Unleashing the Potential of Transformer Flow for Photorealistic Face Restoration

Kepeng Xu[1] , **Li Xu**[1] , **Gang He**[1*] , **Wei Chen**[2] , **Xianyun Wu**[1] , **Wenxin Yu**[3,4]

[1]Xidian University
[2]Beihang University
[3]Southwest University of Science and Technology
[4]Fujiang Laboratory

Abstract

Face restoration is a challenging task due to the need to remove artifacts and restore details. Traditional methods usually use generative model prior to achieve face restoration, but the restored results are still insufficient in terms of realism and details. In this paper, we introduce OmniFace, a novel face restoration framework that leverages Transformer-based diffusion flow. By exploiting the scaling property of Transformer, OmniFace achieves high-resolution restoration with exceptional realism and detail. The framework integrates three key components: (1) a Transformer-driven vector estimation network, (2) a representation aligned ControlNet, and (3) an adaptive training strategy for face restoration. The inherent scaling law of Transformer architectures enables the restoration of high-quality faces at high resolution. The controlnet combined with pre-trained diffusion representation can be easily trained. The adaptive training strategy provides a vector field that is more suitable for face restoration. Comprehensive experiments demonstrate that OmniFace outperforms existing techniques in terms of restoration quality across multiple benchmark datasets, especially in restoring photographic-level texture details in high-resolution scenes.

1 Introduction

Blind face restoration aims to recover high-quality face images from low-quality inputs afflicted by unknown degradations such as low resolution[Zeng *et al.*, 2023], artifacts[Zhang *et al.*, 2022], noise[Yang *et al.*, 2020], and blur[Lai *et al.*, 2022]. Early method[Zhu *et al.*, 2016] directly trained neural networks to regress high-quality face images. In recent years, the quality of restoration has significantly improved through the utilization of various face priors. Geometric priors, including face landmarks[Chen *et al.*, 2018], parsing maps[Chen *et al.*, 2021], and heatmaps[Yu *et al.*, 2018], are essential for accurately restoring the shapes of face components. Generative priors have also played a crucial role; for instance, methods[Wang *et al.*, 2021a] leveraging StyleGAN learn to restore fine details, while approaches[Zhou *et al.*, 2022a] utilizing VQGAN model priors enhance the overall quality of face restoration.

Currently, diffusion-based face restoration methods have achieved state-of-the-art performance, typically comprising two stages. The first stage employs neural networks such as SwinIR[Liang *et al.*, 2021] to remove artifacts and noise, while the second stage controls the generation of image content based on the low-quality input. For example, DiffBIR[Lin *et al.*, 2023] leverages Stable Diffusion v1-5's conditional generation capabilities to perform face restoration using a U-Net model for latent space denoising. Similarly, FlowIE[Zhu *et al.*, 2024] constructs Noise-GT paired datasets using pre-trained Stable Diffusion models and learns image restoration through a second Rectified Flow[Liu *et al.*, 2022].

Despite these advancements, existing methods exhibit several limitations:

Insufficient texture & lack of realism: Previous methods rely on the U-Net architecture as a score (noise, vector) estimation network. As the number of model parameters increases, the performance of U-Net tends to plateau [Peebles and Xie, 2023a], which makes it difficult to restore more detailed faces. Moreover, the results produced by the Unet-based diffusion face restoration method tend to have an oil painting texture and are not realistic.

Challenges in Training ControlNet: Previous methods trained ControlNet from scratch, which is feasible for models with a small number of parameters. However, as model size continue to grow, training ControlNet from scratch on specific datasets becomes increasingly difficult and costly, making them prone to mode collapse and overfitting. As illustrated in Fig. 2, simply training a ControlNet from scratch for face restoration leads to mode collapse issues.

Suboptimal Training Strategies: The loss function weighting strategies are directly inherited from the original diffusion model[Esser *et al.*, 2024]. However, face restoration task require not only the generation of fine details but also the balance of aesthetic quality. Therefore, directly adopting existing training strategies is ineffective for training gradient estimators tailored to this specific task.

To address the challenges in face restoration, we propose **OmniFace**, a novel framework built around three core in-

*Corresponding author

Figure 1: Qualitative results. Even though input faces are severely degraded, our OnmiFace produces high-quality faces with faithful details.

novations: (1) utilizing the scaling law properties of Transformer with large parameter counts to enhance face texture quality, (2) Propose *C-Projector* to control the Transformer diffusion model for face restoration, and (3) designing adaptive loss constraints specifically tailored to face restoration task.

Enhancing Face Texture with Transformer Scaling Law The first innovation in OmniFace lies in leveraging the scaling laws of Transformer, which enable significant improvements in texture quality for high-resolution face restoration. Compared to GAN-based and diffusion-based methods, OmniFace fully exploits the large parameter counts of Transformer, achieving superior clarity and realism, as shown in Fig. 1. OmniFace takes full advantage of the scaling laws, delivering high-fidelity face textures.

