MVBeautyFusion:A Continuous-view Fusion Framework for Face Structure Beautification

Zijin Guo,Suya Li,Zhenping Xie*
School of Artificial Intelligence and Computer Science, Jiangnan University
Jiangsu,Wuxi,China
6233112015@stu.jiangnan.edu.cn, 7233115010@stu.jiangnan.edu.cn, xiezp@jiangnan.edu.cn

Abstract

Facial beautification has gained increasing attention with the advancement of generative models and visual editing techniques. However, existing methods, while effective for static images, often fail to maintain structural consistency and temporal smoothness in dynamic or multi-view scenarios, leading to noticeable artifacts and geometric distortion. To address these challenges, we propose MVBeautyFusion, a continuous-view fusion framework for face structure beautification. The first stage employs a Stable Diffusion–based generator enhanced with ControlNet and structural masks to achieve high-fidelity enhancement while preserving facial geometry. The second stage introduces a lightweight multi-view feature fusion network that enforces cross-frame structural alignment through deformable sampling and content-aware temporal attention. Extensive experiments demonstrate that MVBeautyFusion outperforms state-of-the-art methods such as BFVR and TokenFlow in both structural stability and visual continuity, producing realistic and temporally consistent beautification results with minimal computational cost.

Keywords: Face Beautification, Multi-view Feature fusion, Diffusion Models, Video Face Enhancement.

1. Introduction

Facial beautification technology, as a key application in the field of digital vision, has been widely adopted across diverse scenarios such as social entertainment, live e-commerce, and virtual/augmented reality (VR/AR). On social media platforms like TikTok and Kuaishou, users leverage real-time beautification features to enhance selfies and short videos, increasing content appeal. In live streaming contexts, hosts rely on beautification software for skin smoothing, facial feature adjustment, and dynamic filter overlays to improve audience engagement. Moreover, high-precision beautification techniques are increasingly required in virtual try-on systems, online education, and medical aesthetic consultations to support personalized image rendering and professional use cases. According to market research, the global beauty technology market is projected to reach USD 79.8 billion by 2025, with growing user demand for natural, personalized, and real-time beautification effects continually driving technological innovation.

In recent years, with the rapid advancement of deep learning—particularly generative models—facial beautification has evolved from traditional filters and enhancement techniques to end-to-end generative approaches. These methods can be roughly categorized into three main directions:First, image-to-image translation-based beautification networks treat the task as a translation problem, using conditional generative adversarial networks (cGANs) [29] for style transfer and facial feature enhancement. BeautyGAN [24] is a representative work, achieving makeup style transfer while preserving facial structure, balancing identity fidelity and visual enhancement. SCGAN [28] further introduces semantic disentanglement to independently model key facial regions (e.g., eyes, lips), improving control and naturalness. These methods often adopt encoder-decoder frameworks with semantic supervision and style encoding for visually pleasing and structure-consistent results.Second, cross-frame consistency-aware methods address the connituous stability required in video and real-time communication scenarios. TokenFlow [42] models consistency as temporal token tracking at the latent level, maintaining both style and detail across frames. FateZero [46] proposes a reference-free generative approach using backward latent modeling to ensure editing consistency. StyleUV [51] maps faces under different poses into a unified UV domain for style fusion, solving style shift and misalignment under viewpoint variation.Last but not least, diffusion-based models with structural guidance have shown great promise in both generation quality and controllability. NNSG-Diffusion [25] incorporates nearest-neighbor visual reference and 3D structural cues to enhance facial aesthetics while preserving identity. DiffFAE [48]

*Corresponding author

combines 3D texture rendering and semantic integration with consistency regularization for improved single-image beautification. ControlNet [56], built upon diffusion models such as Stable Diffusion [40, 34, 58], introduces structural guidance from edge maps, depth, and pose to achieve high structural control and visual consistency, opening new possibilities for editable and realistic facial beautification.

Despite the early success of traditional image beautification methods, their limitations have become increasingly evident with the growing complexity of application scenarios. These shortcomings are primarily reflected in the following three aspects:First, existing approaches show limited performance in maintaining detail fidelity and naturalness. Many GAN-based or image translation methods, such as BeautyGAN [24] and SCGAN [28], are capable of learning style mappings but often produce overly smoothed skin textures or visual artifacts, especially in facial details. Furthermore, GAN models [13] are highly dependent on the training data distribution and tend to generalize poorly to unseen poses or lighting conditions, resulting in instability in real-world applications.Second, most mainstream methods focus on single-frame processing without modeling temporal information, leading to inconsistencies in multi-view or video scenarios. For instance, in cases of head turning, fast movement, or frequent lighting changes, these methods may suffer from misaligned facial features or abrupt changes in editing style, seriously affecting visual smoothness and user trust.Third, some beautification techniques involve high computational costs during inference. Methods that process each frame independently or rely on large generative models, such as TokenFlow [42], face significant trade-offs between inference speed and real-time processing requirements, making them unsuitable for deployment in mobile or live streaming scenarios.

To tackle the aforementioned challenges, inspired by the remarkable performance of diffusion models in image generation and editing, we propose MVBeautyFusion, a novel framework with multi-view feature fusion for continuous smoothing face beautification. This approach follows a two-stage pipeline of structure-constrained generation followed by continuous smoothing feature fusion, enabling stable propagation of facial details and stylization across dynamic sequences, thereby significantly enhancing editing consistency and smoothness under varying viewpoints and motions.In the first stage, each frame is processed using a diffusion model guided by ControlNet [56], combined with a binary skin mask obtained through dual fusion of SAM-based structural segmentation [22] and skin detection in the YCrCb color space [23, 43, 21, 8]. This setup produces base beautified images with accurate facial contours and lighting restoration. By incorporating both Canny edge [4] and monocular depth [38, 39] branches in ControlNet, the model's perception of edge geometry and 3D

Figure 1. The first-stage beautification results generated by the ControlNet-guided diffusion model. Incorporating refined skin masks based on SAM and skin color detection further improves naturalness and structural fidelity.

facial structure is enhanced, resulting in more natural and realistic beautification, as illustrated in Figure 1.In the second stage, to ensure continuous smoothing of appearance across consecutive frames, we introduce a feature-level fusion network that combines a deformable alignment module with a temporal attention mechanism. The temporal attention [45], as an extension of self-attention [47], captures long-range dependencies across time, supporting consistent semantic propagation. The deformable alignment module leverages learnable offsets over standard convolution [9] to accommodate non-rigid facial motion across views. To further refine frame-wise smoothness, a visibility-guided strategy [2, 26] is integrated, adaptively emphasizing salient facial regions under motion and occlusion. Together, these components enable seamless feature fusion across frames, achieving high-level multi-view consistency and stylistic smoothness in the beautification results.

In summary, our contributions are threefold:

- We propose MVBeautyFusion, a framework for multi-view continuous smoothing face beautification. By combining structure-guided diffusion generation with temporally-aware feature fusion, our model significantly improves visual consistency and realism.
- We design a cross-frame appearance consistency module, which leverages deformable feature sampling and temporal attention to enhance the model's ability to align and fuse facial regions across multiple frames and views.
- Extensive experiments on portrait datasets such as FNpic and Facevid demonstrate that our approach outperforms baseline methods in terms of image reconstruction quality and multi-view continuous smoothness.

