep convolutional features with facial geometric priors through a hierarchical fusion strategy. Their approach introduced a novel geometric attention mechanism to dynamically adjust the contribution of different facial regions based on learned aesthetic patterns, achieving superior performance while maintaining model interpretability through explicit geometric constraints. These hybrid models offer promising results, yet they are still challenged by high computational costs and limited model flexibility.In contrast, Yan and Ye (2025) introduced the Adaptive Cross-Cultural Beauty Fusion Network (ACBF-Net), a hybrid deep learning framework that adapts to cultural differences by combining global visual features with culture-specific priors. This results in more accurate and fair predictions across diverse demographics, as demonstrated in experiments using cross-cultural datasets.

Geometric-driven methods offer the advantage of providing stability and interpretability by relying on easily measurable facial features. However, they are limited by the need for precise landmark localization and shallow feature representation. Deep learning-based methods excel in capturing complex, high-dimensional features from raw pixel data, leading to improved accuracy and performance. Nevertheless, their "black-box" nature makes them difficult to interpret. Hybrid fusion methods, which combine the strengths of both approaches, present a promising direction by enhancing predictive performance through the integration of both geometric and deep features. Yet, these methods still face challenges related to computational complexity and model flexibility. The current research is focusing on finding efficient ways to fuse deep features and geometric features to overcome these limitations and enhance facial beauty prediction.

2.2 Personalized facial beauty prediction

In contrast to generic prediction modeling population-level consensus, personalized facial beauty prediction (PFBP) addresses the inherent subjectivity of aesthetic preferences across individuals–crucially enabling applications like cosmetic recommendation and user-centric attractiveness enhancement.

Whitehill and Movellan (2008) laid the groundwork by introducing a support vector regression (SVR) model for personalized attractiveness assessment, although their approach was constrained by shallow features and a limited dataset. Lebedeva et al. (2021) enhanced prediction accuracy by employing deep convolutional neural networks (CNNs) to extract more abstract, high-dimensional features, incorporating individual preferences to better capture subjective attractiveness ratings. In 2023, Lebedeva et al. further pushed the boundaries by introducing a meta-learning framework, enabling models to quickly adapt to individual preferences with minimal data, showcasing superior generalization and learning efficiency compared to traditional deep learning methods (Lebedeva et al., 2023). The same year, Lin et al. (2023) proposed MetaFBP, an advanced model integrating meta-learning mechanisms that adapted more effectively to diverse aesthetic judgments by incorporating a learnable user adaptation module. These studies collectively highlight the progression from traditional machine learning approaches to more sophisticated deep learning and meta-learning techniques, although challenges remain in terms of computational efficiency, feature selection, and handling low- or zero-shot learning scenarios. Future research will need to explore combining both deep and geometrical features for a more holistic and practical approach to personalized facial attractiveness prediction.

While personalized facial attractiveness prediction has seen substantial advancements, including the incorporation of deep learning and meta-learning techniques, several limitations remain. These methods still face challenges in computational efficiency, as the need for individual user training or fine-tuning makes large-scale deployment costly. Additionally, many approaches overly rely on deep visual features, neglecting the significance of geometrical features such as facial symmetry and proportionality, which are strongly correlated with perceived attractiveness.

3 Method

We propose a deep learning framework for personalized facial beauty prediction, illustrated in Figure 1. The architecture comprises four core components: First, the Feature Extraction Module employs a dual-path architecture: the Vision Mamba

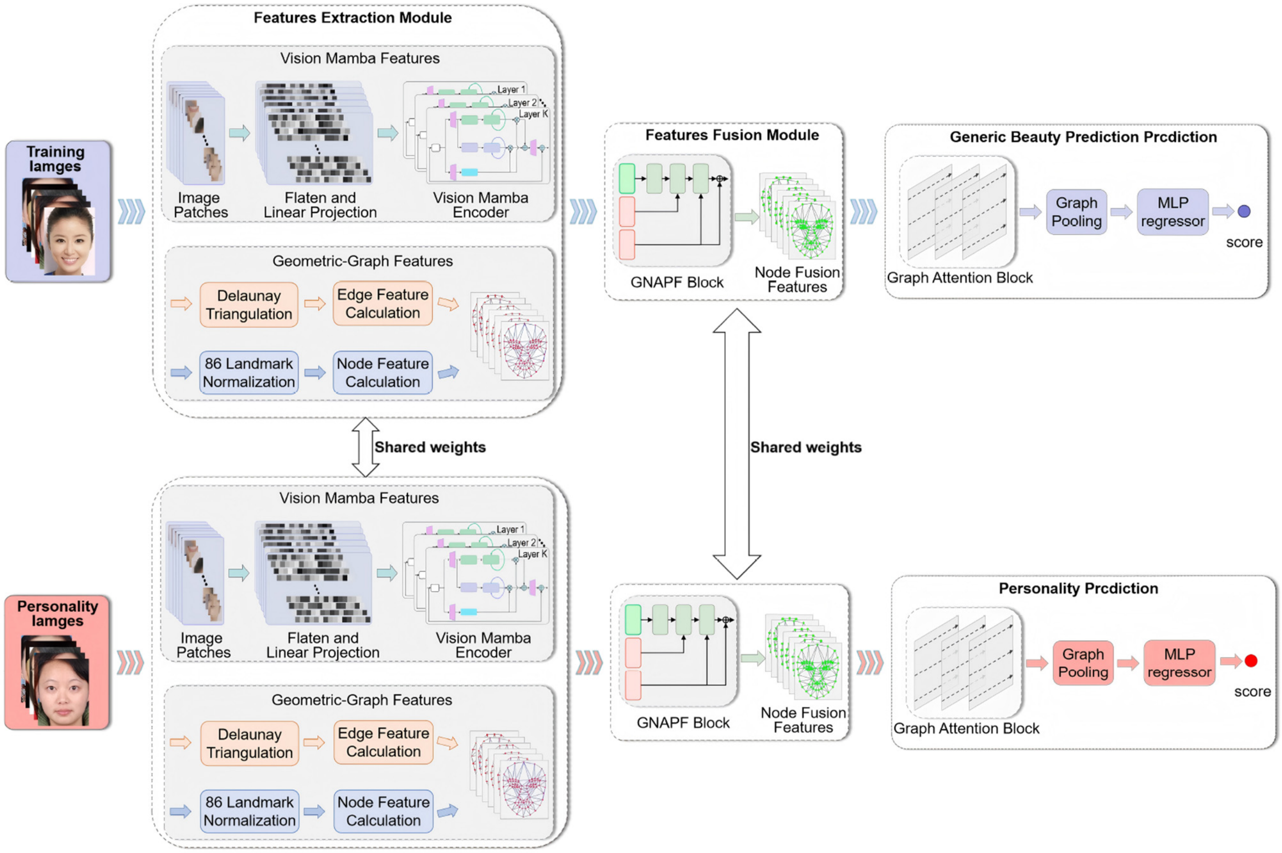

FIGURE 1
Architecture of the proposed personalized facial beauty prediction framework. Image(s) taken from Liang et al. (2018).

Block captures hierarchical visual representations from facial images via state-space modeling, while the Geometric-Graph Block constructs topology-aware graphs from 86 anthropometric landmarks to encode spatial relationships. Second, the Graph Node Attention Projection Fusion (GNAPF) module bridges the cross-modal semantic gap by aligning and integrating geometric and visual features through attention-based projection mechanisms. Third, the fused representations are decoded using a Graph Attention Network (GAT), which leverages attention-based graph operations to predict continuous beauty scores. Finally, to enable personalized aesthetic assessment, a dedicated Personalization Adaptation Module incorporates user-provided ratings of seed images to extract subjective preference priors, which are then used to fine-tune the core model parameters, allowing the system to generate individual-specific beauty predictions.

3.1 Feature extraction module

In the feature extraction module, we extract two complementary feature representations from the raw facial images for personalized facial beauty prediction: global semantic features via the Vision Mamba architecture, and geometric graph features constructed based on facial landmarks.

3.1.1 Vision Mamba features extraction block

We utilize Vision Mamba as our global feature extractor, leveraging its Cross-Scan State Space Modeling (CSSM) framework to capture comprehensive facial attributes. The processing pipeline begins by converting input images $\mathbf{I} \in \mathbb{R}^{224^2\times 3}$ into D-dimensional patch embeddings ($D = 192$) with positional encodings. These embeddings are then processed through L stacked Mamba blocks, each featuring dual-parallel processing pathways.

Within each block, input features $\mathbf{X}$ undergo simultaneous transformation through complementary mechanisms: (1) A selective state-space path with dynamic parameterization $(\Delta, \mathbf{B}, \mathbf{C} = \tau(\mathbf{X}))$ and discretized state updates $(\overline{\mathbf{A}}, \overline{\mathbf{B}}$ via ZOH). (2) A linear attention path using kernel-based approximation for efficient local feature extraction.