***C-Projector*: Effective Control via Parameter Reuse-Based ControlNet** The second innovation addresses a key challenge in controlling Diffusion Transformer (DIT) for face restoration. As shown in Fig. 2, direct control of DIT often leads to mode collapse and inconsistent results, limiting their applicability to face restoration task. To overcome this, we propose a parameter reuse-based ControlNet, which effectively stabilizes and controls the denoising Transformer. By reusing parameters, the ControlNet ensures robust operation and prevents mode collapse, enabling the generation of consistent and realistic face detail.

Adaptive Training Strategy We propose an adaptive training strategy that dynamically balances perceptual constraints and structural constraints. Different from the original loss function weighting strategy in the diffusion model, we count the values of each denoising step in the diffusion process on multiple loss functions, and adjust the corresponding

Figure 2: The original Transformer-based ControlNet generates significant texture artifacts in face restoration, and large-scale networks suffer from overfitting and mode collapse during training. As shown in Fig. 4, the weight distribution exhibits numerous anomalous values.

loss weights according to the magnitude of the loss function at different time steps, so as to better constrain the trajectory of the face diffusion flow.

OmniFace unifies these innovations into a cohesive framework for face restoration. By fully leveraging Transformer scaling laws, implementing the parameter reuse-based ControlNet, and adopting adaptive vector field constraints, OmniFace achieves state-of-the-art face restoration performance.

2 Related Works

2.1 Deep Learning for Image Enhancement

Deep learning has shown remarkable progress in image enhancement tasks, including inpainting, super-resolution, HDR reconstruction, deraining, and artifact removal. Various methods have been proposed using CNNs, Transformers, and GNNs to improve visual quality. Lightweight and edge-guided models have been developed for image inpainting [Li *et al.*, 2020], while temporal and attention-based techniques have advanced HDR video generation [He *et al.*, 2022; Xu *et al.*, 2023; Zhang *et al.*, 2024b; Zhang *et al.*, 2024a; Zhang *et al.*, 2023]. Graph-based methods explore patch similarities for effective deraining [Wang *et al.*, 2024]. Other works focus on video enhancement, compression, and debanding using adaptive and feature-aware networks [He *et al.*, 2021; Xu *et al.*, 2024b; Xu *et al.*, 2024a].

2.2 GAN-based Methods

Generative Adversarial Networks (GANs) have been widely explored for providing high-quality face priors in face restoration task. Methods such as [Zhou *et al.*, 2022b; Wang *et al.*, 2021b; Gu *et al.*, 2022] effectively utilize GAN-based priors to address blind face restoration (BFR) and achieve satisfactory results. However, GAN-based approaches often suffer from instability during training and require meticulous hyperparameter tuning, limiting their practical applicability.

2.3 Diffusion-based Methods

Diffusion models have emerged as a powerful generative framework due to their stability and ability to produce high-quality, diverse images [Ho *et al.*, 2020]. Unlike GANs, they avoid mode collapse by iteratively denoising random noise. Recent works have also improved their sampling efficiency without compromising visual fidelity.

In face restoration, diffusion models have shown strong performance in both zero-shot [Kawar *et al.*, 2022] and supervised settings [Lin *et al.*, 2023]. Methods like FlowIE further enhance inference speed by distilling large diffusion models through flow matching strategies.

However, existing diffusion-based face restoration methods often lack the scalability and detail-preserving capabilities of convolutional architectures. To address these issues, we propose an adaptive training strategy that leverages the scaling benefits of Transformers, achieving more realistic and detailed face restoration results.

3 Method

In this section, we introduce OmniFace, a face enhancement method based on Transformer Flow Matching. This method proposes a diffusion representation reuse ControlNet *C-Projector* to effectively control Transformer to restore high-quality faces. We begin by providing a brief overview of the background on Transformer Diffusion and Flow Matching, followed by a detailed description of our designed conditional network, adaptive training strategy. The overall network architecture of the proposed OmniFace method is illustrated in the Fig.3.

3.1 Preliminary

Diffusion Transformer

In [Peebles and Xie, 2023b], the authors propose Diffusion Transformer (DIT), which integrates transformer architectures into the diffusion modeling framework to enhance scalability and performance in generative task. By leveraging the strengths of transformers, such as modeling long-range dependencies and capturing complex data distributions, DIT addresses the limitations of traditional CNN-based diffusion models.

The core of DIT is a Transformer-Based Denoising Network, which replaces the conventional CNN denoiser with a transformer architecture that utilizes multi-head self-attention to capture global context and intricate patterns. Positional Encoding Integration ensures the retention of spatial information, crucial for task like face restoration [Dosovitskiy, 2020]. Additionally, its Scalable Architecture Design enables efficient handling of larger models and higher-resolution inputs with minimal computational overhead, making DIT a powerful and scalable framework for generative modeling.

Flow Matching

Flow Matching models are generative models that use Normalizing Flows (NFs) to transform complex distributions into simple ones via invertible transformations. Continuous Normalizing Flows (CNFs) extend this idea by modeling these transformations with ordinary differential equations (ODEs), making it easier to capture time-varying data distributions.

