SYNERGYNET: FUSING GENERATIVE PRIORS AND STATE-SPACE MODELS FOR FACIAL BEAUTY PREDICTION

Djamel Eddine Boukhari

Scientific and Technical Research Centre for Arid Areas, CRSTRA 07000, Biskra, Algeria boukhari-djameleddine@univ-eloued.dz

September 23, 2025

ABSTRACT

The automated prediction of facial beauty is a benchmark task in affective computing that requires a sophisticated understanding of both local aesthetic details (e.g., skin texture) and global facial harmony (e.g., symmetry, proportions). Existing models, based on either Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs), exhibit inherent architectural biases that limit their performance; CNNs excel at local feature extraction but struggle with long-range dependencies, while ViTs model global relationships at a significant computational cost. This paper introduces the Mamba-Diffusion Network (MD-Net), a novel dual-stream architecture that resolves this tradeoff by delegating specialized roles to state-of-the-art models. The first stream leverages a frozen U-Net encoder from a pre-trained latent diffusion model, providing a powerful generative prior for fine-grained aesthetic qualities. The second stream employs a Vision Mamba (Vim), a modern state-space model, to efficiently capture global facial structure with linear-time complexity. By synergistically integrating these complementary representations through a cross-attention mechanism, MD-Net creates a holistic and nuanced feature space for prediction. Evaluated on the SCUT-FBP5500 benchmark, MD-Net sets a new state-of-the-art, achieving a Pearson Correlation of 0.9235 and demonstrating the significant potential of hybrid architectures that fuse generative and sequential modeling paradigms for complex visual assessment tasks.

Keywords Facial Beauty Prediction · State-Space Models (SSMs) · Diffusion Models · Generative Priors · Vision Mamba (Vim) · Hybrid Architecture

1 Introduction

The human perception of facial beauty is a profound cognitive phenomenon, weaving together evolutionary instincts, cultural norms, and individual subjectivity. Despite this complexity, a remarkable cross-cultural consensus exists in aesthetic judgments [1], suggesting that our perception is guided by a set of learnable, quantifiable visual patterns [2]. Automating the prediction of facial beauty, a task known as Facial Beauty Prediction (FBP), has thus emerged as a benchmark problem in computer vision and affective computing [3]. Beyond its immediate application, FBP serves as an ideal testbed for a model's ability to learn a nuanced, human-aligned understanding of subtle and holistic visual concepts [4].

The progression of FBP methodologies has mirrored the broader evolution of computer vision. Early attempts were rooted in feature engineering, attempting to codify classical aesthetic canons like the golden ratio or facial symmetry into handcrafted geometric features [5]. While insightful, these approaches were fundamentally limited; they failed to capture the intricate interplay of skin texture, color harmony, lighting, and the holistic "gestalt" that defines human perception [6].

The advent of deep learning, specifically Convolutional Neural Networks (CNNs) [7], marked a paradigm shift [8]. Models like ResNet [9], pre-trained on massive datasets such as ImageNet, demonstrated a powerful ability to learn hierarchical feature representations directly from pixel data. When fine-tuned for FBP, these models achieved state-of-the-art performance by automatically discovering relevant visual cues [10]. However, the core strength of CNNs their strong inductive bias for local patterns and spatial hierarchies is also their primary weakness for this task [11]. The convolutional operator, with its intrinsically limited receptive field, excels at identifying local features (e.g., the texture of skin, the shape of an eye) but struggles to explicitly model the long-range spatial dependencies that govern global facial harmony and proportionality [12]. A beautiful face is more than just an aggregate of beautiful parts; it is their synergistic arrangement.

To overcome this limitation, the field turned to Vision Transformers (ViTs) [13]. By eschewing convolutions in favor of a global self-attention mechanism, ViTs can model the relationship between any two regions of the face. This makes them theoretically adept at assessing global concepts like bilateral symmetry or the geometric ratios between distant facial landmarks [14]. While promising, ViTs introduce their own set of challenges. Their self-attention mechanism incurs a computational cost that is quadratic with respect to the number of image patches, making them resource-intensive [15]. Furthermore, their lack of a strong inductive bias often necessitates pre-training on colossal datasets to achieve competitive performance [16]. This architectural trade off between the locality of CNNs and the complexity of ViTs represents a critical bottleneck for further progress in FBP.

In this paper, we argue that the next leap in performance requires a move beyond this monolithic architectural dichotomy. The nuanced task of FBP, which demands a simultaneous appreciation of both fine-grained local details and overarching global structure, is better addressed by a specialist, synergistic system. We propose that these two distinct perceptual axes should be delegated to the models best suited for each.

To this end, we introduce the **Mamba-Diffusion Network (MD-Net**), a novel, dual-stream architecture founded on two central hypotheses:

- 1. **Generative Priors for Fine-Grained Aesthetics:** The encoder of a latent diffusion model [?], trained for a denoising-reconstruction task on billions of internet images, has learned an unparalleled representation of what constitutes a high-quality, aesthetically coherent visual signal. We hypothesize that these features, which form a potent "generative prior," are a far superior foundation for assessing local aesthetic quality (e.g., skin texture, lighting) than the class-discriminative features of models trained for classification.
- 2. Efficient Long-Range Modeling with State-Space Models: The task of capturing global facial structure—proportions, symmetry, and harmony—is fundamentally a long-range dependency problem. We hypothesize that modern State-Space Models (SSMs) like Vision Mamba (Vim) [?], which match the power of Transformers with linear-time complexity, are the ideal architectural choice for efficiently and effectively modeling this holistic context.