Overall, MVBeautyFusion offers a unified beautification solution that effectively balances editing quality, continuous

smoothing, and computational efficiency. It is well-suited for a variety of practical applications, including multi-view video processing, facial editing, and special effects rendering. Experimental results demonstrate that our approach significantly outperforms existing methods across multiple real-world datasets, achieving high-quality and controllable portrait beautification without requiring additional training.

2. Related work

2.1. Diffusion Models for Image Generation and Editing

In recent years, diffusion models [40, 35] have emerged as a core approach for image generation, demonstrating outstanding performance particularly in text-to-image (T2I) generation tasks [6, 32, 37]. For example, Imagen [41] achieves high-quality text-driven image synthesis by progressively refining image resolution through a cascaded diffusion structure.To improve inference efficiency and scalability, Stable Diffusion [40] transfers the diffusion process to the latent space, significantly reducing memory and computational cost. While text-prompt-driven image editing enables personalized and semantically guided control, early approaches often lack precise structural guidance, leading to content drift or loss of details.To address this, ControlNet [56] extends Stable Diffusion by incorporating structural condition branches. By introducing controllable priors such as edge maps [4], depth maps [38, 39], and human poses [5], ControlNet enables more accurate editing while preserving the original image structure.Subsequent studies, including T2I-Adapter [30] and Prompt-to-Prompt [17], further enhance local controllability in editing scenarios, supporting structure-preserving synthesis under multimodal input conditions.

2.2. Portrait Editing and Style Consistency Modeling

Portrait image editing, an important subtask in generative modeling, is widely applied in virtual avatars, social media, and short video platforms. Early face editing methods typically relied on GAN-based feature spaces for style transfer and attribute manipulation. For example, StarGAN [7] and AttGAN [16] achieved cross-domain multi-attribute translation and controllable semantic editing, respectively, but struggled to handle pose variations and continuous smoothing in editing scenarios.

To address these limitations, DECA [12] and IMavatar [59] employed 3D facial meshes or implicit representations to build motion-driven facial models, improving multi-view consistency to some extent. However, their complex 3D fitting procedures lead to low training and inference efficiency. More lightweight alternatives like FaceVerse [53] fuse 2D and 3D information but remain limited in style expressiveness.

Recently, diffusion-based methods such as InstantID [49] and PhotoMaker [54] have gained attention by integrating multiple reference facial images and text prompts, enabling local detail control while preserving identity consistency. Additionally, some approaches incorporate control signals into the diffusion process to achieve controllable editing effects, exemplified by DragGAN [31].

Another research direction focuses on modeling continuous smoothing. TokenFlow [42] transforms style preservation along the temporal dimension into temporal token mappings within feature space, enhancing video generation consistency through cross-frame token tracking. Subsequent works like Tune-A-Video [52] and VideoComposer [20] introduce conditional guidance and cross-frame content fusion mechanisms to further improve multi-frame consistency and local reconstruction.

3. Method

3.1. Fusion Mask-Guided Single-View Image Beautification

In the first stage, we propose a refined control mechanism that combines ControlNet [56] with a skin segmentation strategy, as illustrated on the left side of Figure 2. This approach aims to enhance structural preservation and aesthetic consistency in single-view facial image beautification at the basic enhancement phase. The overall pipeline is built upon the Stable Diffusion framework, incorporating multiple ControlNet conditioning signals to guide the generation process, while leveraging facial semantic information to enforce region-specific constraints.

3.1.1 ControlNet Input Design

To better guide Stable Diffusion in preserving facial structure and contours during the beautification process, we introduce two ControlNet [56] conditioning inputs: edge information (Canny) [4] and depth information [38, 39]. These inputs respectively capture facial contour edges and 3D structural details.

Canny Edge Guidance Features Edge information is extracted from the image I using the Canny algorithm [3] implemented via OpenCV:

$$E_{\text{canny}} = \text{Canny}(I; \theta_{\text{low}}, \theta_{\text{high}}) \tag{1}$$

Where θ_{low} and θ_{high} are the dual-threshold parameters for edge detection. This operation highlights strongly structured edge regions such as the eyebrows, jawline, and nasal wings.

Depth-Guided Features The depth map is extracted using a pretrained Transformer-based depth estimation network [39], which captures the spatial distribution and depth

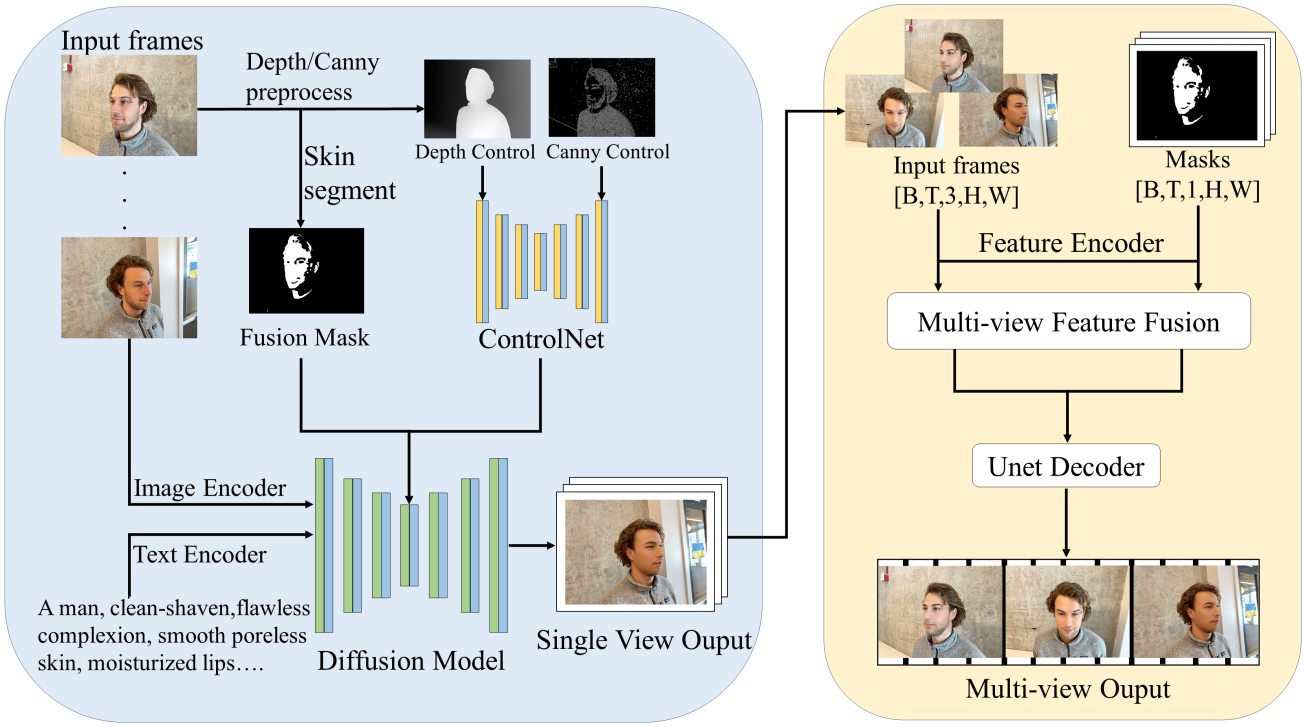

Figure 2. Architecture of MVBeautyFusion framework.The first stage performs single-view beautification using ControlNet combined with refined skin masks from skin segmentation.The second stage builds a feature-level fusion network based on temporal attention and deformable alignment to perform multi-view smoothing beautification on the results from the first stage.

relationships of the facial region in image I:

$$D_{\text{depth}} = \text{DepthEstimator}(I) \tag{2}$$

Where higher intensity values indicate regions closer to the camera (e.g., the nose tip and forehead), while lower values correspond to farther areas (e.g., behind the ears or the background).