In Flow Matching (FM), the goal is to learn a vector field $v(x, t)$ that describes the evolution of data points over time. The data generation process is represented as:

$$\frac{dx(t)}{dt} = v(x(t), t),$$

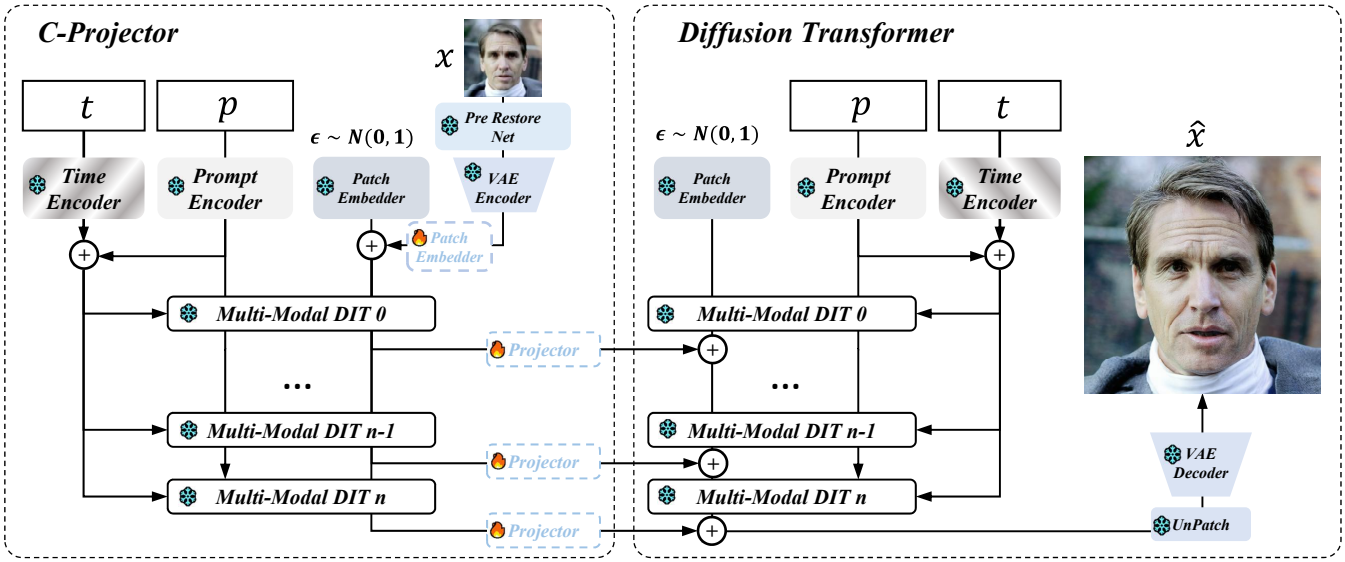

Figure 3: Framework. The structure of the proposed *C-Projector* is shown on the left, while the denoising Transformer from the diffusion model is depicted on the right. The *C-Projector* consists solely of the blue-colored Projector layer and the embedding layer for low-quality images, both of which are trained.

where $x(t)$ is the data at time t, and $v(x(t), t)$ is the vector field governing its evolution. The training objective minimizes the difference between the real data trajectory and the model's trajectory, often using a path-matching loss:

$$\mathcal{L}_{\text{FM}} = \int_0^T \|v(x(t), t) - \hat{v}(x(t), t)\|^2 dt,$$

where $\hat{v}(x(t), t)$ is the vector field from real data. After training, new samples can be generated by solving the ODE:

$$x(t + \Delta t) = x(t) + v(x(t), t)\Delta t.$$

This method improves the stability and efficiency of training, particularly in applications like diffusion models, where it helps refine the data generation process.

3.2 Framework

We propose a novel face restoration network, illustrated in Fig. 3. The architecture consists of two main components: a Diffusion Transformer Network and a projection-based ControlNet, called the *C-Projector*.

To leverage the Transformer's scaling laws for enhanced texture generation, we incorporate a large-scale Denoising Transformer (DIT) into our framework. The DIT's strong detail generation capabilities in high-parameter settings enable the restoration of intricate face features, ensuring high quality face.

To effectively control the Transformer architecture, we design the *C-Projector* $\mathcal{F}_c$. This ControlNet project the multimodal hidden features h_l into control features c, allowing fine-grained manipulation of the restoration process. By utilizing parameter reuse strategies, the *C-Projector* can efficient training and rapid convergence.