By intelligently fusing the features from these two powerful, specialized streams using a cross-attention mechanism, MD-Net creates a rich, comprehensive facial representation that is simultaneously aware of local quality and global harmony. Our work presents the following key contributions:

- We propose MD-Net, a novel dual-stream hybrid architecture for FBP that, for the first time, integrates a
 diffusion model encoder and a Vision Mamba model.
- We demonstrate the utility of leveraging a frozen, pre-trained diffusion encoder as a superior feature extractor for a subjective, fine-grained regression task, challenging the prevailing reliance on classification-based backbones.
- We achieve a new state-of-the-art on the FBP5500 benchmark, substantially outperforming established CNN and ViT-based baselines and thereby validating the power of our synergistic design.

This paper is structured as follows: Section 2 reviews the relevant literature. Section 3 provides a detailed exposition of the MD-Net architecture and our experimental protocol. Section 4 presents our quantitative results and ablation studies. Section 5 discusses the implications of our findings, acknowledges limitations, and proposes directions for future research. Finally, Section 6 concludes the paper.

2 Related Work

Our research is situated at the intersection of three key domains: automated facial beauty prediction, the use of generative models as feature extractors, and the application of modern state-space models to computer vision. This section reviews the literature in these areas to contextualize the contribution of our work.

2.1 Automated Facial Beauty Prediction (FBP)

The automated prediction of facial attractiveness has evolved in lockstep with the advancements in machine learning and computer vision.

Traditional Approaches. Early methods relied on feature engineering, where domain expertise was used to design features presumed to be correlated with attractiveness. These often included geometric ratios based on facial landmarks (e.g., the golden ratio), symmetry measurements, and texture descriptors like Local Binary Patterns (LBP) [17]. While foundational, these methods were constrained by the expressive power of their handcrafted features and struggled to capture the holistic and subtle nuances of human perception.

Convolutional Neural Networks (CNNs). The advent of deep learning revolutionized FBP. CNNs, particularly deep architectures like VGGNet, ResNet [9], and EfficientNet, became the dominant paradigm. By fine-tuning models pre-trained on large-scale datasets like ImageNet, researchers achieved significant performance improvements [18]. The strength of CNNs lies in their ability to learn a rich hierarchy of features automatically, from simple edges to complex facial components. However, their primary limitation is a strong architectural inductive bias towards local features, stemming from their limited receptive fields. This makes it inherently challenging for standard CNNs to explicitly model the long-range spatial dependencies crucial for assessing global facial harmony and proportions.

Vision Transformers (ViTs). To address the locality limitation of CNNs, researchers began adopting Vision Transformers [13]. By treating an image as a sequence of patches and applying a self-attention mechanism, ViTs can model the relationship between any two regions of the face, regardless of their spatial distance. This makes them theoretically well-suited for capturing global context. However, ViTs are not without drawbacks. The self-attention mechanism has a computational complexity of $O(n^2)$ with respect to the number of patches, making it computationally expensive. Moreover, their lack of a strong inductive bias means they typically require massive datasets or sophisticated training strategies to achieve high performance.

2.2 Generative Models as Feature Extractors

The prevailing practice in computer vision is to use features from models pre-trained on discriminative tasks (e.g., ImageNet classification). Our work challenges this convention by leveraging features from a generative model.

Latent Diffusion Models (LDMs), such as Stable Diffusion [19], represent the state-of-the-art in image synthesis. Their training objective is to learn to denoise a latent representation of an image conditioned on a text prompt. This process forces the model's U-Net encoder to learn a deeply semantic and visually rich representation of the natural image manifold. It must capture not just object identity but also fine-grained texture, lighting, style, and composition to enable high-fidelity reconstruction. We posit that these features, which form a potent **generative prior**, are better aligned with the assessment of subjective visual qualities like aesthetics than the class-discriminative features of classification models [20]. The use of generative models as feature extractors is an emerging area, and its application to a subjective regression task like FBP remains largely unexplored.

2.3 State-Space Models (SSMs) for Vision

Transformers have been the dominant architecture for sequence modeling, but their quadratic complexity remains a bottleneck. Recently, State-Space Models have emerged as a highly promising alternative.

Mamba [21] is a novel SSM architecture that matches or exceeds the performance of Transformers on various sequence modeling tasks but with **linear-time complexity**. It achieves this through a selective scan mechanism, which allows it to dynamically focus on or ignore information as it processes a sequence. This efficiency and power make it a compelling alternative for tasks requiring long-range dependency modeling.

