SODiff: Semantic-Oriented Diffusion Model for JPEG Compression Artifacts Removal

Tingyu Yang[*1], **Jue Gong**[*1], **Jinpei Guo**[2], **Wenbo Li**[3], **Yong Guo**[4], **Yulun Zhang**[†1]

[1]Shanghai Jiao Tong University
[2]Carnegie Mellon University
[3]Joy Future Academy
[4]Max Planck Institute for Informatics

Abstract

JPEG, as a widely used image compression standard, often introduces severe visual artifacts when achieving high compression ratios. Although existing deep learning-based restoration methods have made considerable progress, they often struggle to recover complex texture details, resulting in over-smoothed outputs. To overcome these limitations, we propose SODiff, a novel and efficient semantic-oriented one-step diffusion model for JPEG artifacts removal. Our core idea is that effective restoration hinges on providing semantic-oriented guidance to the pre-trained diffusion model, thereby fully leveraging its powerful generative prior. To this end, SODiff incorporates a semantic-aligned image prompt extractor (SAIPE). SAIPE extracts rich features from low-quality (LQ) images and projects them into an embedding space semantically aligned with that of the text encoder. Simultaneously, it preserves crucial information for faithful reconstruction. Furthermore, we propose a quality factor-aware time predictor that implicitly learns the compression quality factor (QF) of the LQ image and adaptively selects the optimal denoising start timestep for the diffusion process. Extensive experimental results show that our SODiff outperforms recent leading methods in both visual quality and quantitative metrics.

Code — https://github.com/frakenation/SODiff

1 Introduction

JPEG (Wallace 1991) is a widely adopted lossy compression standard. It applies discrete cosine transform on each 8×8 blocks and quantizes high-frequency components.This method significantly reduces storage size while maintaining visually acceptable quality. However, high compression ratios cause over-quantization, which produces artifacts including blocking and ringing. Previous deep learning-based methods (Dong et al. 2015; Jiang, Zhang, and Timofte 2021; Liang et al. 2021; Zhang et al. 2018b) achieve considerable success in JPEG artifacts removal tasks. However, when facing such severe compression conditions, they often suffer from blocked details and distorted structures due to their lack of detail generation capability.

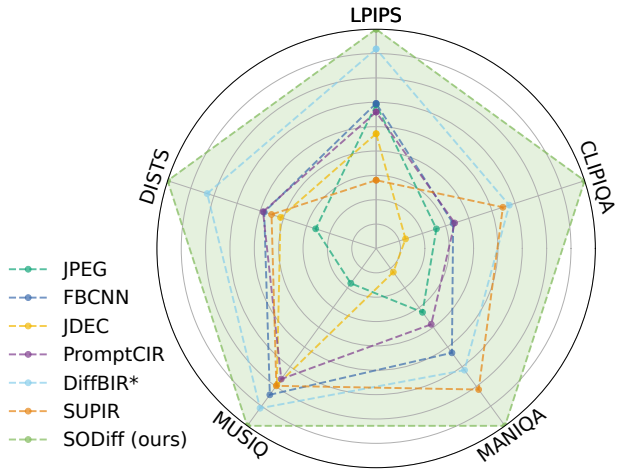


Figure 1: JEPG compression artifacts removal performance of recent methods on DIV2K-val (QF=5).

In recent years, diffusion models (DM) (Rombach et al. 2022; Song, Meng, and Ermon 2021; Ho, Jain, and Abbeel 2020) based on latent space denoising mechanisms demonstrate excellent performance in image generation. Text-to-image (T2I) models (Stability AI 2022; Podell et al. 2023) generate high-quality images that match textual descriptions by encoding input text into semantic information to guide the denoising process. Pre-trained diffusion models possess powerful image generation prior knowledge. Related studies (Lin et al. 2024; Yu et al. 2024) show their superior performance in image restoration. Considering the time overhead of multi-step diffusion inference, one-step diffusion (OSD) methods (Wu et al. 2024a; Wang et al. 2024b; Gong et al. 2025b) significantly reduce inference time. OSD methods can maintain restoration quality and provide a new technical pathway for JPEG compression artifacts removal.