Let the input low-quality image be denoted as $x \in \mathbb{R}^{H \times W \times C}$. The image x is first processed by the pre-restoration network $\mathcal{F}_p$, resulting in an initial restoration feature:

$$\tilde{x} = \mathcal{F}_p(x).$$

Subsequently, $\tilde{x}$ is encoded by the encoder $\mathcal{E}$ of a Variational Autoencoder (VAE) to obtain the latent feature z_l:

$$z_l = \mathcal{E}(\tilde{x}).$$

The latent feature z_l, random noise $\epsilon \in \mathbb{R}^{h \times w \times c}$, text condition $p \in \mathbb{R}^m$, and time step $t \in \mathbb{R}$ are sequentially input into N multimodal Transformer modules $\{\mathcal{T}_i\}_{i=1}^N$, generating hidden features $h_l^{(i)}$:

$$h_l^i = \mathcal{T}_i(h_l^{i-1}, \epsilon, p, t), \quad i = 1, \ldots, N.$$

where $h_l^0 = z_l$. Each hidden feature h_l^i is then projected by the Projector module $\mathcal{F}_c$ to obtain the control feature c^i:

$$c^i = \mathcal{F}_c(h_l^i), \quad i = 1, \ldots, N.$$

Within the Denoising Transformer model $\mathcal{D}_T$, noise ϵ, text condition p, and time step t are input into N Transformer layers. The output of each layer y_i is combined with the control feature c and input to the next Transformer layer:

$$y^i = \mathcal{D}_T^i(y^{i-1}) + +c^i, \quad i = 1, \ldots, N.$$

where $y^0 = \epsilon$. The final output y_n represents high-quality latent features, which are subsequently decoded by the VAE decoder $\mathcal{D}$ to produce the restored image:

$$y = \mathcal{D}(y^n).$$

This architecture combines the powerful generative capabilities of Transformer model with the efficient control mechanism provided by *C-Projector*, achieving high-fidelity face image restoration.

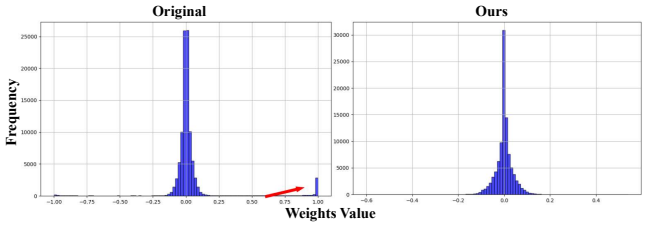

Figure 4: Comparison of weight distributions between the original ControlNet (left) and our proposed *C-Projector* (right). The *C-Projector* leverages pre-trained diffusion representations for effective training, avoiding overfitting and mode collapse. In contrast, the original ControlNet exhibits anomalous large values (highlighted by the red arrow), indicative of overfitting and instability.

3.3 *C-Projector*: Efficient Control via Parameter Reuse

During the training of Transformer-based conditional diffusion model on the FFHQ dataset, we noticed significant texture anomalies. The original models tended to generate excessive textures, which degraded restoration quality. Specifically, the hidden features of trained model exhibited large weights, a sign of overfitting (Fig. 2).

To solve this, we introduce the *C-Projector* $\mathcal{F}_c$, a ControlNet that suppresses texture anomalies. The *C-Projector* uses a parameter reuse strategy by initializing the weights of its N Transformer modules with those from a pre-trained Denoising Transformer $\mathcal{T}$.

Each Transformer module in the *C-Projector* add a Projector layer $\mathcal{P}_i$ that maps low-quality image features x to guidance features c^i in the diffusion space:

$$c^i = \mathcal{P}_i(x) = \mathbf{W}_i(\mathcal{T}(x)) + \mathbf{b}_i,$$

where $\mathbf{W}_i$ and $\mathbf{b}_i$ are learnable parameters.

Additionally, we apply Low-Rank Adaptation (LoRA) to the Transformer modules in $\mathcal{F}_c$ to further improve control with a lightweight model. LoRA introduces low-rank matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ to update the model weights as follows:

$$\theta'_{\mathcal{F}_c^i} = \theta_{\mathcal{T}} + \Delta\theta_{\mathcal{F}_c^i}, \quad \Delta\theta_{\mathcal{F}_c^i} = \mathbf{A}_i\mathbf{B}_i.$$

The LoRA method can further enhance the details and realism of the restored face.

3.4 Time Adaptive Loss Weighting Strategy for Face Restoration in Diffusion Model

We propose an adaptive loss weighting strategy for face restoration task in diffusion models. Traditional diffusion models use fixed loss weights, which are not ideal for task requiring specific attributes, such as face identity and fine-grained details. To address this, we introduce a dynamic mechanism that adjusts the loss weights based on the model's performance at each timestep, ensuring that critical aspects of face restoration, like perceptual quality and structural accuracy, are prioritized.

Loss Function and Dynamic Weight Adjustment

Traditional diffusion models rely on static loss weights (e.g., MSE) across all timesteps. However, these fixed weights are not well-suited for face restoration, which requires a more flexible approach. Our method incorporates four key loss functions: *mean squared error (MSE)*, *face recognition loss*, *perceptual loss (LPIPS)*, and *structural similarity loss (SSIM)*. The weight for each loss function is adaptively adjusted based on its magnitude at each timestep.