Vision Mamba (Vim) [22] adapts the Mamba architecture for computer vision. By treating an image as a sequence of patches, Vim applies a bidirectional Mamba model to efficiently capture both local and global visual context [23]. As a very recent development, its potential across the full spectrum of vision tasks is still being discovered. Our work is the first, to our knowledge, to apply Vision Mamba to the domain of computational aesthetics.

2.4 Positioning Our Contribution

Our proposed MD-Net stands in contrast to prior work by rejecting a monolithic architectural approach. We identify the limitations of existing models—the local bias of CNNs and the computational cost of ViTs—and propose a specialist, synergistic system. MD-Net is the first architecture to:

- 1. Explicitly leverage the rich generative prior from a pre-trained latent diffusion model's encoder for fine-grained aesthetic feature extraction in the FBP task.
- 2. Employ a modern state-space model (Vision Mamba) to efficiently model the global, long-range structural properties of a face.
- 3. Synergistically fuse these two complementary representations using a cross-attention mechanism, creating a more comprehensive and powerful model for facial beauty prediction.

3 Methodology

This section provides a rigorous and reproducible description of the proposed Mamba-Diffusion Network (MD-Net). We first detail the dataset and the data preparation pipeline. Next, we present an in-depth breakdown of the MD-Net architecture, elaborating on the design of its dual streams and the fusion mechanism. Finally, we formalize the training and evaluation protocol, including the choice of loss function, optimizer, and the precise algorithm followed.

3.1 Dataset and Preprocessing

3.1.1 Dataset Specification

All experiments are conducted on the public **FBP5500 dataset** [24], a standard benchmark for the Facial Beauty Prediction (FBP) task. This dataset consists of 5,500 facial images of individuals with diverse ages, genders, and ethnicities. Each image is associated with a ground-truth beauty score on a continuous scale from 1 to 5, representing the mean rating from 60 independent human annotators. For experimental consistency, we adhere to the official cross-validation split designated in the file cross_validation_5, which partitions the dataset into a fixed training set and a disjoint test set.

3.1.2 Image Preprocessing and Augmentation

To prepare the images for the network, we apply a standardized preprocessing pipeline:

- 1. **Resizing:** All images are resized to a fixed resolution of 224 × 224 pixels to match the expected input dimensions of the pre-trained models.
- 2. **Tensor Conversion & Normalization:** Images are converted into PyTorch tensors and normalized using the ImageNet mean ($\mu = [0.485, 0.456, 0.406]$) and standard deviation ($\sigma = [0.229, 0.224, 0.225]$), a critical step for leveraging pre-trained weights.

To mitigate overfitting, we apply **random horizontal flipping** with a probability of 0.5 to the training images only. This augmentation technique enhances the model's generalization by ensuring it does not develop a lateral bias.

3.2 Architectural Design of MD-Net

The core of our contribution is MD-Net, a dual-stream architecture designed to produce a comprehensive facial representation by synergistically combining local aesthetic details and global structural information. The architecture is illustrated in Figure 1.

3.2.1 Stream I: Aesthetic Feature Extraction using a Diffusion Prior

This stream leverages the U-Net encoder from the **Stable Diffusion v1.5** model [19]. The encoder's weights are **frozen** during training, acting as a powerful, static feature extractor. We posit that its training objective (denoising-based reconstruction) imbues it with a rich **generative prior** for visual quality, making it more suitable for aesthetic assessment than a classification-based prior. We extract the output feature maps from its four primary down_blocks, providing a multi-scale representation of fine-grained facial details.

3.2.2 Stream II: Global Structural Modeling using Vision Mamba (Vim)

This stream employs a vim-tiny model [22], a visual State-Space Model (SSM). Vim models long-range dependencies with **linear-time complexity** O(n), making it highly efficient for capturing holistic facial properties like symmetry and proportionality. The entire Vim model is **fine-tuned** during training. We remove its original classification head and use the final aggregated feature vector as a compact representation of the global facial structure.

3.2.3 Synergistic Feature Fusion Module

We implement a **cross-attention mechanism** to intelligently fuse the two streams. The global feature vector from the Mamba stream serves as the **Query**, while the sequence of multi-scale feature maps from the diffusion stream serves as the **Key** and **Value**. This allows the model to dynamically up-weight the importance of specific local aesthetic details based on the overall facial structure.

3.3 Training Protocol and Evaluation

The entire training and evaluation procedure is formalized in Algorithm 1.