JPEG compression varies with quality factors (QF) and quantization table specifications. This diversity produces variable degradation patterns. Previous studies (Guo et al. 2025) achieve good results by strategically incorporating QF prior information into the guidance prompts of diffu-

[*]These authors contributed equally.
[†]Corresponding author: Yulun Zhang, yulun100@gmail.com

sion models. Pre-trained T2I models can achieve optimal performance under text information guidance. On this basis, using textual descriptions as auxiliary information for image restoration proves to be effective (Wu et al. 2024a,b; Yu et al. 2024). Existing methods commonly employ additional vision-language models to generate text prompts online during inference. However, such practice poses two critical problems: (1) The rich information contained in low-quality (LQ) images is significantly diluted during conversion to text prompts. (2) Inference efficiency is constrained by the performance of pre-trained vision-language models.

To address the aforementioned issues, we propose SODiff, a novel semantic-oriented one step diffusion (OSD) model for JPEG compression artifacts removal. At its core is the semantic-aligned image prompt extractor (SAIPE), which employs a Swin Transformer architecture-based image feature extractor to obtain guidance information. In generating embedded "image prompts", SAIPE achieves alignment with textual description information embedded by the text encoder. Our core motivation is that pre-trained T2I models possess optimal generation prior knowledge. This prior performs best under semantic guidance. SAIPE can utilize the powerful generation prior capabilities of pre-trained models by aligning text and image embeddings.

Furthermore, for LQ images with different compression degrees, the extent of texture detail performs differently from visual information loss. The denoising process at different timesteps in the pre-trained diffusion model corresponds to blocked semantic structures at early steps and refined details at later steps. We propose a quality factor-aware timestep predictor that implicitly learns the QF of LQ images. Such a predictor can select appropriate timesteps for injecting into the diffusion UNet during the denoising process to achieve better restoration effects.

In general, our contributions can be summarized as:

- We propose SODiff, a novel OSD model for JPEG compression artifacts removal, which explores the guiding role of rich embedded visual features in the diffusion process at the textual semantic level.
- We design a semantic-aligned image prompt extractor (SAIPE) that can extract rich information from LQ images to guide the diffusion process, enabling it to extract "distilled semantic guidance" while preserving the feature priors of the images themselves.
- We design a quality factor (QF)-aware time predictor that attempts to use compression QF as a timestep predictor for diffusion models, selecting the most suitable noise for LQ images with different compression degrees.
- Our SODiff can reconstruct missing details under severe compression with high fidelity. SODiff outperforms recent leading methods on multiple existing datasets.

2 Related Works

2.1 JPEG Artifacts Removal

Over-compressed JPEG files produce noticeable artifacts. Deep learning-based JPEG artifacts removal methods have achieved remarkable success in recent years. AR-CNN (Dong et al. 2015) is the first attempt to apply super-resolution networks for artifact mitigation. Transformer-based methods (Han et al. 2024; Zhang et al. 2019) also demonstrate their effectiveness in this task. GAN network is introduced to this task (Zhang et al. 2021; Galteri et al. 2017), aiming to improve the perceptual quality of restored images through generative adversarial networks. To simultaneously utilize the rich information in both pixel and frequency domains, dual-domain convolutional network methods (Guo and Chao 2016; Zhang et al. 2018b) emerge. Prior information of JPEG compression is proven to provide strong guidance for the restoration process. The ranker-guided framework (Wang et al. 2021) uses compression quality rankings as auxiliary information. For blind removal scenarios where the QF is unknown, FBCNN (Jiang, Zhang, and Timofte 2021) balances artifact removal and detail preservation by predicting an adjustable quality factor. PromptCIR (Li et al. 2024) explores the application of prompt learning for this task. Based on the powerful image generation priors of pre-trained T2I models, CODiff (Guo et al. 2025) achieves excellent results by implicitly learning the QF of LQ images to guide detail recovery within OSD.