Formally, at timestep t, the total loss is:

$$\begin{aligned}\mathcal{L}_t = &\lambda_{\text{mse}}(t) \cdot \mathcal{L}_{\text{mse}}(\hat{x}, x) + \lambda_{\text{face}}(t) \cdot \mathcal{L}_{\text{face}}(\hat{x}, x) \\ &+ \lambda_{\text{lpips}}(t) \cdot \mathcal{L}_{\text{lpips}}(\hat{x}, x) + \lambda_{\text{ssim}}(t) \cdot \mathcal{L}_{\text{ssim}}(\hat{x}, x),\end{aligned} \tag{1}$$

where $\hat{x}, x$ respectively predict the latent and high-quality latent, and the weights are computed as:

$$\lambda_{\text{loss}}(t) = \frac{\mathcal{L}_{\text{loss}}(t)}{\mathbb{E}[\mathcal{L}_{\text{loss}}]}, \tag{2}$$

where $\mathbb{E}[\mathcal{L}_{\text{loss}}]$ is the average of the respective loss function across previous timesteps. This ensures that the loss weights are adapted based on the historical performance of each loss term.

Time-Step Dependent Weighting

Different stages of the restoration process require different focuses: early timesteps prioritize structural restoration, while later steps refine identity and details. Our adaptive strategy adjusts the weights based on the loss magnitudes, allowing the model to prioritize important features as needed.

For example, when face recognition loss is lagre, the loss function increases the weight on identity preservation, ensuring that the face's identity is preserved as the restoration progresses. Similarly, if perceptual or structural losses are significant, the model adapts to refine the generated texture and structure. In the case of MSE, when large pixel discrepancies exist, the model increases the MSE weight to focus on overall pixel-level accuracy.

Advantages of Adaptive Weighting

The adaptive loss weighting strategy allows the model to dynamically focus on the most critical aspects of face restoration, improving both the efficiency and quality of the restoration process. This approach facilitates faster convergence and enhances identity preservation and face detail recovery. Ultimately, our method enables more precise and high-quality face restoration in diffusion models, making it particularly effective for task that require both structural accuracy and fine-grained preservation.

4 Experiments

4.1 Datasets, Metrics

Train Dataset. We train our model using the FFHQ dataset[Karras, 2019], which contains 70,000 high-quality (HQ) face images. During training, all images are resized to 1024×1024. To generate low-quality (LQ) images for training pairs, we apply a degradation model. Specifically, we first convolve the HQ image with a Gaussian kernel to introduce blur, followed by downsampling by a factor of r. Gaussian noise with strength δ is then added, and JPEG compression

Method	PSNR	SSIM	LPIPS	MUSIQ	MANIQA	NIQE	CLIP-IQA	FID(FFHQ)	NIMA	TOPIQ(IAA)
RealESRGAN	26.295	**0.704**	0.520	33.865	0.341	10.355	0.423	97.660	4.658	4.621
CodeFormer	26.350	0.700	0.369	57.128	0.300	5.217	0.591	71.111	5.227	5.051
PASD	26.201	0.696	0.398	70.442	0.397	4.532	0.649	70.952	**5.347**	5.024
OnmiFace	**26.408**	0.641	**0.363**	**71.662**	**0.428**	**4.632**	**0.688**	**66.176**	4.960	**5.068**

Table 1: Quantitative Results in CelebA-Val 1000 datasets.

Method	PSNR	SSIM	LPIPS	MUSIQ	MANIQA	NIQE	CLIP-IQA	FID(FFHQ)	NIMA	TOPIQ(IAA)
RealESRGAN	26.240	**0.740**	0.498	31.925	0.335	9.932	0.412	79.897	4.501	4.280
CodeFormer	26.306	0.735	0.371	56.161	0.291	5.275	0.572	52.238	**4.927**	4.757
PASD	25.863	0.726	0.386	69.252	0.407	4.864	0.656	52.779	4.588	4.697
OnmiFace	**26.340**	0.667	**0.362**	**70.666**	**0.415**	**4.655**	**0.677**	**42.279**	4.705	**4.778**

Table 2: Quantitative Results in FFHQ Val datasets.