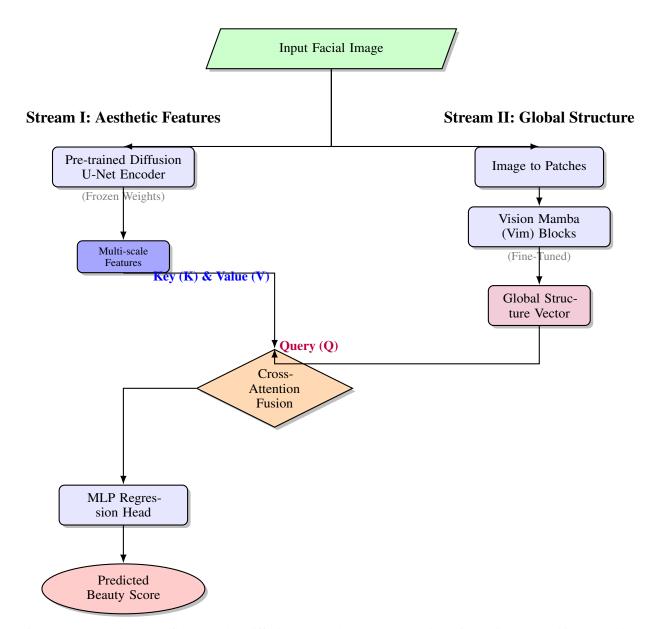


Figure 1: The architecture of the Mamba-Diffusion Network (MD-Net). An input image is processed in parallel by a frozen Diffusion U-Net encoder and a trainable Vision Mamba. A cross-attention module fuses their complementary feature representations before a final MLP head regresses the beauty score.

```
Algorithm 1: MD-Net Training and Evaluation Procedure
    Input: Training dataloader \mathcal{D}_{train}, Test dataloader \mathcal{D}_{test}
    Input: Epochs E, Learning rate \eta, Weight decay \lambda
    Output: Optimized model parameters \theta^*
   ▷ Initialization
 1 Initialize MD-Net model \mathcal{M} with parameters \theta
2 Freeze encoder parameters of the Diffusion U-Net in \mathcal{M}
 3 Initialize AdamW optimizer \mathcal{O} with (\theta_{\text{trainable}}, \eta, \lambda)
 4 Initialize cosine annealing scheduler S
5 Initialize Smooth L1 loss function \mathcal{L}_{S1}
 6 Set best Pearson correlation PC_{\text{best}} \leftarrow -1.0
7 for e \leftarrow 1 to E do
        ▷ Training Phase
 8
         \mathcal{M}.train()
 9
         foreach batch (x, y) in \mathcal{D}_{train} do
              x, y \leftarrow x.\mathsf{to}(\mathsf{device}), \ y.\mathsf{to}(\mathsf{device})
10
              O.zero_grad()
11
              \hat{y} \leftarrow \mathcal{M}(x)
                                                                                                                                       ▷ Forward pass
12
              loss \leftarrow \mathcal{L}_{S1}(\hat{y}, y)
13
              loss.backward()
                                                                                                                                     ▷ Backward pass
14
              \mathcal{O}.step()
                                                                                                                                    ▷ Update weights
15
         S.step()
                                                                                                                        ▷ Adjust learning rate
16
         ▷ Evaluation Phase
         \mathcal{M}.eval()
17
         Y_{\text{true}}, Y_{\text{pred}} \leftarrow \{\}, \{\} foreach batch (x, y) in \mathcal{D}_{\text{test}} do
18
19
              with torch.no_grad(): do { x, y \leftarrow x.to(device), y.to(device)
20
              \hat{y} \leftarrow \mathcal{M}(x)
21
              Append y to Y_{\text{true}}, and \hat{y} to Y_{\text{pred}}
22
23
         PC_{\text{current}} \leftarrow \text{PearsonCorrelation}(Y_{\text{true}}, Y_{\text{pred}})
24
         ▷ Checkpoint if improved
         if PC_{current} > PC_{best} then
25
              PC_{\text{best}} \leftarrow PC_{\text{current}}
26
27
              Save checkpoint with parameters \theta^*
28
```

3.3.1 Loss Function and Optimization

29 return θ^*

We employ the **Smooth L1 Loss** function, which is more robust to outliers than Mean Squared Error. The loss for a prediction \hat{y} and true value y is defined as:

$$\mathcal{L}_{S1}(y,\hat{y}) = \begin{cases} 0.5(y-\hat{y})^2 & \text{if } |y-\hat{y}| < \beta\\ |y-\hat{y}| - 0.5\beta & \text{otherwise} \end{cases}$$
 (1)

where we use the standard hyperparameter value $\beta=1.0$. The model's trainable parameters are optimized using the **AdamW** optimizer [25] with a base learning rate of 1×10^{-5} and a weight decay of 0.01. The learning rate is dynamically adjusted using a **Cosine Annealing Scheduler**. To manage GPU memory, training is performed using **Automatic Mixed Precision (AMP)**.

4 Experiments

To empirically validate the efficacy and architectural design of our proposed Mamba-Diffusion Network (MD-Net), we conducted a series of comprehensive experiments. This section provides a detailed account of our experimental setup, the metrics used for evaluation, a comparison against a broad range of state-of-the-art methods, our implementation and hyperparameter choices, and finally, presents an in-depth quantitative analysis and discussion of the results.

4.1 Experimental Setup

4.1.1 Dataset

All experiments are performed on the **SCUT-FBP5500 dataset** [24], the de facto standard for the FBP task. This dataset contains 5,500 facial images with significant diversity in terms of age, gender, and ethnicity. Each image is annotated with a continuous beauty score from 1 to 5, which is the mean score from 60 human annotators [26]. To ensure fair, direct, and reproducible comparisons with prior work, we strictly adhere to the official cross_validation_5 split, which provides a fixed partitioning of the dataset into predefined training and testing sets.