2.2 Diffusion Models

Pre-trained text-to-image (T2I) diffusion models possess powerful natural image generation capabilities through conditional guided denoising processes. Some methods incorporate degraded images into the diffusion process (Lin et al. 2024; Yu et al. 2024; Wu et al. 2024b), achieving excellent image restoration effects by fine-tuning pre-trained networks. DiffBIR (Lin et al. 2024) uses ControlNet (Zhang, Rao, and Agrawala 2023) to incorporate high-fidelity images as guidance information to control the denoising process. SUPIR (Yu et al. 2024) employs SDXL (Podell et al. 2023) to generate high-resolution restoration results with impressive amounts of details. One-step Diffusion (OSD) models (Wu et al. 2024a; Wang et al. 2024b; Zhang et al. 2024)encode LQ images into latent space as the starting point, achieving restoration effects in a single step. This aprach avoids the lengthy inference time and large parameter count of multi-step diffusion. Among them, OSEDiff (Wu et al. 2024a) applies variational score distillation, while S3Diff (Zhang et al. 2024) leverages degradation-aware LoRA fine-tuning. These methods provide reference paradigms for severe JPEG compression artifacts removal, but additional task-specific problems need to be considered.

3 Methods

3.1 Semantic-Aligned Image Prompt Extractor

Previous research (Wang et al. 2025; Gong et al. 2025a) demonstrates the effectiveness of extracting image features as guidance information for image restoration. The extracted visual features should retain both reconstruction and semantic information than pure text. To this end, we first train a semantic-aligned image prompt extractor (SAIPE). In the OSD stage, SAIPE can extract key image prompts from compressed LQ image I_L to guide the diffusion process.

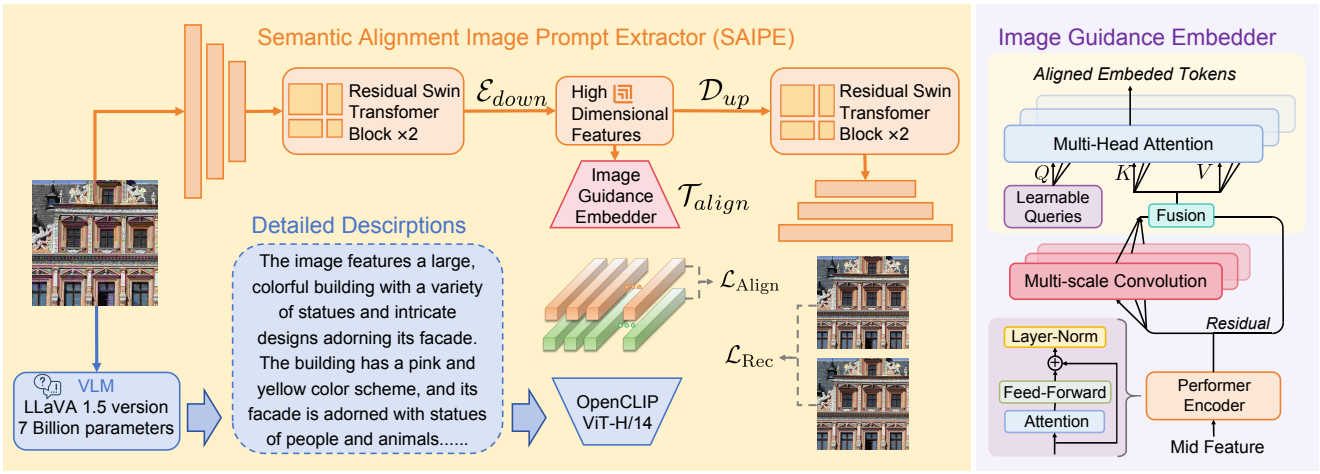


Figure 2: Training Framework of SAIPE. SAIPE employs a dual-branch training approach with a shared feature extractor $\mathcal{E}_{down}$. The LQ input image I_L is processed to generate intermediate features F_{mid}, which feed into two branches. The reconstruction decoder $\mathcal{D}_{up}$ outputs $\hat{I}_{rec}$ and is optimized via reconstruction loss $\mathcal{L}_{rec}$ against the high-quality ground truth I_H. The image guidance embedder $\mathcal{T}_{align}$ produces semantic image embeddings e_{img}. To strengthen semantic associations, LLaVA-v1.5-7B generates descriptive text from I_L, which is embedded to obtain e_{text}. The alignment loss $\mathcal{L}_{align}$ enforces semantic consistency between e_{img} and e_{text} in the embedding space, enabling effective guidance for pre-trained T2I diffusion models.