CelebA-512	MUSIQ	MANIQA	NIQE	CLIP IQA	FID	NIMA	TOPIQ
RealESRGAN	38.90	0.33	8.79	0.45	64.08	4.82	4.69
CodeFormer	58.75	0.30	5.13	0.60	49.02	5.25	5.07
PASD	69.51	0.39	4.94	0.63	50.98	5.29	5.20
OnmiFace	**74.37**	**0.43**	**4.72**	**0.72**	**46.96**	**5.53**	**5.35**
Child							
RealESRGAN	54.35	0.27	5.17	0.48	120.09	4.73	4.52
CodeFormer	53.62	0.23	5.38	0.50	110.94	4.63	4.41
PASD	64.04	0.32	5.36	0.59	126.67	4.98	4.81
OnmiFace	**71.29**	**0.39**	**4.71**	**0.72**	**104.47**	**5.40**	**5.19**
LFW							
RealESRGAN	56.23	0.31	5.43	0.47	49.03	4.79	4.65
CodeFormer	60.27	0.29	4.96	0.57	51.11	4.69	4.73
PASD	68.87	0.37	5.18	0.61	40.09	4.90	4.92
OnmiFace	**73.66**	**0.41**	**4.83**	**0.71**	**38.61**	**5.20**	**5.16**
WebPhoto							
RealESRGAN	38.09	0.32	7.40	0.44	105.89	4.55	4.24
CodeFormer	55.16	0.27	5.31	0.58	87.87	4.82	4.61
PASD	68.04	0.38	5.69	0.60	108.12	5.01	**4.87**
OnmiFace	**70.62**	**0.38**	**5.08**	**0.68**	**79.53**	**5.04**	4.78
Wider							
RealESRGAN	21.75	0.33	11.08	0.37	124.70	4.31	3.90
CodeFormer	50.87	0.28	5.47	0.59	56.31	4.94	4.76
PASD	64.17	0.36	5.15	0.57	49.88	5.00	4.79
OnmiFace	**72.04**	**0.40**	**4.76**	**0.71**	**35.70**	**5.19**	**4.97**

Table 3: Quantitative results on five real low-quality face datasets.

with quality factor q is applied. Finally, the LQ image is resized back to 1024×1024. The parameters for the degradation process, including the Gaussian kernel size σ, downsampling ratio r, noise strength δ, and JPEG quality factor q, are randomly sampled from predefined ranges to ensure diversity and complexity in the training data. The degradation model is consistent with [Zhou *et al.*, 2022a].

Test Datasets. Due to the lack of high-resolution face datasets, we use the last 500 images from the FFHQ dataset as test set, which are also resized to 1024×1024. To evaluate the model under various degradation conditions, we apply the same degradation process to these test images as used during training. Additionally, we evaluate our method on several other datasets, including CelebA-HQ[Karras, 2017], LFW-Test[Wang *et al.*, 2021b], WebPhoto-Test[Wang *et al.*, 2021b], WIDER-Test[Zhou *et al.*, 2022b], and Child[Wang *et al.*, 2021b]. Although the images in these datasets are of lower resolution, they cover varying levels of degradation, ranging from mild to severe, providing a comprehensive evaluation of the model's performance under different real-world conditions.

Method	CelebA HQ	FFHQ	CelebA 512	Child	LFW	Web Photo	Wider
RealESRGAN	1.89	1.83	2.04	2.97	2.70	1.96	1.32
CodeFormer	3.51	3.26	3.66	3.07	3.65	2.82	2.49
PASD	4.40	4.20	4.33	3.48	3.99	3.42	3.22
OnmiFace	**4.49**	**4.22**	**4.75**	**4.04**	**4.57**	**3.78**	**4.29**

Table 4: Quantitative comparison using Q-Align, a visual scoring metric leveraging vision-language models (VLM) to assess image quality.

Method	CelebA HQ	FFHQ	CelebA 512	Child	LFW	Web Photo	Wider
RealESRGAN	2.01	1.77	2.26	2.25	2.18	1.88	1.48
CodeFormer	2.82	2.40	2.88	2.19	2.31	2.28	2.29
PASD	3.08	2.30	3.21	2.63	2.61	**2.60**	2.40
OnmiFace	**3.08**	**2.53**	**3.43**	**2.97**	**2.88**	2.59	**2.70**

Table 5: Quantitative comparison of methods on aesthetic quality assessment using Q-Align(Aesthetic).

Evaluation Metrics. For datasets with ground truth (GT) images, we use PSNR, SSIM, and LPIPS as the basic evaluation metrics. These metrics provide a quantitative assessment of image quality, with PSNR measuring pixel-wise reconstruction accuracy, SSIM capturing structural similarity, and LPIPS evaluating perceptual similarity. In addition, to ensure comprehensive evaluation, for datasets without ground truth, we employ several no-reference metrics, including MUSIQ[Ke *et al.*, 2021], MANIQA[Yang *et al.*, 2022], NIQE[Mittal *et al.*, 2013], CLIPIQA[Wang *et al.*, 2023], FID[Jayasumana *et al.*, 2024], NIMA[Talebi and Milanfar, 2018] and TOPIQ[Chen *et al.*, 2024], to assess image quality. These metrics provide insight into the overall quality and naturalness of the generated images without requiring ground

truth data. Furthermore, we also evaluate the realistic and aesthetic quality[Wu *et al.*, 2024] of the restored faces generated by different methods. This subjective assessment ensures a more holistic evaluation of the models' performance in terms of generating realistic and visually appealing face images.

4.2 Comparisons with State-of-the-Art Methods

We compared the proposed method with state-of-the-art approaches, including RealESRGAN (GAN-based), CodeFormer (VQGAN-based), and PASD (Diffusion-based). Extensive evaluations were conducted on both synthetic and real-world datasets.