The core computations integrate these pathways through residual connections and nonlinear enhancement:

$$\mathbf{Y}_{\text{ssm}} = \text{SSM}(\mathbf{X}) \tag{1}$$

$$\mathbf{Y}_{\text{att}} = \frac{\phi(\mathbf{Q})(\phi(\mathbf{K})^\top \mathbf{V})}{\sqrt{d}} \qquad (\text{with } \phi(x) = \text{ELU}(x) + 1) \tag{2}$$

$$\mathbf{Z} = \text{LayerNorm}(\mathbf{Y}_{\text{ssm}} + \mathbf{Y}_{\text{att}}) \tag{3}$$

$$\mathbf{X}_{\text{out}} = \text{MLP}(\mathbf{Z}) + \mathbf{Z} \tag{4}$$

Bidirectional context modeling processes sequences in forward/reverse orientations, with spatial aggregation yielding the final 384-dimensional facial descriptor:

$$\mathbf{f}_{\text{mamba}} = \text{Pool}\big([\mathbf{X}_{\text{forward}}; \mathbf{X}_{\text{backward}}]\big) \tag{5}$$

This CSSM-based extraction holistically encodes facial topology, expressions, and illumination characteristics through selective global-local feature integration.

3.1.2 Geometric-graph features extraction block

To address the limitations of global features and handcrafted descriptors in modeling facial aesthetics, we propose a Delaunay triangulation-based geometric graph feature. This approach synergistically integrates anatomical constraints with semantic features through topologically consistent mes structures, as illustrated in Figure 2.

3.1.2.1 Landmark normalization

Input image $\mathbf{I} \in \mathbb{R}^{224\times224\times3}$ with 86 landmarks $\mathcal{P}$ undergoes dual normalization:

$$\text{Global}: \quad \mathbf{p}'_i = s \cdot \mathbf{R}(\theta) \cdot (\mathbf{p}_i - \mathbf{c}) \tag{6}$$

$$\text{Local}: \quad \mathbf{p}_k^{\text{local}} = \mathbf{R}_{\text{region}} \cdot (\mathbf{p}_k - \mathbf{c}_{\text{region}}) \tag{7}$$

where $\mathbf{c}$ denotes the facial centroid, $s = \| \ \mathbf{p}_{36} - \mathbf{p}_{45} \ \|^{-1}$ scales by inter-pupil distance, and θ aligns with nasal bridge orientation $\overrightarrow{\mathbf{p}_{27}\mathbf{p}_{33}}$. This ensures pose-invariant representation.

3.1.2.2 Anatomically constrained Delaunay triangulation

The normalized points $\mathcal{P}'$ are triangulated with biological boundary constraints preserving facial topology:

$$\begin{aligned} \mathcal{B}_{\text{face}} &= \{(0,1), \ldots, (16,0)\}, \\ \mathcal{B}_{\text{eye}} &= \{(36,37), \ldots, (41,36)\}, \\ \mathcal{B}_{\text{lip}} &= \{(48,49), \ldots, (59,48)\} \end{aligned} \tag{8}$$

Each vertex v_i encodes biological attributes in a 10-dimensional feature vector:

$$\mathbf{f}_v^{(i)} = \left[x'_i,\ y'_i,\ \kappa_i,\ \rho_i,\ o,\ w_i\right]^T \tag{9}$$

where κ_i denotes discrete curvature, ρ_i represents local point density, and o is the anatomical region encoding.

3.1.2.3 Deformation-robust edge features

Edge features incorporate biomechanical properties:

$$\mathbf{f}_e^{(ij)} = \left[\frac{l_{ij}}{l_{\text{median}}},\ \theta_{ij},\ \alpha_{ij},\ \beta_{ij},\ \gamma_{ij},\ t_{ij}\right]^T \tag{10}$$

Critical is the *rigid ratio* $\gamma_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\| / \|\mathbf{p}_i^{\text{neu}} - \mathbf{p}_j^{\text{neu}}\|$, preserving invariance under expression variations.

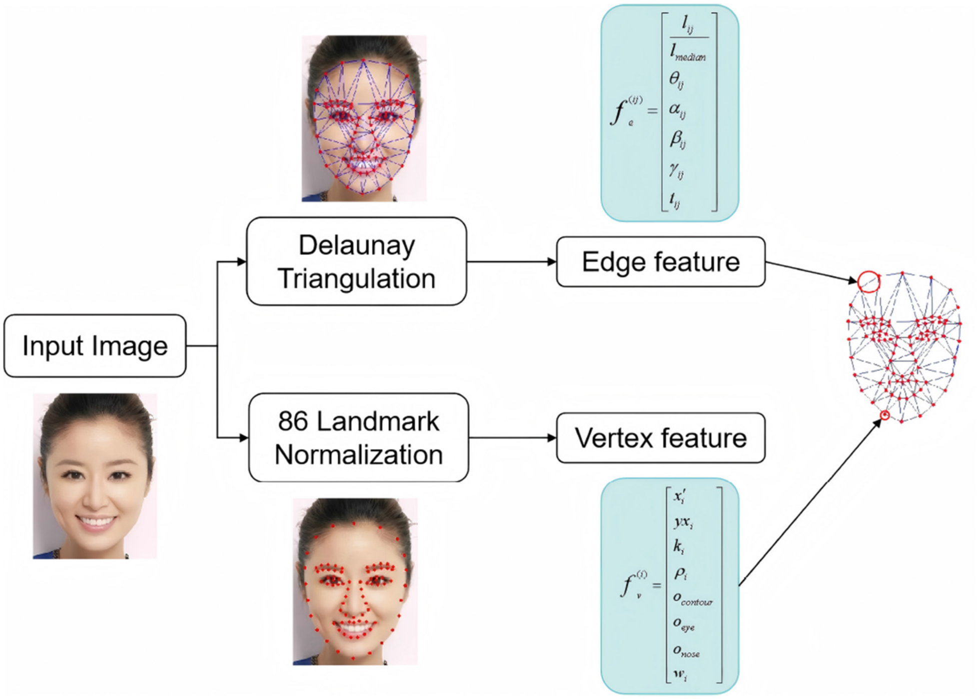
$f_e^{(ij)} = [\frac{l_{ij}}{l_{median}}, \theta_{ij}, \alpha_{ij}, \beta_{ij}, \gamma_{ij}, t_{ij}]$
Delaunay Triangulation
Edge feature
Input Image
86 Landmark Normalization
Vertex feature
$f_v^{(i)} = [x_i', yx_i, k_i, \rho_i, o_{contour}, o_{eye}, o_{nose}, w_i]$

FIGURE 2
Facial geometric graph feature extraction pipeline. For an input face image, two parallel branches process edge and vertex geometry: Delaunay triangulation generates edge features from inter-landmark connections, while 86 normalized landmarks yield vertex features encoding local structure. Separate feature computation is followed by fusion into a geometric graph, which captures inherent facial structural patterns for downstream analysis. Image(s) taken from Liang et al. (2018).

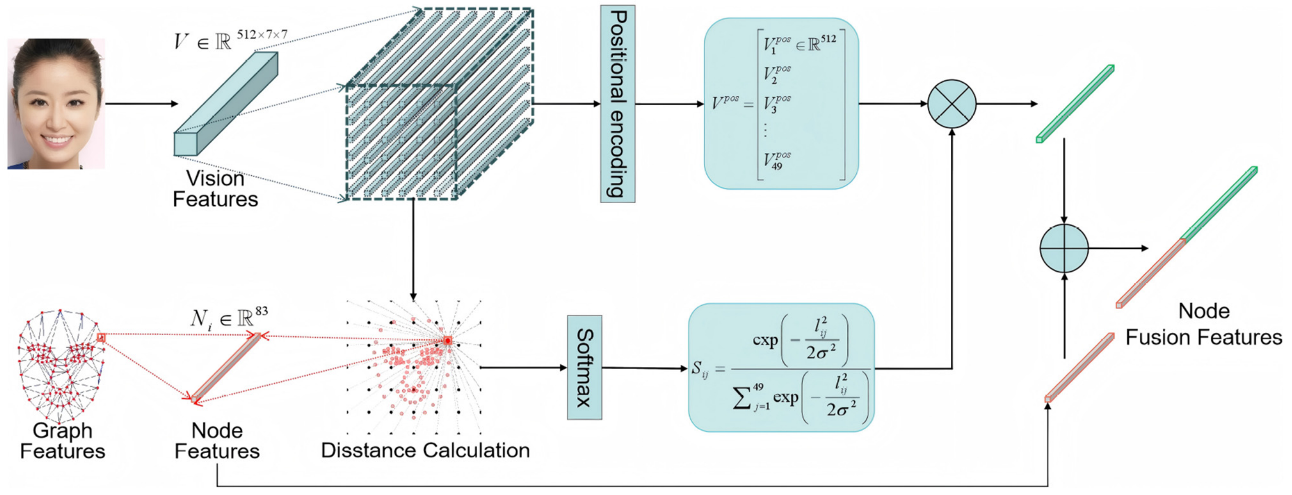
$V \in \mathbb{R}^{512\times7\times7}$
Vision Features
Positional encoding
$V^{pos} = [V_1^{pos} \in \mathbb{R}^{512}, V_2^{pos}, V_3^{pos}, \vdots, V_{49}^{pos}]$
Node Fusion Features
$N_i \in \mathbb{R}^{83}$
Graph Features
Node Features
Disstance Calculation
Softmax
$S_{ij} = \frac{\exp\left(-\frac{l_{ij}^2}{2\sigma^2}\right)}{\sum_{j=1}^{49}\exp\left(-\frac{l_{ij}^2}{2\sigma^2}\right)}$

FIGURE 3
Architecture of Graph Node Attention Projection Fusion (GNAPF) block. Image(s) taken from Liang et al. (2018).