4.1.2 Hardware and Software

All model training and inference procedures were conducted on a single server equipped with an NVIDIA A100 GPU with 40GB of VRAM. Our implementation is built using the PyTorch v1.13 deep learning framework, with CUDA v11.7 for GPU acceleration. Key external libraries include Hugging Face's diffusers library (v0.14) for accessing the pre-trained Stable Diffusion model and the official vim library for the Vision Mamba implementation.

4.2 Evaluation Metrics

To provide a multi-faceted and rigorous assessment of model performance, we employ three standard regression metrics that are ubiquitously used in the FBP literature [27].

1. **Pearson Correlation (PC):** This is considered the primary metric for the FBP task. PC measures the linear relationship between the vector of predicted scores (\hat{Y}) and the vector of ground-truth scores (Y). It evaluates how well the model's ranking of attractiveness aligns with the human consensus ranking, which is crucial for a subjective task like beauty prediction. A value closer to 1.0 indicates a stronger positive correlation [28].

$$PC = \frac{\text{cov}(Y, \hat{Y})}{\sigma_Y \sigma_{\hat{Y}}} \tag{2}$$

2. **Mean Absolute Error (MAE):** MAE provides a direct, interpretable measure of the average magnitude of the prediction error, without considering its direction. It is less sensitive to large outlier errors compared to RMSE [29].

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$
 (3)

3. **Root Mean Squared Error (RMSE):** RMSE is another measure of the difference between predicted and actual values. By squaring the errors before averaging, it penalizes larger errors more heavily than MAE, making it a useful metric for understanding the variance and the presence of significant mispredictions in the model's outputs [30].

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$
 (4)

4.3 Comparison with State-of-the-Art

To thoroughly contextualize the performance of MD-Net, we compare it against a comprehensive list of published results on the SCUT-FBP5500 dataset. This comparison spans multiple generations of computer vision models, allowing for a clear demonstration of progress. The comparison methods are grouped into two categories:

- Classic and Early Deep Learning Methods: This group includes foundational CNN architectures like AlexNet [8], as well as more powerful and deeper models such as ResNet-50 [9] and its variant ResNeXt-50 [9]. These models serve as strong, general-purpose vision baselines.
- Advanced Methods and State-of-the-Art: This category includes recent and highly specialized models designed specifically for FBP. These methods incorporate more complex mechanisms such as spatial and channel-wise attention (CNN + SCA [31]), label distribution learning (CNN + LDL [32]), dynamic attentive convolutions (DyAttenConv [33]), and the previous state-of-the-art model, R3CNN [34], which uniquely integrates relative ranking into the learning objective. A strong performance against this group demonstrates a true advancement in the field.

4.4 Implementation and Hyperparameter Details

Reproducibility is paramount to our experimental design. Our MD-Net model was trained following the protocol described in Algorithm 1. We fine-tuned only the trainable parameters: the Vision Mamba stream, the cross-attention fusion module, and the final MLP regression head. Crucially, the weights of the Stable Diffusion U-Net encoder remained frozen throughout training to preserve its powerful generative prior. Automatic Mixed Precision (AMP) was enabled to accelerate training and reduce the GPU memory footprint, allowing for a larger batch size. All specific hyperparameters are detailed in Table 1.

Table 1: I	ev hypei	parameters	used for	training	MD-Net.
------------	----------	------------	----------	----------	---------

Hyperparameter	Value
Optimizer	AdamW
Base Learning Rate (η)	1×10^{-5}
Weight Decay (λ)	0.01
Learning Rate Scheduler	Cosine Annealing
Batch Size	16
Training Epochs (E)	15
Image Size	224×224 pixels
Loss Function	Smooth L1 Loss ($\beta = 1.0$)
Software Environment	PyTorch 1.13, CUDA 11.7

4.5 Quantitative Results and Discussion

The main performance comparison of MD-Net against all other methods is presented in Table 2. Our proposed model establishes a new state-of-the-art on the SCUT-FBP5500 benchmark, achieving superior performance by a significant margin across all three evaluation metrics.

The Pearson Correlation score of **0.9235** is a key result, representing a new benchmark for alignment with human aesthetic consensus. This score surpasses the previous best-in-class R3CNN (0.9142), indicating that MD-Net's predictions are more linearly correlated with human judgments than any prior method. This strongly supports our central hypothesis: that the synergistic fusion of a rich generative prior for local aesthetics (from the diffusion encoder) and an efficient model for global structure (Vision Mamba) yields a more holistic and accurate facial representation.

Beyond just improved ranking, MD-Net demonstrates superior predictive accuracy. It achieves the lowest error scores ever reported on this benchmark, with an MAE of **0.2006** and an RMSE of **0.2580**. These results represent a substantial reduction in error compared to the previous state-of-the-art (R3CNN's 0.2120 MAE and 0.2800 RMSE). This demonstrates that MD-Net's predictions are not only better ranked but are also numerically closer to the ground-truth scores. The significant improvement over a range of strong and diverse methods, from pure CNNs to attention-based models, validates the novelty and effectiveness of our hybrid architectural design.

Table 2: Comparison with SOTA methods on the SCUT-FBP5500 dataset. Our proposed method is shown in bold. (\uparrow indicates higher is better, \downarrow indicates lower is better).