Quantitative Results. Table 1, Table 2 and Table 3 presents quantitative comparison on synthetic and real datasets. Our proposed OmniFace outperforms existing methods in terms of image quality metrics such as FID, and MUSIQ. Moreover, OmniFace achieves state-of-the-art performance on perceptual evaluation metrics, including CLIP-IQA and TOPIQ, demonstrating the superior subjective quality of the generated results. Furthermore, we used the latest TOPIQ(Face) metric to evaluate face quality. The results show that the proposed method significantly outperforms previous approaches in generating high-quality face reconstructions, as shown in Table 6. To better evaluate the quality of results produced by different methods, we introduced VLM-based Q-Align for image quality assessment. The proposed method achieves the best performance on the Q-Align metric, as shown in Table 4.

Visual Results. Fig. 1 provides a qualitative comparison, highlighting the limitations of competing methods. For example, CodeFormer introduces artifacts around the eyes, and PASD generates less realistic skin textures. Additionally, all baseline methods struggle to produce photorealistic hair. Leveraging the detail generation capabilities of Transformer, our proposed method excels in producing highly realistic hair and skin textures, achieving visually compelling results.

Method	CelebA HQ	FFHQ	CelebA 512	Child	LFW	Web Photo	Wider
RealESRGAN	0.23	0.20	0.25	0.47	0.48	0.23	0.08
CodeFormer	0.68	0.66	0.70	0.51	0.67	0.59	0.61
PASD	0.79	0.76	0.82	0.66	0.77	0.72	0.68
OnmiFace	**0.82**	**0.79**	**0.90**	**0.82**	**0.87**	**0.78**	**0.85**

Table 6: Quantitative comparison of methods using TOPIQ(Face), a specialized metric for assessing face quality, across multiple benchmarks.

4.3 Aesthetic evaluation

We further evaluated the aesthetic quality of the results produced by the proposed method using the Q-Align(Aesthetic) metric. As shown in Table 5, the proposed method achieves the best performance across all datasets, demonstrating its ability to generate visually pleasing and high-quality outputs.

4.4 Ablation studies

To evaluate the contributions of the proposed *C-Projector* ControlNet and the adaptive loss strategy, we conducted extensive ablation experiments. Quantitative results, shown in Table 7, demonstrate that removing either component leads to a significant drop in performance across multiple evaluation metrics, including MUSIQ, CLIP-IQA, and TOPIQ. Furthermore, qualitative comparisons in Fig. 5 and Fig. 6 highlight the importance of these modules. The removal of the *C-Projector* results in noticeable artifacts, while excluding the adaptive loss strategy compromises the realism and richness of textures, particularly in hair and skin. These findings validate the effectiveness of both components in enhancing face restoration quality.

Method	MUSIQ	MANIQA	CLIP IQA	NIMA	TOPIQ
OmniFace	76.98	0.69	0.81	5.38	5.34
W/o Adaptive Loss	74.59	0.63	0.75	4.96	5.08
W/o *C-Projector*	65.12	0.35	0.62	4.66	4.82

Table 7: Quantitative ablation results of *C-Projector* ControlNet and adaptive loss strategy.

Figure 5: Qualitative comparison of face restoration results with and without the *C-Projector*.

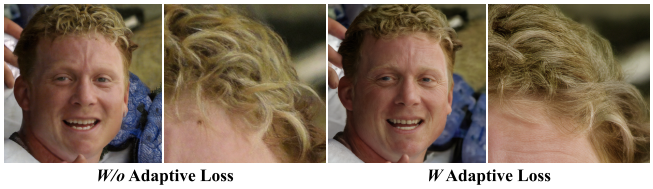

Figure 6: Comparison of face restoration results without (left) and with (right) the proposed adaptive loss. Our adaptive loss enables the recovery of richer details, such as the hair texture, while maintaining overall consistency and realism.

5 Conclusion

We proposed OmniFace, a Transformer Flow based face restoration framework that leverages diffusion flow to achieve high-resolution, photorealistic results. OmniFace outperforms state-of-the-art methods across multiple benchmarks and metrics, with ablation studies confirming the importance of its *C-Projector* and adaptive loss strategy. Our work highlights the potential of scaling Transformer for face restoration and sets a new benchmark in realistic face restoration.

Acknowledgements

This paper is supported by the construction project of the Fujiang Laboratory Nuclear Medicine Artificial Intelligence Research Center (No. 2023ZYDF074) , supported by the Fundamental Research Funds for the Central Universities under Grant ZYTS25270. **Kepeng Xu and Li Xu** contributed equally to this work. Please ask Dr. Gang He for correspondence.

References

[Chen *et al.*, 2018] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2492–2501, 2018.

[Chen *et al.*, 2021] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11896–11905, 2021.

[Chen *et al.*, 2024] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Esser *et al.*, 2024] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[Gu *et al.*, 2022] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*, 2022.

[He *et al.*, 2021] Gang He, Shan Wu, Simin Pei, Li Xu, Chang Wu, Kepeng Xu, and Yunsong Li. Fm-vsr: Feature multiplexing video super-resolution for compressed video. *IEEE Access*, 9:88060–88068, 2021.