3.1.2.4 Biomechanical energy constraints

An energy minimization framework maintains structural integrity:

$$\mathcal{L}_{\text{rigid}} = \sum_{\Delta_{ijk}} \left| \frac{\|\mathbf{p}_i - \mathbf{p}_j\|}{\|\mathbf{p}_j - \mathbf{p}_k\|} - r_{ijk}^0 \right|^2 \tag{11}$$

$$\mathcal{L}_{\text{elastic}} = \sum_{e_{ij}} \lambda_{ij}(l_{ij} - l_{ij}^0)^2 \tag{12}$$

Tissue-specific elasticity ($\lambda_{\text{bone}} = 1.0$, $\lambda_{\text{skin}} = 0.2$) prevents unnatural distortions.

3.2 Feature fusion module

The Graph Node Attention Projection Fusion (GNAPF) block integrates vision mamba features with facial graph representations through spatially-aware attention mechanisms. As illustrated in Figure 3 and Algorithm 1, this process operates in two coordinated stages centered on graph nodes.

3.2.1 Visual feature projection to graph nodes

Visual features $\mathbf{V} \in \mathbb{R}^{512\times7\times7}$ extracted from Vision Mamba undergo spatially-conditioned projection onto facial graph nodes $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\|\mathcal{V}\| = 86$ keypoints. The projection employs distance-based attention to preserve geometric relationships:

$$e_{ij} = -\lambda \cdot \|\mathbf{p}_i - \mathbf{g}_j\|_2 \tag{13}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{49}\exp(e_{ik})} \tag{14}$$

$$\mathbf{V}_j^{\text{pos}} = \mathbf{V}_j + \text{PE}(\mathbf{g}_j) \tag{15}$$

$$\tilde{\mathbf{v}}_i = \sum_{j=1}^{49} \alpha_{ij}\mathbf{V}_j^{\text{pos}} \tag{16}$$

```
Input: Visual feature map: $\mathbf{V} \in \mathbb{R}^{512\times7\times7}$
Keypoint positions: $\mathbf{P} = \{\mathbf{p}_1, \cdots, \mathbf{p}_{86}\} \in \mathbb{R}^{86\times2}$
Graph node features: $\mathbf{X} \in \mathbb{R}^{86\times d_g}$
Grid positions: $\mathbf{G} \in \mathbb{R}^{7\times7\times2}$
Output: Fused features $\mathbf{F} \in \mathbb{R}^{86\times(d_g+d_v)}$
// Spatial projection of visual features
$\tilde{\mathbf{V}} \leftarrow \mathbf{0}^{86\times d_v}$          // Initialize projected features
for $i \leftarrow 1$ to 86 do
    for $j \leftarrow 1$ to 49 do
        $\alpha_{ij} \leftarrow \exp(-10 \cdot \|\mathbf{p}_i - \mathbf{g}_j\|_2)$
    end
    $\tilde{\mathbf{v}}_i \leftarrow \sum_j \alpha_{ij} \cdot (\phi(\mathbf{V}_j) + \text{PE}(\mathbf{g}_j))$
    $\tilde{\mathbf{V}}[i] \leftarrow \tilde{\mathbf{v}}_i$
end
// Node feature augmentation
for $i \leftarrow 1$ to 86 do
    $\mathbf{F}_i \leftarrow [\mathbf{X}_i; \tilde{\mathbf{V}}_i]$          // Feature concatenation
end
return $\mathbf{F}$
```

Algorithm 1. Visual-geometric feature fusion.

where $\mathbf{p}_i \in \mathbb{R}^2$ denotes the coordinate of the i-th facial keypoint, $\mathbf{g}_j \in \mathbb{R}^2$ represents the spatial position of the j-th visual feature grid, and PE(·) is the positional encoding layer.

3.2.2 Concatenative feature fusion

The augmented node features are generated through direct concatenation:

$$\mathbf{F}_i = [\underbrace{\mathbf{X}_i}_{\substack{\text{Original}\\\text{graph feature}}} ; \underbrace{\tilde{\mathbf{v}}_i}_{\substack{\text{Projected}\\\text{visual feature}}}] \in \mathbb{R}^{d_g+d_v} \tag{17}$$

The concatenation operation synthesizes complementary facial representations by merging: (1) geometric information characterizing facial structure through $\mathbf{X}_i$, (2) visual-textural attributes describing appearance properties via $\widetilde{\mathbf{v}}_i$, and (3) spatial relationships implicitly maintained through position encoding mechanisms.

The resulting feature matrix $\mathbf{F}$ maintains the original graph topology while enriching each node with complementary visual information. This spatially-aligned representation serves as input for downstream graph processing layers.

3.3 Facial beauty prediction module

```
Input:  Graph input features: $\mathbf{X} \in \mathbb{R}^{N\times(d_g+d_v)}$
     // Node features
Output: Predicted beauty score: $\hat{y} \in [0, 5]$
// Step 1: Graph Attention Network (GAT) Encoding
$\mathbf{H} \leftarrow$ GAT($\mathbf{X}$)       // Update node features via GAT
// Step 2: Global Graph Pooling
$\mathbf{f}_{\text{graph}} \leftarrow \sum_{i=1}^{N} \alpha_i \mathbf{h}_i$
    where $\alpha_i = \text{softmax}(\mathbf{w}^\top \mathbf{h}_i)$    // Weighted sum of
     node features
// Step 3: MLP Regression
$\hat{y} \leftarrow$ MLP($\mathbf{f}_{\text{graph}}$)      // Predict facial beauty score
return $\hat{y}$
```

Algorithm 2. Facial beauty prediction from graph representation.

3.3.1 Generic beauty prediction

Facial graph features are updated using the Graph Node Attention Projection Fusion (GNAPF) block, which effectively integrates vision features with geometric representations. As illustrated in Figure 1 and Algorithm 2, these enhanced graph features are then processed by Graph Attention Networks (GAT), followed by global graph pooling and a fully connected MLP regression to predict facial beauty scores.

Given an input graph consisting of 86 nodes, where each node is characterized by a feature vector $\mathbf{h}_i \in \mathbb{R}^d$ (with $d = d_v + d_g$), the following procedure is applied to perform facial beauty score prediction:

To effectively model the relationship between nodes while considering their local structure, we adopt a Graph Attention Network (GAT). GAT introduces a self-attention mechanism that enables each node to selectively aggregate information from its neighbors. Specifically, the attention coefficient α_{ij} between node i and node j is computed as:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_j]\right)\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_k]\right)\right)} \tag{18}$$

Here, $\mathbf{W}$ is a shared linear transformation applied to each node, $\mathbf{a}$ is the learnable attention vector, and $\parallel$ denotes vector concatenation.

The updated feature representation of node i is obtained by aggregating the features of its neighbors weighted by the attention scores:

$$\mathbf{h}_i^{GAT} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\mathbf{W}\mathbf{h}_j\right) \tag{19}$$

This step allows the model to focus on more relevant neighboring nodes by assigning them higher attention weights during feature aggregation.

Following the GAT layer, we perform a global pooling operation to convert the set of node representations into a single fixed-length graph-level representation. We utilize either mean or max pooling across all nodes to extract holistic information:

$$\mathbf{h}^{pool} = \text{Pooling}(\{\mathbf{h}_i^{GAT} \mid i = 1, 2, \ldots, 86\}) \tag{20}$$

This operation yields a compact representation that captures the overall structure and semantic content of the entire graph, which is essential for the final prediction task.

The pooled graph representation $\mathbf{h}^{pool}$ is subsequently passed through a fully connected Multi-Layer Perceptron (MLP) to predict a continuous facial beauty score. The regression function can be defined as:

$$s = \text{MLP}(\mathbf{h}^{pool}) \tag{21}$$

where $s \in [0, 5]$ denotes the predicted beauty score. The MLP consists of one or more hidden layers with non-linear activation functions, and a final output layer that maps the feature vector to a scalar score.

3.3.2 Personalized beauty prediction

The introduced model incorporates a personalization module into a generic beauty prediction network, enabling adaptation to subjective aesthetic preferences. The main idea is to capture individual aesthetic preferences by leveraging user feedback on a small set of images, where each user provides beauty ratings for N images (in this study, $N = 10$). By learning the deviations between individual aesthetics and general public beauty perception, the model fine-tunes the weights of the generic prediction network to provide personalized beauty predictions that better align with the user's aesthetic taste.

The personalization procedure involves the following sequential operations:

Each user is asked to provide ratings for a small set of images. These ratings are used to infer the user's individual aesthetic preferences and how they differ from the general public's perception of beauty.