Finally, the two feature maps $E_{\text{canny}}, D_{\text{depth}}$ are fed into separate ControlNet branches, which operate in parallel with the main backbone of Stable Diffusion, serving as structural and geometric guidance.

3.1.2 Skin Region Mask Extraction and Fusion

To prevent beautification effects from propagating to non-facial regions (e.g., background or clothing), we further introduce a fine-grained skin region mask, which is directly imposed as a hard constraint on the diffusion generation process.The mask is generated by combining structural segmentation using the Segment Anything Model (SAM) [22] with skin color detection in the YCrCb color space [23, 43, 21, 8].

SAM-Based Base Region Segmentation The SAM model processes the input image I to generate an initial foreground (facial) region mask M_{SAM}:

$$M_{\text{SAM}} = \text{SAM}(I) \tag{3}$$

Skin Color Detection Refinement Mechanism To further eliminate interference from non-skin regions (such as hair and clothing), we employ a skin color detection method based on the YCrCb color space [33] to generate a coarse skin region mask M_{skin}. This mask is then fused with the SAM mask through intersection to produce the final region mask used, denoted as M_{face}:

$$M_{\text{face}} = M_{\text{SAM}} \cap M_{\text{skin}} \tag{4}$$

Specifically, by taking the intersection of the two masks, we construct a binary skin mask that is used to directly constrain the diffusion sampling region. This mask operates in conjunction with the ControlNet guidance signals during the diffusion process, effectively suppressing the propagation of generative effects to the background and other non-target regions, while preserving fine structural and textural details within key facial areas. As a result, this hard-constrained mechanism substantially improves temporal consistency, beautification stability, and the overall visual naturalness of the generated results.

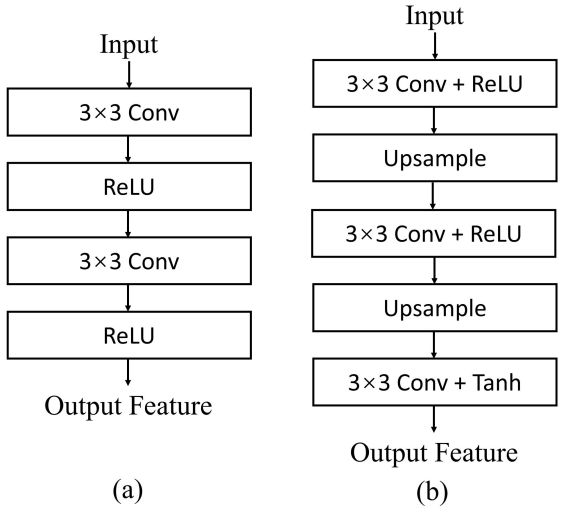

Figure 3. The encoder and decoder architecture of the cross-frame appearance fusion network. Figure (a) (left) shows the Feature Encoder module, while Figure (b) (right) presents the Decoder module.

3.1.3 Beautified Image Generation Process

The original image I, along with the guidance features E_{canny} and D_{depth}, is fed into the ControlNet-augmented Stable Diffusion model for generation:

$$\hat{I}_{\text{beauty}} = \text{StableDiffusion}(I; E_{\text{canny}}, D_{\text{depth}}, M_{\text{face}}) \quad (5)$$

Where the ControlNet branches respectively perform feature extraction and control weight prediction for the two modalities, guiding the diffusion process to preserve facial structure and ensure consistency in the beautification direction.

3.2. Cross-Frame Appearance Fusion Network

3.2.1 Feature Encoder Module

As shown in Figure 3(a), the input sequence is defined as $I = I_t \mid t = 1, ..., T$, where each frame $I_t \in \mathbb{R}^{3\times H\times W}$ is an RGB image. The encoder $E(\cdot)$ maps each frame to a corresponding feature map F_t:

$$F_t = E(I_t), \quad F_t \in \mathbb{R}^{C\times H\times W} \quad (6)$$

The encoder consists of two 3×3 convolutional layers followed by ReLU activation:

$$F_t = \text{ReLU}(\text{Conv}_2(\text{ReLU}(\text{Conv}_1(I_t)))) \quad (7)$$

This module extracts local semantic information from each frame, providing a unified semantic space for cross-frame feature fusion.

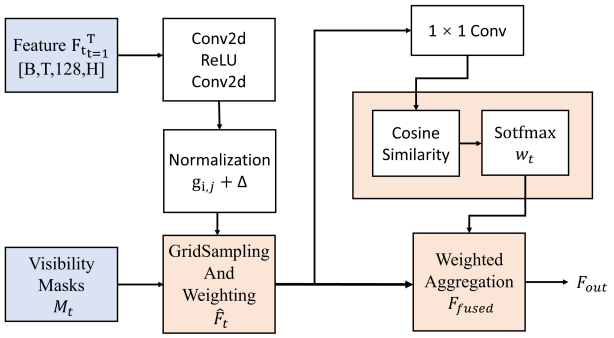

Figure 4. Multi-view Feature Fusion module architecture in the cross-frame appearance fusion network.

3.2.2 Multi-view Feature Fusion Module

This feature fusion module aims to achieve cross-frame alignment and information integration along the temporal dimension. It comprises two core submodules:First, the Deformable Sampling module extends standard convolutional sampling by introducing learnable offsets to the kernel positions, effectively adapting to geometric deformations and spatial misalignments between frames [9, 60].Second, the Temporal Attention mechanism acts as a temporal extension of the self-attention mechanism [47] over video frame sequences, enabling adaptive weighting based on the semantic relevance of features across frames.By explicitly modeling the correspondence between facial regions across frames, this module effectively mitigates cross-frame inconsistencies caused by viewpoint variation, occlusion, and lighting discrepancies, thereby enhancing the continuous smoothing and coherence of the editing results.

Deformable Sampling Mechanism As shown in Figure 4, for each frame feature map F_t, we employ a small convolutional network $O(\cdot)$ to estimate offset fields, obtaining N deformable sampling offset maps $\Delta_t^{(n)} \in \mathbb{R}^{2\times H\times W}$, where $n = 1, ..., N$ denotes the index of sampling points. The offset output has dimensions $[B, N, 2, H, W]$, reconstructed from $2N$ channels generated by a three-layer convolutional network. Here, B represents the batch size, H and W denote spatial resolution, N is the number of sampling points per location, and 2 corresponds to the two-dimensional offsets $(\Delta x, \Delta y)$ for each sampling point. In other words, for every spatial position in each frame, $\mathcal{O}(\cdot)$ predicts N offset vectors indicating the 2D displacements of the N sampling points at that position.