[He *et al.*, 2022] Gang He, Kepeng Xu, Li Xu, Chang Wu, Ming Sun, Xing Wen, and Yu-Wing Tai. Sdrtv-to-hdrtv via hierarchical dynamic context feature mapping. In *Proceedings of the 30th ACM international conference on multimedia*, pages 2890–2898, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Jayasumana *et al.*, 2024] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024.

[Karras, 2017] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[Karras, 2019] Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019.

[Kawar *et al.*, 2022] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

[Ke *et al.*, 2021] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

[Lai *et al.*, 2022] Wei-Sheng Lai, YiChang Shih, Lun-Cheng Chu, Xiaotong Wu, Sung-Fang Tsai, Michael Krainin, Deqing Sun, and Chia-Kai Liang. Face deblurring using dual camera fusion on mobile phones. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 2022.

[Li *et al.*, 2020] Siyuan Li, Lu Lu, Kepeng Xu, Wenxin Yu, Ning Jiang, and Zhuo Yang. Lpi-net: Lightweight inpainting network with pyramidal hierarchy. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27*, pages 442–449. Springer International Publishing, 2020.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021.

[Lin *et al.*, 2023] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023.

[Liu *et al.*, 2022] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[Mittal *et al.*, 2013] Anish Mittal, Rajiv Soundararajan, and Alan Conrad Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20:209–212, 2013.

[Peebles and Xie, 2023a] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[Peebles and Xie, 2023b] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[Talebi and Milanfar, 2018] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

[Wang *et al.*, 2021a] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[Wang *et al.*, 2021b] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[Wang *et al.*, 2023] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

[Wang *et al.*, 2024] Cong Wang, Wei Wang, Chengjin Yu, and Jie Mu. Explore internal and external similarity for single image deraining with graph neural networks. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 1371–1379. ijcai.org, 2024.

[Wu *et al.*, 2024] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *International Conference on Machine Learning (ICML)*, 2024. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi.

[Xu *et al.*, 2023] Kepeng Xu, Li Xu, Gang He, Wenxin Yu, and Yunsong Li. Towards robust sdrtv-to-hdrtv via dual inverse degradation network. *arXiv e-prints*, pages arXiv–2307, 2023.

[Xu *et al.*, 2024a] Kepeng Xu, Zijia Ma, Li Xu, Gang He, Yunsong Li, Wenxin Yu, Taichu Han, and Cheng Yang. An end-to-end real-world camera imaging pipeline. In *ACM Multimedia 2024*, 2024.

[Xu *et al.*, 2024b] Kepeng Xu, Li Xu, Gang He, Zhiqiang Zhang, Wenxin Yu, Shihao Wang, Dajiang Zhou, and Yunsong Li. Beyond feature mapping gap: Integrating real hdrtv priors for superior sdrtv-to-hdrtv conversion. *arXiv preprint arXiv:2411.10775*, 2024.

[Yang *et al.*, 2020] Lingbo Yang, C. Liu, P. Wang, Shanshe Wang, P. Ren, Siwei Ma, and W. Gao. Hifacegan: Face renovation via collaborative suppression and replenishment. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[Yang *et al.*, 2022] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022.

[Yu *et al.*, 2018] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–233, 2018.

[Zeng *et al.*, 2023] Kangli Zeng, Zhongyuan Wang, Tao Lu, Jianyu Chen, Jiaming Wang, and Zixiang Xiong. Self-attention learning network for face super-resolution. *Neural Networks*, 160:164–174, 2023.

[Zhang *et al.*, 2022] Puyang Zhang, Kaihao Zhang, Wenhan Luo, Changsheng Li, and Guoren Wang. Blind face restoration: Benchmark datasets and a baseline model. In *arXiv:2206.03697*, 2022.

[Zhang *et al.*, 2023] Hengsheng Zhang, Li Song, Wenyao Gan, and Rong Xie. Multi-scale-based joint super-resolution and inverse tone-mapping with data synthesis for uhd hdr video. *Displays*, 79:102492, 2023.

[Zhang *et al.*, 2024a] Hengsheng Zhang, Xinning Chai, Yuhong Zhang, Rong Xie, and Li Song. Hdrtvformer: Efficient sdrtv-to-hdrtv via affine transformation and spatial-aware transformer. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 2785–2789. IEEE, 2024.

[Zhang *et al.*, 2024b] Hengsheng Zhang, Xueyi Zou, Guo Lu, Li Chen, Li Song, and Wenjun Zhang. Effihdr: An efficient framework for hdrtv reconstruction and enhancement in uhd systems. *IEEE Transactions on Broadcasting*, 70(2):620–636, June 2024.

[Zhou *et al.*, 2022a] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.

[Zhou *et al.*, 2022b] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022.

[Zhu *et al.*, 2016] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 614–630. Springer, 2016.

[Zhu *et al.*, 2024] Yixuan Zhu, Wenliang Zhao, Ao Li, Yansong Tang, Jie Zhou, and Jiwen Lu. Flowie: Efficient image enhancement via rectified flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–22, 2024.