The collected user ratings are then used to learn the deviation between the user's aesthetic preferences and the general beauty perception encoded in the generic beauty prediction network. This deviation is modeled as a vector $\mathbf{d}_u$ that reflects the user's bias toward specific beauty attributes, such as symmetry, facial features, and expressions.

$$\mathbf{d}_u = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i^{user} - \hat{y}_i^{generic}) \tag{22}$$

where $\hat{y}_i^{user}$ is the beauty score predicted by the personalized model for the i-th image, and $\hat{y}_i^{generic}$ is the beauty score predicted by the generic model.

Based on the learned deviation vector $\mathbf{d}_u$, the weights of the generic model are fine-tuned to adjust for the user's specific aesthetic preferences. This fine-tuning is performed through a small learning rate to ensure the generic network retains its ability to predict beauty scores for the general population while adapting to individual biases.

By learning from a small set of user-specific ratings and adjusting the weights of the generic model, the proposed system can deliver personalized beauty predictions that better reflect the user's individual aesthetic preferences, while still maintaining the general applicability of the model.

Given five seed images, the ratings provided by user A are:$Y_A = (1, 2, 3, 4, 5)$, and the predictions from the universal model are:$\hat{Y}_U = (1.2, 2.5, 3.3, 3.8, 4.4)$.

Below is an example to illustrate this process:

The deviation vector is calculated as: $V_d = Y_A - \hat{Y}_U$, which is (-0.2, -0.5, -0.3, 0.2, 0.6). Let α be the weight used to integrate the deviation vector into the universal model prediction:$\hat{Y}_A = \hat{Y}_U + \alpha \cdot \text{Deviation Vector}$ For this example, setting $\alpha = 0.5$, the updated predictions are:$\hat{Y}_A = (1.1, 2.25, 3.15, 3.9,\ 4.7)$

The model is fine-tuned using the Mean Squared Error (MSE) between user A's ratings and the model's predictions:$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(Y_{A,i} - \hat{Y}_{A,i})^2$, For the given example, the MSE is calculated as:

$$\mathcal{L}_{\text{MSE}} = \tfrac{1}{5}((1-1.1)^2 + (2-2.25)^2 + (3-3.15)^2 + (4-3.9)^2 \\ +(5-4.7)^2) = 0.039$$

After fine-tuning, the model provides a personalized prediction for user A:$\hat{Y}_A = (1.1, 2.25, 3.15, 3.9, 4.7)$ The updated prediction for a user's beauty score is then calculated as:

$$\hat{y}_i^{personalized} = f(\mathbf{h}_i, \mathbf{W}^{generic} + \mathbf{d}_u) \tag{23}$$

where $\mathbf{h}_i$ is the feature vector for the i-th image, and f represents the model's output function.

4 Experiments and results

4.1 Experimental settings

4.1.1 Dataset and subsets

The SCUT-FBP5500 dataset contains 5500 frontal faces, aged from 15 to 60, with a neutral expression. It can be divided into four subsets with different races and genders to assess model performance across different demographic groups. We divide the dataset into four subsets: 2,000 Asian females, 2,000 Asian males, 750 Caucasian females, and 750 Caucasian males. Most of the images of the SCUT-FBP5500 were collected from the Internet. All the images are labeled with beauty scores ranging from 1 to 5.

4.1.2 Train/test splitting

For each demographic subset and the full dataset, we adopt a five-fold cross-validation strategy. Each fold splits the data into 80% for training and 20% for testing, ensuring non-overlapping images between the training and test sets. We use fixed random seeds to guarantee consistency across all experiments.

4.1.3 Training settings

All models are trained for 100 epochs with a batch size of 32 using the Adam optimizer (L2 weight decay = 1e-5) to mitigate overfitting. The initial learning rate is set to 1e-4, and it is reduced by a factor of 0.1 if the validation Pearson Correlation (PC) does not surpass the historical best for 50 consecutive epochs. Early stopping is applied with a patience of 10 epochs (i.e., training stops if the validation PC does not improve for 10 consecutive epochs). To ensure reproducibility, fixed random seeds are used for data splitting, model initialization, and all training processes. The same training settings are adopted for all baseline models to guarantee a fair comparison.

4.1.4 Evaluation metrics

We use three metrics for facial beauty prediction–Pearson Correlation (PC), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

For generic prediction, PC quantifies the linear correlation between predicted and ground-truth beauty scores (range: [-1, 1]), with higher values indicating stronger consistency in capturing overall aesthetic trends.

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{24}$$

MAE measures the average absolute difference between predictions and true scores, reflecting mean prediction error (lower values = better accuracy).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{25}$$

RMSE, the square root of the mean squared error, emphasizes large errors to assess model stability (smaller values = more robust performance against extreme cases).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{26}$$

4.1.5 Repeated trials and significance testing (paired *t*-tests)

To ensure the statistical reliability of the results, we conducted five repeated experiments for the top three performing methods identified in the initial evaluation. This adjustment was made in response to the need for significance testing, which aims to verify the robustness of performance differences among high-ranking models. The results of these repeated experiments are used to conduct significance testing and are reported alongside the main performance tables, including corresponding p-values where appropriate.

4.2 Feature extraction module analysis

Comparison of the features used in this study with other geometric and visual features, along with their predictive performance on the Asian female subset of the SCUT-FBP5500 dataset, is presented in Table 1. As quantitatively demonstrated in Table 1, three core conclusions can be drawn from the analysis:

First, our proposed facial geometric graph features achieve a statistically significant improvement in Pearson correlation (PC) of 6%–8% over conventional geometric descriptors. This improvement suggests that geometric graph features may capture distinct aesthetic dimensions of facial structure–such as the topological interplay between landmarks and local shape curvature–that isolated geometric measurements are less likely to reveal.

Second, the deep features extracted using the visual Mamba mechanism in this study exhibit a significant improvement over the traditional ResNet-50 features. This highlights the superior ability of the visual Mamba mechanism in modeling long-range spatial dependencies, which are crucial for the perception of overall beauty.

Third, our proposed combined feature method (Ours) shows favorable performance over baseline approaches across all four demographic subsets of the SCUT-FBP5500 dataset, achieving the highest PC scores. This result underscores the potential effectiveness of our fusion strategy, which integrates geometric graph features with visual embeddings to capture both structural precision and texture harmony. This approach appears to be more effective than the baselines' simpler fusion of these modalities, which may not fully leverage their cross-modal synergy. Such integration aligns with human beauty judgments, which rely on both structural proportions and textural qualities, suggesting the benefit of a unified representation for automated assessment.

To investigate the impact of fusion between different sizes of Vision Mamba features and facial geometric graph features on facial beauty prediction, we conducted experiments with four different feature sizes. The results are presented in Figure 4. As shown, the 7 × 7 feature size yielded the best performance, indicating that this size provides an optimal balance between local details and global context. Therefore, all subsequent experiments in this study will be conducted using this feature size.

Moreover, as shown in Table 2, statistical significance tests were conducted to evaluate the performance differences among the feature extraction strategies. The results confirm that the improvements achieved by the proposed feature strategy are not only consistent across subsets but also statistically significant when compared to conventional approaches. This supports the robustness and validity of our design choices from a statistical perspective, indicating that the performance gains observed are unlikely to be due to random chance.

4.3 Feature fusion module analysis

To rigorously evaluate the performance of our proposed Graph Node Attention Projection Fusion (GNAPF) mechanism, we conducted comparative experiments with two heterogeneous feature fusion baselines, designed to isolate the contribution of our attention-driven, spatially aware fusion strategy.

The first baseline, Flatten-Concat, employs a straightforward fusion strategy: Graph features consist of two components—86 node (landmark) features, each encoded as a 10-dimensional vector, resulting in a total of 860 dimensions (86 × 10), and 168 edge (triangulation) features, each represented as a 6-dimensional vector, contributing 1,008 dimensions (168 × 6). These node and edge features are first concatenated along the feature dimension to form a combined graph feature vector, which is then flattened into a single one-dimensional vector (860 + 1,008 = 1,868 dimensions). This flattened graph vector is subsequently concatenated with the 192-dimensional visual features extracted from Vision Mamba, resulting in a final fused feature vector of 2,060 dimensions (1,868 + 192).

The second baseline, AvgAgg-Concat, introduces a dimensionality reduction technique through aggregation: The graph features are independently aggregated by computing the mean of the 86 node features (each 10-dimensional), which results in a 10-dimensional vector, and the mean of the 168 edge features (each 6-dimensional), yielding a 6-dimensional vector. These two averaged representations (10 + 6 = 16 dimensions) are then concatenated with the 192-dimensional Vision Mamba visual features, producing a final fused feature vector of 208 dimensions (16 + 192).

Experimental results are presented in Table 3. As shown, the proposed GNAPF mechanism achieves statistically significant improvements over both baseline fusion approaches, indicating its effectiveness in capturing and integrating the complementary information from facial geometric structures and deep visual representations. Rather than relying on definitive claims of superiority, we frame these gains in terms of favorable performance trends supported by statistical evidence. In addition to evaluating fusion strategies, we also investigated the influence of position encoding on prediction performance. The inclusion of position encoding consistently led to improved results, suggesting that spatial contextualization plays an important role in enhancing the model's capacity to model facial attractiveness.