To achieve inter-frame structural alignment, we first define a standard normalized 2D grid coordinate $g_{i,j} \in [-1, 1]^2$, representing the center position of the grid cell at row i and column j in the $H \times W$ spatial resolution. This grid is static and applicable to all image frames, serving as

the sampling reference. Then, by element-wise addition of the n-th offset field $\Delta_t^{(n)}$ predicted by the offset network to the grid positions, we obtain the deformable sampling locations $p_{t,i,j}^{(n)}$ for each pixel along the n-th sampling path. Next, bilinear interpolation sampling is performed on the original feature map F_t at position $p_{t,i,j}^{(n)}$, yielding the resampled feature in the n-th direction:

$$\widetilde{F}_t^{(n)} = \text{GridSample}(F_t, g_{i,j} + \Delta_t^{(n)}) \tag{8}$$

To account for the visibility of the current pixel, we introduce a visibility mask $M_t \in [0,1]^{B\times 1\times H\times W}$ output by a semantic segmentation model. This mask is sampled at the same offset positions and used as a weighting factor for feature fusion. The final fused feature is defined as the weighted average of the sampled features across all directions:

$$\hat{F}_t = \frac{1}{N}\sum_{n=1}^{N}\left(\widetilde{F}_t^{(n)} \cdot \text{GridSample}(M_t, p_t^{(n)})\right) \tag{9}$$

Where $\text{GridSample}(M_t, p_t^{(n)})$ denotes the visibility weight sampled at the offset position.

Content-Aware Temporal Attention Mechanism To further enhance semantic smoothness across time, we introduce a lightweight temporal attention mechanism [45], which extends self-attention to sequential temporal data [47]. By dynamically modulating feature importance across frames, it effectively ensures inter-frame style consistency and stability. Specifically, the module uses the intermediate frame feature $\hat{F}t\text{center}$ as the Query, while the other frame features $\hat{F}_t$ serve as the Key and Value, which are linearly mapped to Q and K_t respectively via 1×1 convolutions:

$$Q = Q_{\text{conv}}(\hat{F}_{t_{\text{center}}}), \quad K_t = K_{\text{conv}}(\hat{F}_t) \tag{10}$$

Where Q denotes the query vector of the center frame, initiating the attention query; K_t represents the key vector of frame t, used to compute attention matching. Next, the point-wise cosine similarity is calculated at each spatial location to measure the semantic correlation between the current frame and the reference frame. The similarity score α_t is further modulated by multiplying with the corresponding frame's visibility mask M_t to suppress attention values in occluded regions:

$$\alpha_t = \text{cosine\_similarity}(Q, K_t)\cdot M_t, \quad \alpha_t \in \mathbb{R}^{1\times H\times W} \tag{11}$$

The similarity scores α_t are normalized across the temporal dimension using softmax to obtain cross-frame attention weights w_t, which are then used to compute the fused feature representation:

$$F_{\text{fused}} = \sum_{t=1}^{T} w_t \cdot \hat{F}_t \tag{12}$$

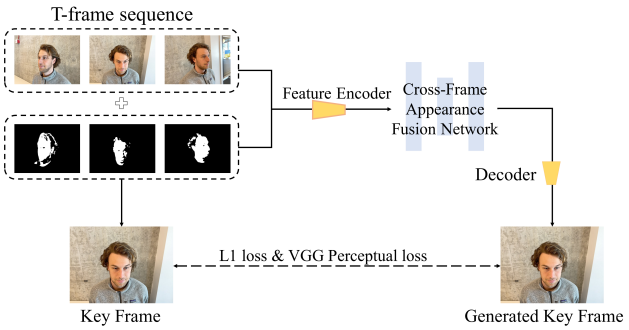

Figure 5. Training process of the cross-frame appearance fusion network. The model extracts features from a sequence of T frames ($T \geq 3$) and performs cross-frame aggregation through the network to reconstruct an enhanced central frame image. Joint supervision with L1 and perceptual losses is applied during training.

Finally, a linear projection module $P(\cdot)$ is applied to the fused feature F_{fused} for channel compression or expansion, producing the final output feature:

$$F_{\text{out}} = P(F_{\text{fused}}) \tag{13}$$

This mechanism allows the model to adaptively refine fusion weights based on semantic correlations across temporal frames, substantially improving the continuous smoothing and coherence of beautification styles throughout video sequences. It proves especially effective in handling dynamic scenarios such as facial expression transitions and head pose variations.

3.2.3 Decoder Module

The decoder adopts a UNet-style architecture to map the fused features back to RGB images:

$$\hat{I} = D(F_{\text{out}}), \quad \hat{I} \in \mathbb{R}^{3\times H\times W} \tag{14}$$

As shown in Figure 3b, the architecture consists of three convolutional layers and two $2\times$ upsampling operations, with a final $\tanh$ activation applied to the output:

$$\begin{aligned} D = {} & \text{Conv} \to \text{ReLU} \to \text{Upsample} \\ & \to \text{Conv} \to \text{ReLU} \to \text{Upsample} \\ & \to \text{Conv} \to \tanh \end{aligned} \tag{15}$$

3.3. Loss Function Design

The model employs a combination of perceptual loss and L1 pixel loss to balance structural preservation and perceptual quality.

Perceptual Loss Features are extracted using the first 9 layers of the VGG16 network, denoted as $\phi(\cdot)$. Input images are resized to 256×256 and normalized. The perceptual loss is defined as:

$$L_{\text{percep}} = \left\| \phi(\hat{I}) - \phi(I) \right\|_1 \tag{16}$$

Where $\hat{I}$ denotes the model output, and I represents the target image (ground truth).

Final Loss The overall loss function is defined as follows:

$$L_{\text{total}} = \lambda_1 \cdot \left\| \hat{I} - I \right\|_1 + \lambda_2 \cdot \left\| \phi(\hat{I}) - \phi(I) \right\|_1 \tag{17}$$

Where $\lambda_1 = 1.0$ and $\lambda_2 = 0.1$. As shown in Figure 5, this loss function simultaneously focuses on image structural reconstruction and perceptual quality during training.

3.4. Training Protocol Design

The cross-frame appearance fusion network is trained to enforce temporal consistency across consecutive frames while preserving the per-frame appearance generated by the first stage. To this end, we construct short temporal clips from real continuous video datasets such as VFHQ[55] for training. Each clip consists of T consecutive frames along with their corresponding facial/skin region masks.

For each clip, the per-frame beautified results produced by the fixed first-stage model are used as inputs to the fusion network. Importantly, although the inputs to the fusion network are generated independently per frame, the supervision signal is derived from real temporal structures present in continuous videos. The refinement loss is defined over the entire predicted sequence, encouraging cross-frame consistency and smooth transitions, rather than enforcing strict per-frame reconstruction accuracy. The first-stage model remains frozen throughout training and does not receive gradient updates.