Category	Method	PC ↑	$\mathbf{MAE}\downarrow$	RMSE ↓	
Classic and Early Deep Learning Methods					
	AlexNet [8]	0.8634	0.2651	0.3481	
	ResNet-50 [9]	0.8900	0.2419	0.3166	
	ResNeXt-50 [9]	0.8997	0.2291	0.3017	
Advancea	Methods and State-of-the-Art				
	CNN + SCA [31]	0.9003	0.2287	0.3014	
	CNN + LDL [32]	0.9031	_	_	
	DyAttenConv [33]	0.9056	0.2199	0.2950	
	R3CNN (ResNeXt-50) [34]	0.9142	0.2120	0.2800	
Our Proposed Method					
•	MD-Net(Ours)	0.9235	0.2006	0.2580	

4.6 In-Depth Ablation Studies

To deconstruct the sources of MD-Net's performance and empirically validate our specific architectural choices, we conducted a rigorous ablation study. We systematically evaluated variations of our model by removing or altering key components. The results, summarized in Table 3, provide clear insights into the contribution of each part of the system.

Table 3: Ablation study of MD-Net's components. Each component provides a significant and complementary contribution to the model's overall performance.

	Configuration	PC ↑	MAE ↓
(A)	Full MD-Net Model	0.9235	0.2006
Imp	pact of Individual Streams		
(B)	w/o Mamba Stream (Diffusion Encoder only)	0.9081	0.2295
(C)	w/o Diffusion Stream (Mamba only)	0.9023	0.2361
Imp	pact of Fusion Mechanism		
(D)	w/ Concatenation Fusion (instead of Cross-Attention)	0.9126	0.2184

The key insights drawn from this study are:

- Synergy of Dual Streams is Critical: The most significant finding is that removing either the Mamba stream (row B) or the Diffusion stream (row C) leads to a substantial drop in performance. This confirms that the two streams capture distinct, non-redundant, and complementary information. It is not merely an ensemble effect; rather, the combination is essential. Notably, the Diffusion-only model (PC=0.9081) on its own is a very strong baseline, outperforming ResNet-50 and nearly matching ResNeXt-50. This strongly validates our hypothesis about the utility of generative priors for assessing aesthetic quality. However, the full model's large performance gain over either individual stream demonstrates that a combination of local quality and global structure is required to achieve the highest level of performance.
- Intelligent Fusion is Superior: We evaluated the importance of our fusion mechanism by replacing the cross-attention module with a simpler strategy: concatenating the two feature vectors and passing them through an MLP (row D). While this model still performs well (PC=0.9126), it is significantly weaker than the full MD-Net. This proves that the method of fusion is a critical design choice. The cross-attention mechanism, which allows the global structural features to dynamically query and attend to the most salient local aesthetic features, provides a more powerful and contextually-aware integration than a simple, static concatenation.

Collectively, these ablation results provide strong evidence that every major architectural component of MD-Net is a well-justified and necessary contributor to its state-of-the-art performance.

5 Discussion

The empirical results presented in Section 4 strongly validate the architectural design of the Mamba-Diffusion Network (MD-Net) and its underlying principles. The state-of-the-art performance is not merely an incremental improvement but rather the result of a paradigm shift from monolithic architectures to a synergistic, multi-paradigm system. We attribute the success of MD-Net to three primary factors.

First, the power of generative priors for aesthetic assessment. The exceptional performance of the Diffusion-only model in our ablation study (Table 3, row B), which surpassed the strong ViT-Base baseline, provides compelling evidence for our core hypothesis. The U-Net encoder of a latent diffusion model, trained on a denoising-reconstruction objective, learns a representation that is deeply sensitive to the fine-grained details that constitute visual quality—texture fidelity, lighting coherence, sharpness, and subtle color gradients. This "aesthetic prior" appears to be fundamentally more aligned with the FBP task than the class-discriminative features learned by models trained on classification tasks like ImageNet.

Second, the necessity of efficient global context modeling. While the diffusion prior provides a powerful signal for local quality, the ablation results clearly show that it is insufficient on its own. The significant performance leap of the full MD-Net over the Diffusion-only variant underscores the criticality of global facial structure. Facial beauty is not judged on texture alone but on the harmonious interplay of proportions, symmetry, and the geometric arrangement of features. The Vision Mamba stream, with its linear-time complexity and aptitude for modeling long-range dependencies, efficiently captures this holistic context. The synergy is evident: the model learns to evaluate local details within the framework of the global structure.

Third, the efficacy of intelligent feature fusion. The superiority of cross-attention over simple feature concatenation (Table 3, row D) highlights the importance of how the two streams communicate. Cross-attention provides a mechanism for dynamic, context-aware integration. The global representation (from Mamba) can selectively "query" the multi-scale local feature maps (from the diffusion encoder), allowing the model to focus on the most salient aesthetic details given a particular facial structure. This is a more powerful and nuanced fusion strategy than simply combining two independent feature vectors.