To further validate these observations, we conducted statistical significance tests (paired t-tests) on the performance metrics of different fusion strategies reported in Table 2. The results demonstrate that the improvements achieved by GNAPF are statistically significant when compared with the other two strategies, reinforcing the reliability and generalizability of the proposed fusion mechanism from a statistical standpoint.

As illustrated in Table 4, the four models–Inception-v3, ResNet50, ViT, and ViM–exhibit substantial performance improvements in the personalized facial attractiveness prediction task upon integration with the three fusion strategies (GNAPF, AvgAgg-Concat, and Flatten-Concat). Notably, while the numerical discrepancies in general facial attractiveness prediction across these models (with or without the fusion strategies) are relatively marginal, the performance gains in the personalized setting are remarkably prominent: our proposed method

TABLE 1 Ablation analysis: feature performance on SCUT-FBP5500 subsets.

Feature representation	Asian female			Asian male			Caucasian female			Caucasian male		
	PC	MAE	RMSE	PC	MAE	RMSE	PC	MAE	RMSE	PC	MAE	RMSE
Geometric features												
GeomFeat 1[a]	0.682	0.367	0.458	0.659	0.368	0.467	0.695	0.362	0.453	0.650	0.361	0.476
GeomFeat 2[b]	0.671	0.376	0.472	0.647	0.386	0.485	0.683	0.381	0.475	0.638	0.391	0.492
Geo-Graph	**0.735**	**0.356**	**0.451**	**0.729**	**0.367**	**0.458**	**0.738**	**0.351**	**0.448**	**0.731**	**0.364**	**0.454**
Vision features												
Inception-v3	0.889	0.252	0.316	0.884	0.273	0.346	0.891	0.252	0.311	0.887	0.266	0.328
ResNet50	0.907	0.219	0.289	0.902	0.232	0.301	0.912	0.231	0.281	0.904	0.229	0.284
ViT	0.910	0.215	0.283	0.904	0.230	0.296	0.914	0.227	0.277	0.906	0.226	0.283
ViM[c]	**0.916**	**0.213**	**0.279**	**0.907**	**0.228**	**0.290**	**0.921**	**0.222**	**0.269**	**0.912**	**0.225**	**0.275**
Combined features												
GeomFeat 1+ViM	0.919	0.210	0.276	0.911	0.223	0.286	0.920	0.217	0.261	0.914	0.222	0.273
GeomFeat 2+ViM	0.918	0.211	0.278	0.911	0.224	0.289	0.921	0.219	0.263	0.915	0.219	0.269
Ours	**0.924**	**0.188**	**0.258**	**0.918**	**0.210**	**0.267**	**0.922**	**0.195**	**0.263**	**0.920**	**0.216**	**0.264**

PC, Pearson correlation ↑; MAE, mean absolute error ↓; RMSE, root mean squared error ↓.

[a] GeomFeat 1: geometric features from Swift and Remington (2019).

[b] GeomFeat 2: geometric features from Zhang et al. (2016).

[c] ViM, vision Mamba model (Zhu et al., 2024).

Bold values indicate the optimal performance among the different feature representations.

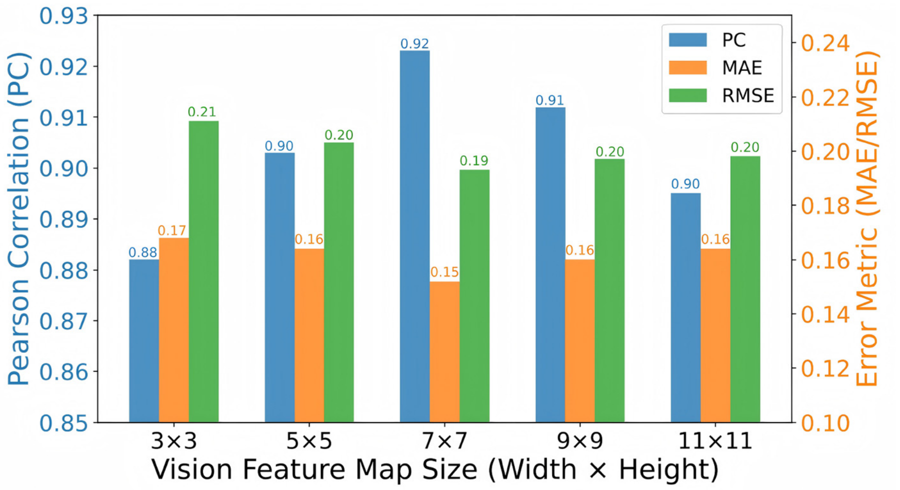

FIGURE 4
Effects of feature map dimensions in Vision Mamba architecture on facial beauty prediction performance.

TABLE 2 Ablation analysis: performance significance comparison between our structural strategy and four alternative strategies on SCUT-FBP5500 subsets subsets.

Methods	Asian female	Asian male	Caucasian female	Caucasian male
Combined features				
Ours vs. GeomFeat 1+ViM	0.019	0.028	0.035	0.023
Ours vs. GeomFeat 2+ViM	0.038	0.019	0.044	0.043
Fusion strategy				
Ours vs. AvgAgg-Concat	0.031	0.042	0.038	0.023
Ours vs. Flatten-Concat	0.035	0.026	0.045	0.037

(backbones integrated with GNAPF) achieves a performance gain of at least 8% compared to both the original backbones and the variants integrated with AvgAgg-Concat/Flatten-Concat. This significant enhancement validates that the GNAPF module can effectively capture and leverage graph-structured information inherent in facial images, which is critical for modeling individual-specific aesthetic preferences. Consequently, this enables the development of a more accurate and personalized facial attractiveness prediction model, fully demonstrating the unique value of the GNAPF module in addressing the personalized prediction task.

Collectively, these findings support the rationale behind our method's design choices: (1) attention-driven fusion facilitates the modeling of cross-modal dependencies between geometric and visual features; (2) spatial contextualization (via position encoding) is crucial for fully leveraging the predictive value of geometric representations. These insights highlight the broader importance of adopting fine-grained, modality-aware integration strategies in cross-modal prediction tasks, where naive concatenation or aggregation often falls short of capturing complex intermodal relationships.

4.4 Generic beauty prediction comparison

As quantitatively demonstrated in Table 5, our generic prediction module achieves state-of-the-art performance across all demographic subgroups while revealing critical insights into bias mitigation.

The framework consistently shows favorable performance across all categories, with the most pronounced gains observed in all the subgroup cohorts (PC = 0.923), where it achieves statistically significant improvements in Pearson correlation compared to the prior best method. This notable margin suggests enhanced capability in modeling nuanced aesthetic attributes prevalent in this demographic, possibly due to our hybrid feature fusion mechanism that jointly optimizes geometric and texture cues.

Ethnically, the model shows robust advantages for Asian subgroups over Caucasian counterparts, a gap of 0.2% that may reflect either training data distribution imbalances or culturally divergent beauty annotation patterns in benchmark datasets. Notably, gender-based analysis reveals universally higher accuracy, with our method achieving a 0.4%–3.4% PC improvement over multi-feature fusion baselines, indicating superior modeling of gender-specific aesthetic traits.

TABLE 3 Fusion strategy ablation: performance comparison on SCUT-FBP5500 subsets.

Fusion strategy	Asian Female			Asian male			Caucasian female			Caucasian male		
	PC	MAE	RMSE	PC	MAE	RMSE	PC	MAE	RMSE	PC	MAE	RMSE
AvgAgg-concat[a]	0.918	0.211	0.272	0.914	0.223	0.288	0.922	0.224	0.271	0.915	0.217	0.272
Flatten-concat[b]	0.917	0.213	0.276	0.912	0.226	0.293	0.920	0.228	0.276	0.914	0.221	0.277
GNAPF w/o PE[c]	0.897	0.242	0.295	0.887	0.253	0.308	0.902	0.237	0.293	0.891	0.252	0.291
Ours	**0.924**	**0.188**	**0.258**	**0.918**	**0.210**	**0.267**	**0.922**	**0.195**	**0.263**	**0.920**	**0.216**	**0.264**

PC, Pearson correlation ↑; MAE, mean absolute error ↓; RMSE, root mean squared error ↓.

[a] AvgAgg-Concat: late-fusion with mean-aggregated geometric features.

[b] Flatten-Concat: naive fusion via flattened graph representation.

[c] GNAPF w/o PE: Graph Neural Attentional Prediction Framework without positional encoding.

Bold values indicate the optimal performance among the different fusion strategies.

4.5 Personalization capability

This study evaluates personalized facial beauty prediction frameworks using the SCUT-FBP5500 dataset, which includes attractiveness ratings from 60 distinct evaluators. To assess personalization, we divided raters into different training and testing sets across five configurations (30, 35, 40, 45, and 50 training raters).