4. Experiments

4.1. Experimental Setup

In practice, we employ a Stable Diffusion v1.5-based image diffusion generation model as the backbone, combined with two ControlNet [56] preprocessors—Canny [4] and Depth [38, 39]—which respectively capture facial edge and depth structural features to guide the high-quality generation of base images.We evaluated our method on datasets including ForgeryNet [15] and Facevid [10], selecting several video clips featuring diverse poses, lighting variations, and facial dynamics, each containing approximately 50–80 frames.

During inference, we use the DDIM sampler with 25 steps. On an NVIDIA 4090 GPU, single-frame generation in the first stage takes about 1.8 seconds per frame, while the second-stage sequence frame fusion processing for a 60-frame sequence takes roughly 40 seconds, with an average GPU memory usage of 13GB. Moreover, the introduction of a sliding window in the second stage allows the method to handle input videos of arbitrary resolution.

4.2. Comparison with Baseline Methods

Qualitative Results We conducted a qualitative evaluation of multi-view consistent editing methods, with a focus on beautification across continuous video frames involving head movements. As shown in Figure 6, we compare the beautification results of our full two-stage pipeline, BeautyGAN[24], TokenFlow[42], and BFVR[50] on different frame sequences. TokenFlow (third row) preserves the motion trajectory of the original video; however, due to its lack of structural transformation modeling, it has limited editing capability and often produces blur or distortion under complex pose changes, as observed in the first images of cases (A) and (C). BeautyGAN (second row), on the other hand, is prone to color drift and style inconsistency. Its makeup transfer mechanism, based on a single reference image, lacks generalization across different viewpoints. BFVR (fourth row) can optimize visual quality while maintaining temporal continuity of the original video, but its overall beautification effect is relatively subtle compared to the original images. In contrast, the results of our first-stage generation validate the effectiveness of our single-view beautification design, while the second-stage feature fusion module significantly enhances visual smoothness and realism under multi-view conditions.

Although the single-view Stable Diffusion model employed in the first stage produces natural beautification effects for individual frames, it often lacks continuous smoothing and leads to temporal discontinuities when extended to multi-frame sequences. As shown in Figure 9, the third beautified frame generated in the first stage depicts the lower lip occluding the teeth, which is clearly inconsistent with adjacent frames. In contrast, our full two-stage approach (Figure 7) demonstrates significantly better consistency and stability when handling multi-view and sequential frames, effectively reducing artifacts, defocus, and regional flickering.

Due to space limitations, we present additional visual results and video comparisons in the supplementary materials.

Quantitative Results We conducted a comprehensive quantitative comparison of several mainstream image editing methods in the context of video facial beautification, with a particular focus on two key aspects: image reconstruction quality and cross-frame consistency. The compared methods include TokenFlow [42] ,BeautyGAN [24] and BFVR[50], along with two variants of our proposed ap-

Figure 6. Comparison with Baseline Methods.

Figure 7. Full Two-Stage Visualization Results of MVBeautyFusion

proach: the single-frame beautification stage (Ours-S1) and the full two-stage fusion method (Ours-Full).

For image quality evaluation, in addition to conventional metrics such as PSNR[19, 1], SSIM[1] , and FID[18] ,

Dataset	Method	Attractiveness↑ VGG16	Attractiveness↑ ResNet18	PSNR↑	SSIM↑	FID↓
	BeautyGAN	0.646	3.303	13.838	0.645	13.802
	Tokenflow	0.607	3.343	24.155	0.715	11.135
	BFVR	0.647	**3.548**	23.908	0.678	15.114
	MVBeautyFusion-S1(Ours)	0.638	3.174	**29.778**	**0.897**	11.758
Old face	MVBeautyFusion-Full(Ours)	**0.653**	3.129	26.663	0.819	**8.509**
	BeautyGAN	0.489	2.641	19.093	0.853	30.441
	Tokenflow	0.430	2.729	**32.497**	**0.917**	31.637
	BFVR	0.410	2.360	26.512	0.851	24.421
	MVBeautyFusion-S1(Ours)	0.512	**2.806**	28.941	0.908	30.751
Young face	MVBeautyFusion-Full(Ours)	**0.515**	2.776	27.974	0.868	**23.433**
	BeautyGAN	0.536	2.552	10.289	0.686	12.816
	Tokenflow	0.517	2.746	**31.973**	**0.906**	25.229
	BFVR	0.530	2.703	27.373	0.853	16.685
	MVBeautyFusion-S1(Ours)	**0.537**	2.762	29.276	0.868	21.240
Asian	MVBeautyFusion-Full(Ours)	0.532	**2.766**	28.722	0.843	**11.739**
	BeautyGAN	0.738	3.615	14.889	0.814	21.615
	Tokenflow	0.716	3.704	30.364	**0.922**	34.019
	BFVR	0.694	3.499	29.066	0.889	**21.269**
	MVBeautyFusion-S1(Ours)	0.734	**3.744**	**31.054**	0.916	23.841
Occidental	MVBeautyFusion-Full(Ours)	**0.738**	3.725	29.617	0.889	24.062

Table 1. Peer comparison results across different age and ethnicity categories. An upward arrow indicates that higher values of the evaluation metric correspond to better results, while a downward arrow indicates the opposite. (MVBeautyFusion-S1 denotes our first-stage method, and MVBeautyFusion-Full denotes the complete two-stage method.)

Method	PSNR Average↑ (dB)	PSNR Standard Deviation↓(dB)	LPIPS Frame↓	LPIPS First↓	CLIP Score↑
BeautyGAN	29.381	0.472	0.240	0.418	0.244
TokenFlow	29.023	1.482	0.256	0.458	0.245
BFVR	29.520	0.874	0.231	0.409	0.238
Ours-S1	28.981	**0.421**	0.264	0.465	0.246
Ours-Full	**30.122**	0.430	**0.222**	**0.401**	**0.253**

Table 2. Quantitative comparison of continuous smoothing between our method and other approaches.

we further adopt an attractiveness score as a perceptual metric. Specifically, the Fréchet Inception Distance (FID) [18]measures the quality of generated images by computing the Fréchet distance between the Gaussian distributions of input and generated images, with lower values indicating better fidelity. Peak Signal-to-Noise Ratio (PSNR) [19, 1]evaluates the pixel-level reconstruction error between generated and input images, while Structural Similarity Index (SSIM)[1] measures the structural similarity to assess the preservation of visual quality. Moreover, to capture human-centric perceptual quality, we employ attractiveness as an additional metric, evaluated by Beholder-GAN[11], which is trained on the SCUT-5500[27] dataset using VGG16[44] and ResNet18[14] backbones.