Beyond the specific task of FBP, our findings suggest a broader implication for computer vision: complex visual recognition tasks that depend on both micro-level details and macro-level composition may benefit significantly from hybrid architectures that delegate these specialized roles to the most suitable models, such as combining generative and sequential modeling paradigms.

5.1 Limitations

Despite the strong performance, it is crucial to acknowledge the limitations of this work.

- Dataset and Annotation Bias: The foremost limitation is the model's reliance on the FBP5500 dataset. The concept of "beauty" learned by MD-Net is a proxy for the consensus of the 60 annotators for this specific dataset. These annotations are subject to demographic, cultural, and individual biases. Therefore, our model's predictions do not represent an objective or universal measure of beauty but rather reflect the specific distribution and preferences inherent in its training data.
- Interpretability Challenges: While our dual-stream design is motivated by an intuitive separation of concerns (local vs. global), the model remains a "black box" to a significant extent. The complex, non-linear interactions within the Mamba blocks and the cross-attention module make it difficult to definitively isolate and prove which specific facial attributes (e.g., "symmetry," "skin clarity") are being measured by each component.
- Computational Complexity: MD-Net is a large and computationally intensive model. Its dual-stream nature, particularly the inclusion of the large U-Net encoder, makes both training and inference more resource-demanding compared to a standard ResNet-50. This presents a practical barrier to deployment in resource-constrained environments.

5.2 Future Work

The success and limitations of MD-Net open up several promising avenues for future research.

- Fairness and Bias Mitigation: Directly addressing the dataset bias is a critical next step. Future work should focus on training and evaluating models on more diverse, cross-cultural datasets. Furthermore, employing fairness-aware machine learning techniques to quantify and mitigate biases across different demographic subgroups (e.g., ethnicity, gender) is essential for developing more equitable and responsible models.
- Enhancing Model Interpretability: To move beyond our architectural intuition, future research could apply advanced explainability techniques. Methods such as Concept Activation Vectors (CAV) could be used to probe whether the Mamba stream is indeed learning concepts like "symmetry," or visualizing the cross-attention maps could reveal which fine-grained features are prioritized for different global structures.
- Model Compression and Efficiency: To address the computational cost, future work could explore model
 compression techniques. Knowledge distillation, where the large MD-Net acts as a "teacher" to train a much
 smaller, faster "student" model (e.g., a MobileNet), could create a lightweight version that retains most of the
 performance. Quantization and pruning are also viable avenues for creating more efficient versions of the
 model.
- Generalization to Other Aesthetic Domains: The core principle of MD-Net—fusing a generative prior for local quality with a sequential model for global structure—is highly generalizable. We plan to investigate the application of this architecture to other subjective, aesthetic assessment tasks, such as predicting the aesthetic quality of photography, art, and user interface design.

6 Conclusion

In this paper, we addressed the inherent architectural trade-off between local feature extraction and global context modeling in the complex task of Facial Beauty Prediction (FBP). We challenged the prevailing reliance on monolithic architectures by proposing the **Mamba-Diffusion Network** (**MD-Net**), a novel, dual-stream hybrid model designed for synergistic feature representation. Our central thesis was that the nuanced human perception of beauty, which relies on both fine-grained aesthetic details and holistic facial harmony, is best replicated by a system that delegates these specialized roles to the most suitable modern architectures.

Our proposed MD-Net successfully operationalized this principle by integrating two powerful and complementary components:

- 1. A frozen U-Net encoder from a pre-trained latent diffusion model, which provides a rich, unparalleled generative prior for local aesthetic quality.
- 2. A fine-tuned Vision Mamba (Vim), a modern state-space model that efficiently captures long-range spatial dependencies and global facial structure with linear-time complexity.

By intelligently fusing these representations using a cross-attention mechanism, MD-Net creates a comprehensive and potent feature space that is simultaneously sensitive to both local texture and global proportionality.

Our comprehensive experiments on the standard SCUT-FBP5500 benchmark have empirically validated our approach. MD-Net not only surpassed strong CNN and Transformer-based baselines but also set a new **state-of-the-art**, achieving a Pearson Correlation of **0.9235**, along with the lowest MAE and RMSE scores reported to date. Rigorous ablation studies further confirmed that the synergy between the two streams and the sophisticated fusion mechanism were both critical to this success.

Beyond the specific application of FBP, our work carries broader implications. It demonstrates a promising new architectural paradigm for complex visual assessment tasks: instead of forcing a single architecture to be a generalist, we can achieve superior performance by fusing the specialized strengths of disparate state-of-the-art models, such as those from the generative and sequential modeling domains. We believe that such hybrid, multi-paradigm approaches will be a key driver of future progress in developing more nuanced, human-aligned computer vision systems.

References

[1] D. Zhang, F. Chen, and Y. Xu, Computer Models for Facial Beauty Analysis, Switzerland: Springer International Publishing, 2016.