As shown in Figure 5 and Table 6, our method consistently shows favorable performance compared to existing approaches. With 50 training raters, it achieves a PC of 0.738, surpassing Lin et al. (0.724) and Lebedeva et al. (0.701). The performance gap widens as the number of training raters decreases, demonstrating the robustness of our method when faced with limited personalized data. This robustness is attributed to our personalized fine-tuning approach, which adapts the generic beauty prediction model based on user-specific aesthetic preferences. By leveraging a small set of user ratings, the model effectively captures individual biases and refines the prediction to align with the user's tastes, reducing overfitting while maintaining general applicability.

A deeper analysis of the top three methods across 10 test raters in the 50-training-rater configuration, as shown in Figure 6, reveals several key findings. Our method consistently achieves the highest PC across all raters, demonstrating its adaptability to diverse aesthetic preferences. In more challenging cases, such as Rater 7, who exhibits highly unique preferences, our approach maintains a significant lead (0.718 vs. 0.695 for Lin et al.).

This work sets a new state-of-the-art benchmark in personalized beauty assessment, particularly in scenarios with limited training raters or diverse aesthetic preferences. Future research will explore cross-cultural generalization and multi-modal preference modeling to further enhance the framework's versatility.

As shown in Figure 7, we compare the prediction outcomes of three different methodologies for the ratings of IMGAF3, evaluated by ten raters. The performance analysis demonstrates that the proposed method exceeds the baseline approach in terms of curve fitting accuracy. This improvement reflects a notable enhancement in both predictive precision and the ability to conduct personalized image assessments, underscoring the method's superior capacity to more accurately capture the preferences of individual raters.

Figure 8 presents a comparative analysis of facial images, general and personalized heatmaps derived from the weighting information of 86 facial keypoints in the Graph Attention Network (GAT), alongside aesthetic scores from two raters (Rater A and Rater B). This illustration underscores the effectiveness of our method in encapsulating individual aesthetic experiences. The results unveil notable subjective preferences: Rater A assigned lower scores to the first two faces and higher scores to the last two faces compared to Rater B, thereby highlighting divergent aesthetic judgments. These differences are further manifested in the personalized heatmaps: Rater A's heatmaps exhibit a more balanced attention distribution across facial regions, signifying a focus on overall harmony, while Rater B's heatmaps emphasize focal features such as the eyes and mouth, indicating a preference for local facial details. Collectively, these findings substantiate that our method can effectively capture and quantify individual aesthetic tendencies, translating them into discernible heatmap patterns and score variations.

TABLE 4 Ablation analysis: method performance before/after integrating GNAPF/avgagg-concat/flatten-concat modules for personalized facial attractiveness prediction (50 raters).

Methods	With GNAPF	w/o GNAPF	AvgAgg-Concat	Flatten-Concat
Inception-v3	0.517	0.467	0.478	0.487
ResNet50	0.604	0.553	0.564	0.543
VIT	0.646	0.561	0.585	0.573
VIM	**0.738 (Ours)**	0.642	0.658	0.637

Bold values indicate the best performance among the compared methods.

TABLE 5 Cross-demographic beauty prediction performance: Pearson correlation coefficients for six methods and statistical significance tests (our method vs. top baselines: anchor-net & REX-INCEP)—*Mean* ± *Std*, *p*-values.

Method	Asian		Caucasian		Gender		Ethnic		Overall
	F	M	F	M	F	M	As	Cau	
Inception-v3	0.889	0.884	0.891	0.887	0.890	0.885	0.889	0.886	0.887
ResNet50	0.907	0.902	0.912	0.904	0.909	0.902	0.905	0.908	0.906
ViT	0.910	0.904	0.914	0.906	0.911	0.904	0.907	0.909	0.908
ACBFN [a]	0.807	0.813	0.794	0.806	0.811	0.804	0.775	0.789	0.818
Anchor-Net[b]	0.915±0.003	0.91 ± 0.005	0.917±0.006	0.913±0.005	0.915±0.004	0.911±0.005	0.912±0.004	0.915±0.006	0.912±0.003
REX-INCEP[c]	0.917±0.005	0.911±0.004	0.918±0.005	0.914±0.006	0.916±0.005	0.912±0.003	0.915±0.004	0.914±0.005	0.913±0.004
DyAttenConv[d]	0.909	0.903	0.913	0.905	0.911	0.903	0.906	0.908	0.906
Ours	**0.924 ± 0.003**	**0.918 ± 0.004**	**0.922 ± 0.004**	**0.920 ± 0.003**	**0.924 ± 0.005**	**0.918 ± 0.004**	**0.921 ± 0.003**	**0.920 ± 0.005**	**0.922 ± 0.003**
p-values									
Ours vs. Anchor-Net	0.028	0.033	0.031	0.042	0.038	0.029	0.037	0.041	0.035
Ours vs. REX-INCEP	0.025	0.029	0.035	0.047	0.027	0.017	0.027	0.032	0.029

F/M, female/male; As/Cau, Asian/Caucasian; PC, Pearson correlation (↑ higher is better).

[a] ACBFN, Adaptive Cross-Cultural Beauty Fusion Network (Yan and Ye, 2025).

[b] Anchor-Net, self-supervised learning model (Bae et al., 2024).

[c] REX-INCEP, robust ensemble method (Bougourzi et al., 2022).

[d] DyAttenConv, dynamic attentive convolution (Sun et al., 2024b).

Bold values indicate the best performance among the compared methods.

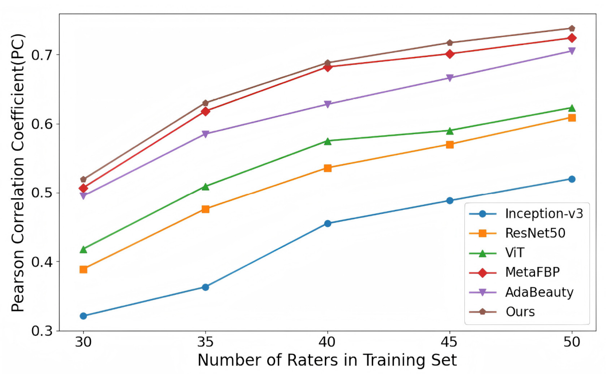

FIGURE 5
Personalized prediction performance metrics.

TABLE 6 Personalized beauty prediction performance on SCUT-FBP5500 dataset.

Method	Number of raters			
	30	35	40	50
Inception-v3	0.359	0.384	0.451	0.517
ResNet50	0.387	0.482	0.542	0.604
Vision transformer	0.410	0.504	0.574	0.626
MetaFBP[a]	0.502	0.623	0.679	0.724
AdaBeauty[b]	0.485	0.583	0.628	0.701
Ours	**0.523**	**0.632**	**0.684**	**0.738**

Performance metrics represent Pearson correlation coefficient (higher values indicate better agreement with human raters).

[a] MetaFBP, meta-learning framework (Lin et al., 2023).

[b] AdaBeauty, adaptive beauty assessment (Lebedeva et al., 2023).

Bold values indicate the best performance among the compared methods.

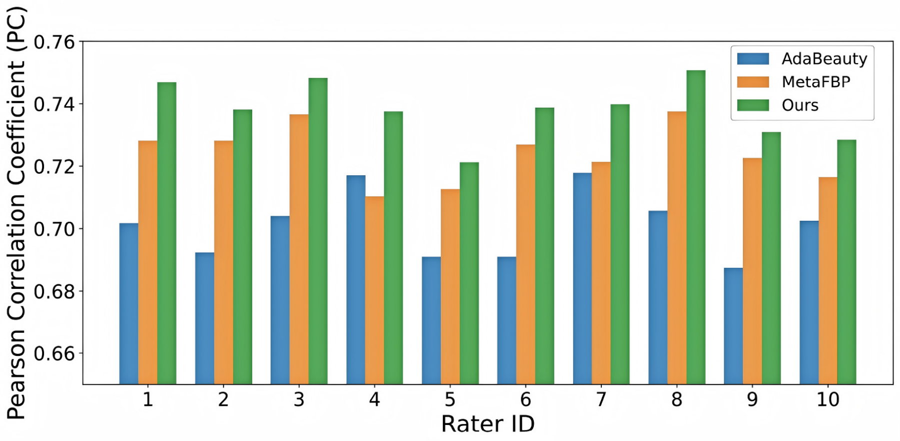

FIGURE 6
Performance comparison of top three methods across 10 test raters in the 50-training-rater configuration.