As shown in Table 1,our proposed MVBeautyFusion consistently outperforms prior methods such as BeautyGAN[24], TokenFlow[42], and BFVR[50] across multiple demographic categories, including age (old/young) and ethnicity (Asian/Occidental). For example, MVBeautyFusion achieves the highest attractiveness scores across both VGG16- and ResNet18-based evaluations, demonstrating its ability to generate visually appealing results. In terms of PSNR[19, 1] and SSIM[1] , MVBeautyFusion also leads in most cases, reflecting superior pixel-level fidelity and structural preservation. Notably, it achieves significantly lower FID[18] values compared with TokenFlow

and BeautyGAN, indicating higher realism and reduced distributional discrepancy. These advantages are consistent across different demographic subgroups, suggesting that MVBeautyFusion generalizes well to diverse populations. Furthermore, even in challenging cases such as old face beautification, MVBeautyFusion maintains superior attractiveness and SSIM[1] scores while achieving competitive FID[18] performance. Taken together, these results confirm that both the first-stage variant (MVBeautyFusion-S1) and the full two-stage model (MVBeautyFusion-Full) deliver consistently higher-quality outputs, with the full model achieving the most balanced improvements in attractiveness, fidelity, and structural consistency.

Temporal consistency of images is comprehensively evaluated using three metrics: PSNR [19, 1], LPIPS [57], and CLIP similarity [36]. Specifically, PSNR reflects the pixel-level fidelity between reconstructed and original images, with higher values indicating better reconstruction quality. LPIPS measures perceptual similarity, where LPIPS-first and LPIPS-frame [57] quantify the similarity between the first frame and subsequent frames, and between consecutive frames, respectively. These jointly account for both inter-frame continuity and overall sequence consistency, with lower LPIPS values corresponding to smoother global temporal coherence and stronger local frame consistency. CLIP similarity evaluates high-level semantic consistency across frames, where higher values indicate better preservation of style and content.

As shown in Table 2, our complete model (Ours-Full) achieves superior performance across all three metrics, demonstrating improved perceptual continuity and semantic consistency. In comparison, although TokenFlow [42] retains certain style transfer capabilities, its lack of structural modeling leads to significantly degraded frame-wise consistency. BeautyGAN [24], while slightly faster in inference speed, performs poorly in temporal stability. BFVR [50] emphasizes consistency in aspects such as lighting, but falls short in maintaining global coherence. Furthermore, even the Ours-S1 variant without temporal fusion outperforms BeautyGAN and other methods, further validating the effectiveness of our first-stage approach in ensuring both image quality and continuous smoothing.

In summary, the proposed two-stage fusion beautification framework not only surpasses existing methods in visual image quality but also exhibits superior stability and robustness in cross-frame and multi-view consistency.

4.3. Ablation Study

To evaluate the contributions of different components in our proposed method, we first conduct experiments to assess whether the refined mask extraction effectively provides regional constraints for the first-stage single-view facial beautification. Then, we compare the single-view beautification output sequence with the final multi-view fusion results to evaluate the impact of the multi-view fusion stage on enhancing the overall beautification quality and continuous smoothing.

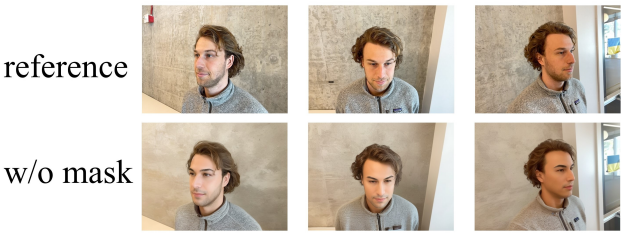

Figure 8. Effectiveness Analysis of Mask Constraints

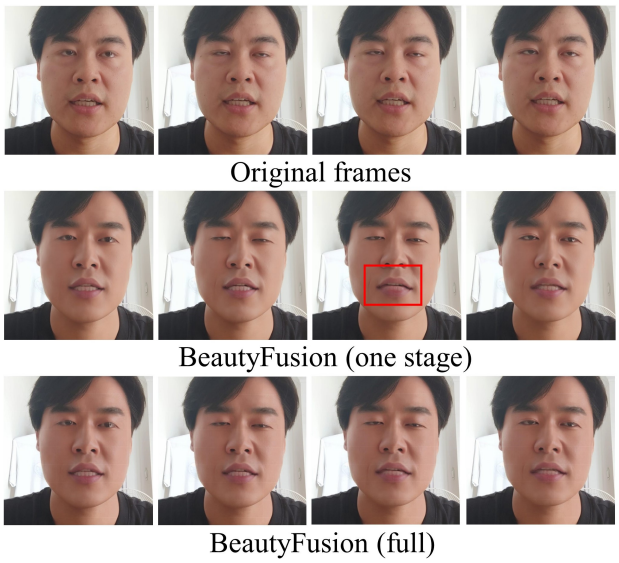

Figure 9. Comparison between the full MVBeautyFusion beautification results and those of the single-stage approach. The single-stage method exhibits mouth closure artifacts, indicating a lack of continuous smoothing across frames.

Effectiveness Analysis of Mask Constraints As shown in Figure 8, we conduct an ablation study on the mask constraints in the first stage. Without strict mask constraints during the single-view beautification stage—i.e., applying only simple beautification via text prompts—the background, which is also involved in image-to-image generation, exhibits noticeable blurring and distortions from any viewpoint. Although the main subject remains largely unaffected, the discrepancies in the background cause visual artifacts and reduce realism, deviating from the original intention of realistic beautification.

Evaluation of Multi-View Fusion Performance As shown in Figure 9, we directly compare the single-view

Dual Structrue Guidance	Fused Skin Mask	Cross-Frame Appearance Fusion Network	Attractiveness↑		LPIPS↓
			VGG16	ResNet18	
×	×	×	0.499	2.744	0.301
✓	×	×	0.512	2.732	0.297
✓	✓	×	**0.537**	2.762	0.264
✓	✓	✓	0.532	**2.766**	**0.222**

Table 3. Quantitative comparison of continuous smoothing between our method and other approaches.

beautified images generated in the first stage with the images after multi-view fusion. Experimental results demonstrate that although the single-view images roughly meet the beautification requirements under a single viewpoint, they exhibit shortcomings in multi-view coherence and frame-to-frame smoothness.

Quantitative Analysis We conduct a quantitative ablation study to systematically evaluate the contribution of each core component in the proposed MVBeautyFusion framework. Specifically, Dual Structure Guidance corresponds to the structural ControlNet[56] constraints employed in the first stage, Fused Skin Mask denotes the integration of SAM-based segmentation and skin color-based masking, and Cross-view Fusion represents the temporal feature fusion module introduced in the second stage.

As shown in Table 3, in terms of attractiveness, configurations without the second-stage fusion module exhibit noticeable instability across frames. Although individual frames may achieve competitive visual quality, the absence of explicit temporal smoothing leads to larger inter-frame variance, resulting in less stable attractiveness scores. In contrast, the complete first-stage design, which jointly incorporates structural guidance and fused skin masks, consistently outperforms settings with only a single constraint, indicating that these components are complementary rather than redundant.

Regarding temporal smoothness, measured by LPIPS[57], all three components contribute positively to reducing temporal inconsistency. Structural guidance and skin mask constraints in the first stage help suppress background interference and geometric distortion, thereby providing a more stable input for subsequent processing. The introduction of the second-stage cross-view fusion module further yields a substantial reduction in LPIPS[57], demonstrating its effectiveness in explicitly modeling cross-frame alignment and appearance consistency.