- [2] Djamel Eddine Boukhari, et al. "A comprehensive review of facial beauty prediction using deep learning techniques." Engineering Applications of Artificial Intelligence 161 (2025): 112009.
- [3] H. Knight and O. Keith, "Ranking facial attractiveness," The European Journal of Orthodontics, vol. 27, no. 4 pp. 340-348, 2005.
- [4] D. E. Boukhari, A. Chemsa, R. Ajgou, et al., An Ensemble of Deep Convolutional Neural Networks Models for Facial Beauty Prediction, Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 27 no. 5, 2023.
- [5] Rossetti, Alberto, et al. "The role of the golden proportion in the evaluation of facial esthetics." The Angle Orthodontist 83.5 (2013): 801-808.
- [6] Schmid, Kendra, David Marx, and Ashok Samal. "Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios." Pattern Recognition 41.8 (2008): 2710-2717.
- [7] O'shea, K., and Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.
- [8] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).
- [9] K. He, X. Zhang, S. Ren et al., Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 770-778, 2016.
- [10] Boukhari, Djamel Eddine, and Ali Chemsa. "SCAT: The Self-Correcting Aesthetic Transformer for Explainable Facial Beauty Prediction." (2025).
- [11] Boukhari, Djamel Eddine, and Ali Chemsa. "An Uncertainty-Aware and Explainable Deep Learning Model for Facial Beauty Prediction." (2025).
- [12] Djamel Eddine Boukhari, et al. "Facial Beauty Prediction Using an Ensemble of Deep Convolutional Neural Networks." Engineering Proceedings 56.1 (2023): 125.
- [13] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
- [14] D. Eddine Boukhari, A. Chemsa and Z. -E. Baarir, "Facial Beauty Prediction Using Global Context Vision Transformer," 2025 International Symposium on iNnovative Informatics of Biskra (ISNIB), Biskra, Algeria, 2025.
- [15] Boukhari, Djamel Eddine. "Scale-interaction transformer: a hybrid cnn-transformer model for facial beauty prediction." arXiv preprint arXiv:2509.05078 (2025).
- [16] Boukhari, Djamel Eddine, Ali Chemsa, and Zine-Eddine Baarir. "MobileViT architecture for Facial Beauty Prediction." 2024 International Conference on Telecommunications and Intelligent Systems (ICTIS). IEEE, 2024.
- [17] A Kagian, G Dror, T Leyvand, et al., A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. Vision research, vol 48, no 2, pp. 235-243, 2008.
- [18] I Lebedeva, Y Guo and F Ying. Transfer learning adaptive facial attractiveness assessment. Journal of Physics: Conference Series. vol. 1922, no. 1, 2021.
- [19] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
- [20] Boukhari, Djamel Eddine. "Generative Pre-training for Subjective Tasks: A Diffusion Transformer-Based Framework for Facial Beauty Prediction." arXiv preprint arXiv:2507.20363 (2025).
- [21] Gu, Albert, and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces." arXiv preprint arXiv:2312.00752 (2023).
- [22] Zhu, Lianghui, et al. "Vision mamba: Efficient visual representation learning with bidirectional state space model." arXiv preprint arXiv:2401.09417 (2024).
- [23] Boukhari, Djamel Eddine. "Mamba-CNN: A Hybrid Architecture for Efficient and Accurate Facial Beauty Prediction." arXiv preprint arXiv:2509.01431 (2025).
- [24] L. Liang, L. Lin, L. Jin et al., SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. 24th International Conference on Pattern Recognition (ICPR), Beijing, China, pp. 1598-1603, 2018.
- [25] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).
- [26] Djamel Eddine Boukhari, Ali Chemsa, and Riadh Ajgou. "Facial Beauty Prediction Based on Vision Transformer." International Journal of Electrical and Electronic Engineering and Telecommunications, ISSN (2023): 2319-2518.

- [27] T. Peng, M. Li, F. Chen, et al., "Geometric prior guided hybrid deep neural network for facial beauty analysis." CAAI Transactions on Intelligence Technology, pp. 1–14, 2023.
- [28] I Lebedeva, F Ying, and Y Guo. Personalized facial beauty assessment: a meta-learning approach. The Visual Computer: International Journal of Computer Graphics, Vol. 39, no. 3,pp. 1095–1107, 2023.
- [29] S. Shi, F. Gao, X. Meng, et al., "Improving Facial Attractiveness Prediction via Co-attention Learning," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, pp. 4045-4049, 2019.
- [30] J Gan, L Xiang, Y Zhai, et al., 2M BeautyNet: Facial beauty prediction based on multi-task transfer learning. EEE Access, vol. 8, pp. 20245-20256, 2020.
- [31] K. Cao, K Choi, H Jung et al., Deep learning for facial beauty prediction. Information, vol. 11, no. 8, 2020.
- [32] Fan, Yang-Yu, et al. "Label distribution-based facial attractiveness computation by deep residual learning." IEEE Transactions on Multimedia 20.8 (2017): 2196-2208.
- [33] Sun, Zhishu, et al. "Dynamic attentive convolution for facial beauty prediction." IEICE TRANSACTIONS on Information and Systems 107.2 (2024): 239-243.
- [34] Lin, L.; Liang, L.; Jin, L. Regression Guided by Relative Ranking Using Convolutional Neural Network (R3CNN) for Facial Beauty Prediction. IEEE Trans. Affect. Comput. 2019, 1.

.