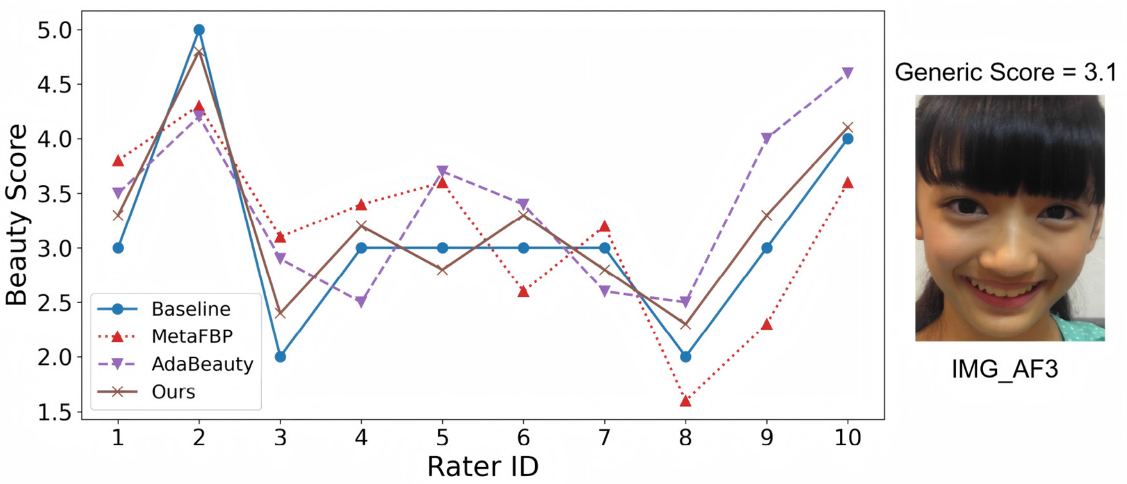

FIGURE 7
Comparison of prediction outcomes for IMGAF3 ratings across three methods.

5 Conclusions and future work

This study introduces an innovative framework for personalized facial beauty evaluation, which combines global visual features with geometric graphs constructed from 86 facial key points. The framework effectively captures both overall aesthetic characteristics and localized structural relationships, allowing the system to finely tune its responses to individual beauty preferences. Comprehensive validation on the diverse SCUT-FBP5500 dataset confirms that the method achieves statistically significant improvements compared to existing approaches. Significantly, the proposed method demonstrates a strong correlation across all test raters, highlighting its flexibility in accommodating subjective preferences.

Our framework underscores the importance of simultaneously modeling global visual patterns and detailed geometric configurations for accurate facial beauty assessment. This fusion strategy delivers robust performance across various ethnicities, ages, genders, and facial expressions, illustrating the versatility of the approach. Looking forward, future work will focus on developing adaptive feature weighting mechanisms to dynamically balance the visual and geometric components based on individual rater characteristics. Additionally, we aim to extend the framework to 3D facial modeling–enhancing real-world applicability by leveraging precise structural analysis principles, as demonstrated in recent deep learning research on craniomaxillofacial multi-structure segmentation (Bao et al., 2024, 2025). We also plan to explore active learning techniques to reduce the need for extensive annotations. Integrating this framework with recommendation systems holds promising potential to address cold-start problems in aesthetic-driven applications. These future advancements will help bridge the gap between

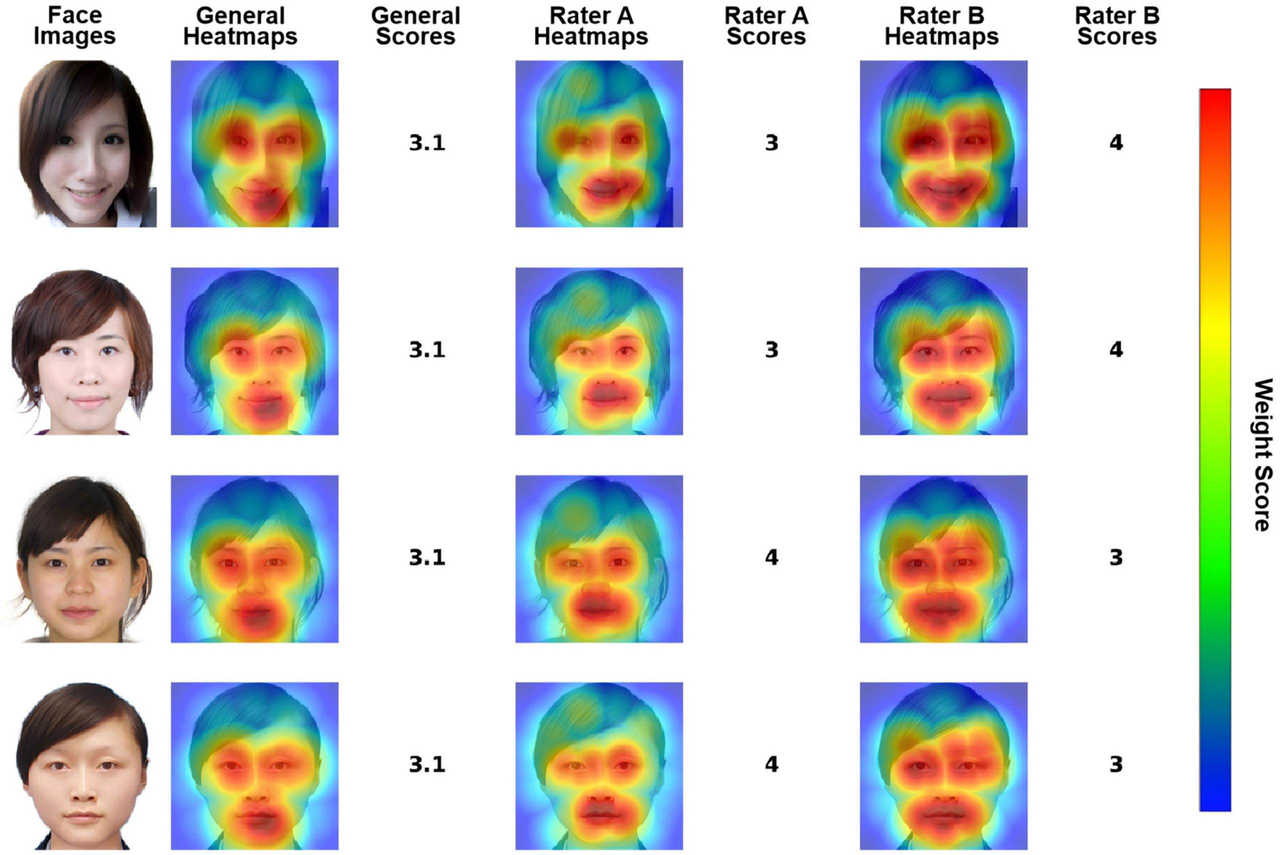

FIGURE 8
Visualization of facial images, general heatmaps, personalized heatmaps (Rater A and Rater B) with corresponding aesthetic scores. Image(s) taken from Liang et al. (2018).

computational beauty assessment and the nuanced complexities of human perception.

Data availability statement

Publicly available datasets were analyzed in this study. The direct link to the SCUT-FBP5500 dataset is https://github.com/HCIILAB/SCUT-FBP5500-Database-Release. The repository name is HCIILAB/SCUT-FBP5500-Database-Release, and there is no corresponding accession number for this dataset.

Ethics statement

The studies involving humans were approved by Ethics Committee of South China University of Technology; Affiliation: South China University of Technology (Guangzhou, Guangdong Province, China). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study, as all facial image data utilized were sourced from the publicly available SCUT-FBP5500 dataset, which was officially released by the HCII Lab of South China University of Technology. According to the official documentation of the SCUT-FBP5500 dataset (accessible via http://www.hcii-lab.net/data/SCUT-FBP5500.html) and its associated published literature (e.g., the dataset's original paper: "SCUT-FBP5500: A Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction"), the research team at South China University of Technology–who initially collected the dataset–had obtained written informed consent from all participants prior to data release. This consent explicitly permitted the use of the anonymized facial data for non-commercial academic research, including studies focused on facial beauty prediction like the current work. As secondary users of this ethically approved, publicly shared dataset, we did not need to separately secure additional written informed consent from the original participants, which aligns with standard ethical practices for reusing pre-validated public research data in academic studies. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

KW: Formal analysis, Validation, Methodology, Writing – original draft, Investigation, Conceptualization. YL: Conceptualization, Writing – review & editing. DH: Project administration, Validation, Writing – original draft, Funding acquisition. JF: Writing – review & editing, Funding acquisition, Project administration. XF: Conceptualization, Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the National Natural Science Foundation of China (Grant 62403473) and partially supported by the Natural Science Basic Research Program of Shaanxi (Program 2024JC-YBQN-0675).

Conflict of interest

YL was employed at Shaanxi Chang'an Computing Technology Co., Ltd.