Overall, the ablation results validate that both stages are essential: the first stage ensures structure-preserving and region-aware beautification, while the second stage plays a critical role in enforcing temporal coherence and visual stability in continuous video sequences.

4.4. Robustness Evaluation under Challenging Conditions

In real-world applications, videos are often captured under adverse imaging conditions, such as insufficient illumination or severe motion blur. These factors introduce noise, blur, and structural degradation, which may negatively affect facial mask estimation and temporal consistency. To further evaluate the generalization capability of the proposed method, we conduct additional robustness experiments under challenging real-world scenarios.

Specifically, we construct two types of test clips from real continuous videos: (a) severe motion blur scenes and (b) extreme low-light scenes. Each clip consists of consecutive frames exhibiting noticeable brightness degradation or motion-induced artifacts, posing significant challenges to both structural guidance and skin-region extraction.

Figure 10(a) presents the results under severe motion blur. Even when facial boundaries become ambiguous or geometrically distorted, the proposed mask fusion strategy is still able to reliably capture facial regions. Combined with the second-stage temporal fusion network, the method produces smooth and coherent outputs with substantially reduced flickering and geometric distortion.

As illustrated in Figure 10(b), under low-light conditions, despite decreased contrast and increased sensor noise, the fused skin mask can still accurately localize facial areas. Benefiting from the structural prior provided by SAM[22] and the refinement in the YCrCb color space[23, 43, 21, 8], the diffusion process remains effectively constrained within valid facial regions, thereby preventing background artifacts and maintaining stable and natural beautification results.

Overall, these experiments demonstrate that the proposed mask fusion mechanism and temporal refinement module exhibit strong robustness and generalization ability in challenging real-world environments.

4.5. Computational Complexity Analysis

To evaluate the computational efficiency of the proposed temporal fusion network, we analyze its model complexity using the third-party profiling tool THOP to measure both the parameter count and theoretical FLOPs. The second-stage network contains only 0.39M parameters and requires approximately 64.85 GFLOPs for processing an input clip. Since the computation is amortized across multiple frames,

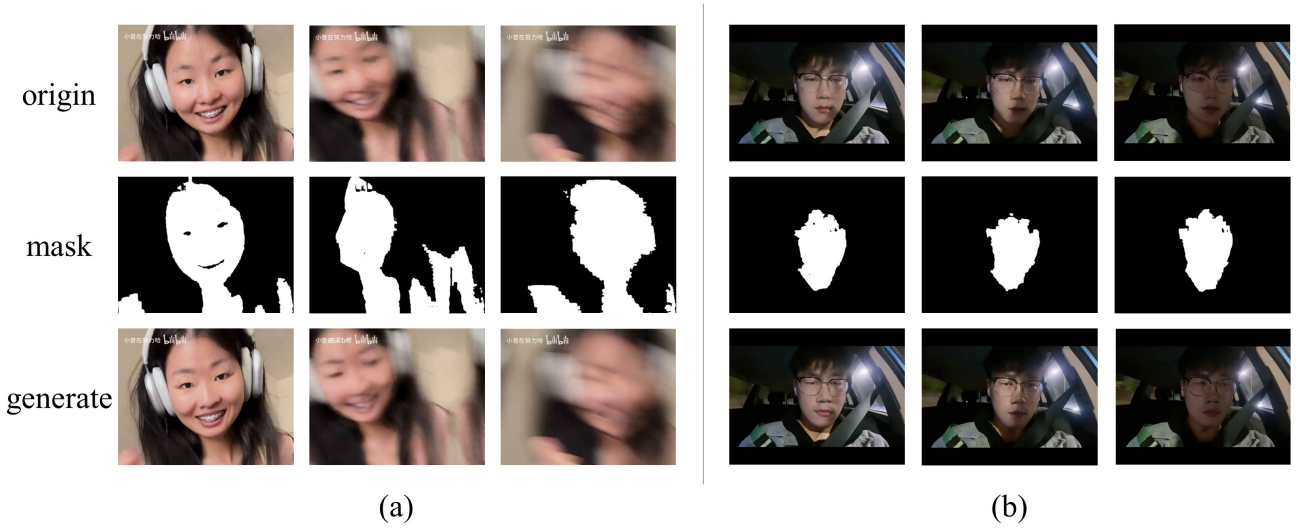

Figure 10. Qualitative robustness evaluation under challenging conditions. (a) Severe motion blur scenes. (b) low-light scenes.

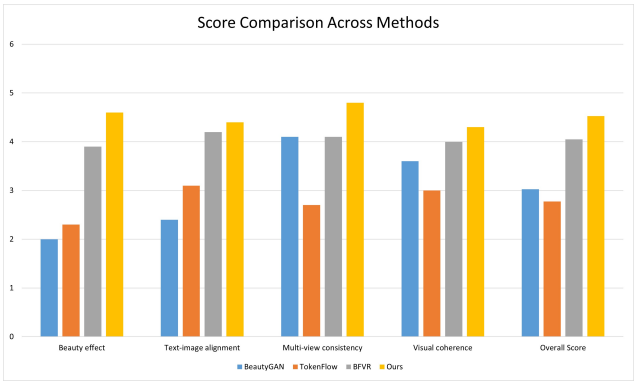

Figure 11. User Study Results.

the per-frame cost is about 13 GFLOPs, which is substantially lower than that of typical UNet backbones or diffusion-based generators. These results demonstrate that the proposed module adopts a compact and efficient architecture, introducing only minimal additional computational overhead while effectively improving temporal consistency.

4.6. User Study

To further validate the proposed method, we conducted a user study collecting 30 valid responses. Participants evaluated the beautified images based on four key criteria: beautification effect, text-image alignment, multi-view consistency, and visual temporal smoothness. The results shown in Figure 11 confirm the quantitative findings presented earlier, demonstrating that our method achieves superior performance across all evaluation metrics.

5. Conclusion

We propose MVBeautyFusion, a novel multi-view face beautification framework that achieves high-quality and continuously smoothed results across frames. In the first stage, ControlNet is guided by skin masks, Canny edges, and depth priors to generate identity-preserving base beautified images. The second stage introduces a multi-view feature fusion module, which integrates deformable sampling and content-aware temporal attention to enhance cross-frame smoothness under varying views. Extensive experiments demonstrate that MVBeautyFusion surpasses existing approaches in beautification quality, continuous smoothing, and multi-view consistency. Due to its lightweight nature and small training scale, it is ideal for real-time and mobile deployment scenarios.

Limitation and future works. Currently, the two-stage model's perception of 3D human body morphology remains somewhat limited.Future work will focus on integrating 3D structure awareness and developing an end-to-end training pipeline to further boost performance.

Acknowledgement

This research was supported by the National Natural Science Foundation of China (No.62272201) and the Qing Lan Project of Jiangsu Province (Zhenping Xie).