The remaining authors declare that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bae, J., Buu, S.-J., and Lee, S. (2024). Anchor-net: distance-based self-supervised learning model for facial beauty prediction. *IEEE Access* 12, 61375–61387. doi: 10.1109/ACCESS.2024.3394870

Bao, J., Tan, Z., Sun, Y., Xu, X., Liu, H., Cui, W., et al. (2025). Deep ensemble learning-driven fully automated multi-structure segmentation for precision craniomaxillofacial surgery. *Front. Bioeng. Biotechnol.* 13:1580502. doi: 10.3389/fbioe.2025.1580502

Bao, J., Zhang, X., Xiang, S., Liu, H., Cheng, M., Yang, Y., et al. (2024). Deep learning-based facial and skeletal transformations for surgical planning. *J. Dent. Res.* 103, 809–819. doi: 10.1177/00220345241253186

Bougourzi, F., Dornaika, F., and Taleb-Ahmed, A. (2022). Deep learning based face beauty prediction via dynamic robust losses and ensemble regression. *Knowl.-Based Syst.* 242:108246. doi: 10.1016/j.knosys.2022.108246

Chen, F., Xiao, X., and Zhang, D. (2018). Data-driven facial beauty analysis: prediction, retrieval and manipulation. *IEEE Trans. Affect. Comput.* 9, 205–216. doi: 10.1109/TAFFC.2016.2599534

Dornaika, F., Wang, K., Arganda-Carreras, I., Elorza, A., and Moujahid, A. (2020). Toward graph-based semi-supervised face beauty prediction. *Expert Syst. Appl.* 142:112990. doi: 10.1016/j.eswa.2019.112990

Eisenthal, Y., Dror, G., and Ruppin, E. (2006). Facial attractiveness: beauty and the machine. *Neural Comput.* 18, 119–142. doi: 10.1162/089976606774841602

Gan, J., Li, L., Zhai, Y., and Liu, Y. (2014). Deep self-taught learning for facial beauty prediction. *Neurocomputing* 144, 295–303. doi: 10.1016/j.neucom.2014.05.028

Gan, J., Luo, H., Xiong, J., Xie, X., Li, H., Liu, J., et al. (2024). Facial beauty prediction combined with multi-task learning of adaptive sharing policy and attentional feature fusion. *Electronics* 13:179. doi: 10.3390/electronics13010179

Gan, J., Xie, X., Zhai, Y., He, G., Mai, C., Luo, H., et al. (2023). Facial beauty prediction fusing transfer learning and broad learning system. *Soft Comput.* 27, 13391–13404. doi: 10.1007/s00500-022-07563-1

Gray, D., Yu, K., Xu, W., and Gong, Y. (2010). "Predicting facial beauty without landmarks," in *European Conference on Computer Vision* (Cham: Springer), 434–447. doi: 10.1007/978-3-642-15567-3_32

Gunes, H., and Piccardi, M. (2006). Assessing facial beauty through proportion analysis by image processing and supervised learning. *Int. J. Hum. Comput. Stud.* 64, 1184–1199. doi: 10.1016/j.ijhcs.2006.07.004

Huang, D., Xia, Z., Li, L., and Ma, Y. (2022). Pain estimation with integrating global-wise and region-wise convolutional networks. *IET Image Process* 17, 637–648. doi: 10.1049/ipr2.12639

Ibrahem, A., Saeed, J., and Abdulazeez, A. (2025). Insights into automated attractiveness evaluation from 2d facial images: a comprehensive review. *Int Arab J. Inf. Technol.* 22, 77–98. doi: 10.34028/iajit/22/1/7

Ibrahem, A. H., and Abdulazeez, A. M. (2025). A comprehensive review of facial beauty prediction using multi-task learning and facial attributes. *ARO-Sci. J. Koya Univ.* 13, 10–21. doi: 10.14500/aro.11850

Juravle, G., and Spence, C. (2024). Beauty is context-dependent: naturalness, familiarity, and semantic meaning influence the appreciation of geometric shapes. *Iperception.* 15:20416695241303004. doi: 10.1177/20416695241303004

Lebedeva, I., Guo, Y., and Ying, F. (2021). "Deep facial features for personalized attractiveness prediction," in *International Conference on Digital Image Processing, Vol. 11878* (Bellingham, WA: SPIE), 118780A. doi: 10.1117/12.2599699

Lebedeva, I., Ying, F., and Guo, Y. (2023). Personalized facial beauty assessment: a meta-learning approach. *Vis. Comput.* 39, 1095–1107. doi: 10.1007/s00371-021-02387-w

Li, L., Yao, Z., Gao, S., Han, H., and Xia, Z. (2024). Face anti-spoofing via jointly modeling local texture and constructed depth. *Eng. Appl. Artif. Intell.* 133:108345. doi: 10.1016/j.engappai.2024.108345

Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., et al. (2018). "Beautygan: instance-level facial makeup transfer with deep generative adversarial network," in *ACM Multimedia* (New York, NY: ACM). doi: 10.1145/3240508.3240618

Liang, L., Lin, L., Jin, L., Xie, D., and Li, M. (2018). "Scut-fbp5500: a diverse benchmark dataset for multi-paradigm facial beauty prediction," in *International Conference on Pattern Recognition* (Beijing: IEEE), 1598–1603. doi: 10.1109/ICPR.2018.8546038

Lin, L., Shen, Z., Yin, J.-L., Liu, Q., Yu, Y., Chen, W., et al. (2023). "Metafbp: learning to learn high-order predictor for personalized facial beauty prediction," *ACM International Conference on Multimedia* (New York, NY: ACM), 6072–6080. doi: 10.1145/3581783.3612319

Londono, J., Ghasmi, S., Lawand, G., Mirzaei, F., Akbari, F., Dashti, M., et al. (2024). Assessment of the golden proportion in natural facial esthetics: a systematic review. *J. Prosthet. Dent.* 131, 804–810. doi: 10.1016/j.prosdent.2022.04.026

Moridani, M. K., Jamiee, N., and Saghafi, S. (2023). Human-like evaluation by facial attractiveness intelligent machine. *Int. J. Cogn. Comput. Eng.* 4, 160–169. doi: 10.1016/j.ijcce.2023.04.001

Peng, T., Li, M., Chen, F., Xu, Y., and Zhang, D. (2023). Learning efficient facial landmark model for human attractiveness analysis. *Pattern Recognit.* 138:109370. doi: 10.1016/j.patcog.2023.109370

Peng, T., Li, M., Chen, F., Xu, Y., and Zhang, D. (2024). Geometric prior guided hybrid deep neural network for facial beauty analysis. *CAAI Trans. Intell. Technol.* 9, 467–480. doi: 10.1049/cit2.12197

Sun, Z., Lin, L., Yu, Y., and Jin, L. (2024a). Learning feature alignment across attribute domains for improving facial beauty prediction. *Expert Syst. Appl.* 249:123644. doi: 10.1016/j.eswa.2024.123644

Sun, Z., Xiao, Z., Yu, Y., and Lin, L. (2024b). Dynamic attentive convolution for facial beauty prediction. *IEICE Trans. Inf.* E107-D, 239–243. doi: 10.1587/transinf.2023EDL8058

Swift, A., and Remington, B. (2019). "The mathematics of facial beauty," in *Injectable Fillers: Facial Shaping and Contouring*, eds. D. H. Jones, and A. Swift (Hoboken, NJ: Wiley), 29–61. doi: 10.1002/9781119046974.ch2

Wang, K., Feng, X., Dornaika, F., Huang, D., and Xia, Z. (2022). "A multi-graph fusion based manifold embedding for face beauty prediction," in *International Conference on Image Processing and Media Computing* (Xi'an: IEEE), 129–134. doi: 10.1109/ICIPMC55686.2022.00032

Whitehill, J., and Movellan, J. R. (2008). "Personalized facial attractiveness prediction," in *IEEE International Conference on Automatic Face & Gesture Recognition* (Amsterdam: IEEE), 1–7. doi: 10.1109/AFGR.2008.4813332

Xiao, Q., Wu, Y., Wang, D., Yang, Y.-L., and Jin, X. (2021). Beauty3dfacenet: deep geometry and texture fusion for 3d facial attractiveness prediction. *Comput. Graph.* 98, 11–18. doi: 10.1016/j.cag.2021.04.023

Xie, D., Liang, L., Jin, L., Xu, J., and Li, M. (2015). "Scut-fbp: a benchmark dataset for facial beauty perception," in *IEEE International Conference on Systems, Man, and Cybernetics* (Hong Kong: IEEE), 1821–1826. doi: 10.1109/SMC.2015.319

Yan, G., and Ye, Y. (2025). "Adaptive cross-cultural beauty fusion network: a hybrid deep learning framework for culturally-aware facial beauty prediction," in *2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL)* (Ningbo: IEEE), 01–04. doi: 10.1109/CVIDL65390.2025.11085585

Zhai, Y., Huang, Y., Xu, Y., Gan, J., Cao, H., Deng, W., et al. (2020). Asian female facial beauty prediction using deep neural networks via transfer learning and multi-channel feature fusion. *IEEE Access* 8, 56892–56907. doi: 10.1109/ACCESS.2020.2980248

Zhang, D., Chen, F., and Xu, Y. (2016). "Beauty analysis fusion model of texture and geometric features," in *Computer Models for Facial Beauty Analysis* (Cham: Springer), 89–101. doi: 10.1007/978-3-319-32598-9_6

Zhang, D., Zhao, Q., and Chen, F. (2011). Quantitative analysis of human facial beauty using geometric features. *Pattern Recognit.* 44, 940–950. doi: 10.1016/j.patcog.2010.10.013

Zhang, L., Zhang, D., Sun, M.-M., and Chen, F.-M. (2017). Facial beauty analysis based on geometric feature: toward attractiveness assessment application. *Expert Syst. Appl.* 82, 252–265. doi: 10.1016/j.eswa.2017.04.021

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X., et al. (2024). Vision mamba: efficient visual representation learning with bidirectional state space model. *arXiv [preprint]*. arXiv:2401.09417. doi: 10.48550/arXiv.2401.09417