References

[1] I. Q. Assessment. From error visibility to structural similarity. volume 13, page 93, 2004. 8, 9, 10

[2] W. Bao, Y. Yang, and X. Wang. Visibility-aware video object segmentation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 11186–11195, 2022. 2

[3] G. Bradski. The opencv library. In *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 2000. 3

[4] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 2, 3, 7

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 7291–7299, 2017. 3

[6] H. Chang, H. Zhang, J. Barber, et al. Muse: Text-to-image generation via masked generative transformers. *ArXiv preprint arXiv:2301.09567*, 2023. 3

[7] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 3

[8] D. Dahmani, M. Cheref, and S. Larabi. Zero-sum game theory model for segmenting skin regions. *Image and Vision Computing*, 99:103925, 2020. 2, 4, 11

[9] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, and J. Sun. Deformable convolutional networks. In *ICCV*, 2017. 2, 5

[10] D. Di, H. Feng, W. Sun, Y. Ma, H. Li, W. Chen, X. Gou, T. Su, and X. Yang. Facevid-1k: A large-scale high-quality multiracial human face video dataset. *arXiv preprint arXiv:2410.07151*, 2024. 7

[11] N. Diamant, D. Zadok, C. Baskin, E. Schwartz, and A. M. Bronstein. Beholder-gan: Generation and beautification of facial images with conditioning on their beauty level. *arXiv preprint arXiv:1902.02593*, 2019. 9

[12] Y. Feng, F. Wu, X. Shao, T. Bolkart, H. Kim, and M. J. Liu. Learning an expressive 3d morphable face model with a deep neural network. *arXiv preprint arXiv:2101.07405*, 2021. 3

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 2672–2680, 2014. 2

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016. 9

[15] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. *arXiv preprint arXiv:2103.05630*, 2021. 7

[16] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want. In *IEEE Transactions on Image Processing*, volume 28, pages 5464–5478, 2019. 3

[17] A. Hertz, R. Mokady, K. Aberman, F. Perazzi, O. Bachmann, T. Dekel, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017. 8, 9, 10

[19] A. Hore and D. Ziou. Image quality metrics: PSNR vs. SSIM. In *2010 International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010. 8, 9, 10

[20] X. Jin, X. Zhao, J. Liang, M. Xu, and C. C. Loy. Videocomposer: Compositional video synthesis with motion controllability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[21] U. Kaya and M. Başaran. A comparative study of classification methods on human skin detection from rgb and ycbcr represented color images. *Eskişehir Technical University Journal of Science and Technology*, 21:40–44, 2020. 2, 4, 11

[22] A. Kirillov, E. Mintun, N. Ravi, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 4, 11

[23] J. Kovač, P. Peer, and F. Solina. Human skin color clustering for face detection. In *International conference on computer as a tool (EUROCON)*, volume 2, page 144–148. IEEE, 2003. 2, 4, 11

[24] D. Li, S. Lin, H. Yang, J. Liu, and X. Wang. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 645–653, 2018. 1, 2, 7, 9, 10

[25] L. Li, J. Hou, W. Liu, Y. Fang, and J. Yan. Diffusion-based facial aesthetics enhancement with 3d structure guidance. *IEEE Transactions on Image Processing*, 2025. 1

[26] Y. Li, W. Bao, W.-S. Zhang, and Q. Yang. Depth-aware flow projection for video frame interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 100–117, 2022. 2

[27] L. Liang, L. Lin, L. Jin, D. Xie, and M. Li. Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. 2018. 9

[28] L. Ma, Y. Wang, R. Wu, and X. Wang. Scgan: Semantically-consistent gan for facial makeup transfer. In *CVPR*, 2022. 1, 2

[29] M. Mirza and S. Osindero. Conditional generative adversarial nets. In *arXiv preprint arXiv:1411.1784*, 2014. 1

[30] C. Mou, Z. He, W. Zhang, H. Zhang, X. Wang, and L. Zhang. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *CVPR*, 2023. 3

[31] X. Pan, Y. Zhang, J. Liang, Q. Zhang, C. C. Loy, and Z. Li. Drag your gan: Interactive point-based manipulation on the generative image manifold. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023. 3

[32] W. Peebles and S. Xie. Scalable diffusion models with transformers. *ICCV*, 2023. 3

[33] D. T. Phan and J. H. Choi. Human skin detection using ycbcr color space. *Proceedings of the International Symposium on Image and Signal Processing and Analysis*, 2003. 4

[34] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent

diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[35] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 10

[37] A. Ramesh, M. Pavlov, and G. e. a. Goh. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 3

[38] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. *CoRR*, abs/2103.13413, 2021. 2, 3, 7

[39] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 2, 3, 7

[40] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3

[41] C. Saharia, W. Chan, S. Saxena, L. Li, T. Salimans, J. Ho, D. J. Fleet, and Q. V. Le. Imagen: Scaling up diffusion models for text-to-image generation. *arXiv preprint arXiv:2205.11487*, 2022. 3

[42] R. Shi, B. Dai, P. Li, Y. Zhang, X. Wang, and Y. Chen. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 1, 2, 3, 7, 9, 10

[43] L. Sigal and S. Sclaroff. Skin color modeling and adaptation. *Technical Report BUCS-TR-2004-014, Boston University*, 2004. 2, 4, 11

[44] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 9

[45] C. Tan, Z. Gao, L. Wu, Y. Xu, J. Xia, S. Li, and S. Z. Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18770–18782, 2023. 2, 6

[46] M. Tao, X. Chen, L. Yuan, N. Yu, L. Zhang, and J. Lu. Fatezero: Fusing attentions for zero-shot text-driven image-to-image translation. In *NeurIPS*, 2022. 1

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 5, 6

[48] Q. Wang, J. Zhang, C. Xu, W. Cao, Y. Tai, Y. Han, Y. Ge, H. Gu, C. Wang, and Y. Fu. Difffae: Advancing high-fidelity one-shot facial appearance editing with space-sensitive customization and semantic preservation. *arXiv preprint arXiv:2403.17664*, 2024. 1

[49] X. Wang, Z. Liu, W. Xu, X. Ren, and X. Lu. Instantid: Zero-shot identity-preserving generation in seconds. In *arXiv preprint arXiv:2312.08070*, 2023. 3

[50] Y. Wang, J. Teng, J. Cao, Y. Li, C. Ma, H. Xu, and D. Luo. Efficient video face enhancement with enhanced spatial-temporal consistency. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2183–2193, 2025. 7, 9, 10

[51] X. Wei, X. Li, M. Ding, X. Zhu, L. Xu, and J. Yu. Styleuv: Facial uv map for identity-preserving makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[52] J. Wu, Y. Yang, X. Chen, Z. Lin, Y. Zhao, and D. Lin. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022. 3

[53] Y. Wu, X. Luo, C. Zhang, J. Yu, et al. Faceverse: a fine-grained controllable 3d face dataset and generation framework. *arXiv preprint arXiv:2301.11743*, 2023. 3

[54] Y. Xia, T. He, Y. Li, Y. Zhang, et al. Photomaker: Controllable identity-preserving text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17000–17010, 2023. 3

[55] L. Xie, X. Wang, H. Zhang, C. Dong, and Y. Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 7

[56] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 3, 7, 11

[57] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 10, 11

[58] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023. 2

[59] Y. Zheng, X. Zhan, J. Lin, D. Tao, and C. C. Loy. Imavatar: Implicit morphable head avatars from videos. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 3

[60] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019. 5