Model Architecture. As shown in Fig. 2, the main structure of SAIPE consists of a shared feature extractor encoder $\mathcal{E}_{down}$, image guidance embedder $\mathcal{T}_{align}$, and reconstruction decoder $\mathcal{D}_{up}$. Among them, the shared feature extractor encoder $\mathcal{E}_{down}$ is designed based on the SwinIR (Liang et al. 2021) architecture. The input 512×512 compressed image I_L first undergoes 4× downsampling and expands the channels to 180. It subsequently passes through two layers of residual Swin Transformer blocks (RSTB). It can be detailed as being processed through Swin Transformer layers (STL) with skip connections to preserve shallow-layer information. Through this process, it outputs intermediate feature representations F_{mid} containing rich hierarchical information. $\mathcal{D}_{up}$ correspondingly adopts a symmetric structure with reduced application of STL. After performing Layer Norm, upsampling is conducted to obtain the predicted restored image $\hat{I}_{rec}$. This process can be represented as:

$$F_{mid} = \mathcal{E}_{down}(I_L) \quad \text{with} \quad \hat{I}_{rec} = \mathcal{D}_{up}(F_{mid}). \tag{1}$$

Although high-dimensional features F_{mid} contain rich visual information, they cannot directly serve as semantic information to guide the diffusion process. Since the most suitable guidance information for pre-trained T2I models is text, designing an adapter that can focus on the semantic information of image features and establish strong associations with text semantic information is crucial. This approach can be analogized to the process of embedding "image prompts", for which we design an image guidance embedder. After F_{mid} passes through an input MLP layer, it is processed through a Performer (Choromanski et al. 2021) encoder, obtaining refined features. Following multi-scale convolutions, these serve as keys and values that enter the multi-head attention pooling layer, where they interact with learnable queries to produce image embeddings e_{img} matching the text encoder. This process can be expressed as:

$$\mathrm{e}_{img} = \mathcal{T}_{align}(F_{mid}). \tag{2}$$

Training Objective. During training, we need to preserve the original reconstruction guidance while performing semantic feature alignment. The purpose of this is to remove the degradation information obtained during the compression process of I_L. Therefore, for the reconstructed image $\hat{I}_{rec}$ after $\mathcal{D}_{up}$, we need to compute the reconstruction loss with the non-compressed image I_H:

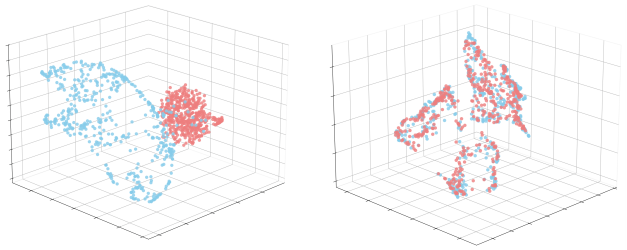
$$\mathcal{L}_{rec} = \mathcal{L}_1(\hat{I}_{rec}, I_H). \tag{3}$$

Although the compressed LQ image I_L has suboptimal image quality, it is still sufficient to guide highly robust large-scale Vision-Language Models (VLM) in generating detailed textual descriptions. Here, to strengthen the semantic features within the image embeddings, we employ LLaVA-v1.5-7B (Liu et al. 2023) to generate corresponding descriptive text from I_L. These detailed descriptions are embedded through the text embedder of the pre-trained T2I model to obtain e_{text}. To align e_{img} and e_{text} in the embedding space, we compute the MSE loss between them, allowing the semantically relevant parts within the image embeddings to be as close as possible to the spatial structure of the text embeddings. Therefore, the alignment loss $\mathcal{L}_{align}$ and the total training loss $\mathcal{L}$ can be expressed as:

$$\mathcal{L}_{align} = \mathcal{L}_{\text{MSE}}(\mathrm{e}_{img}, \mathrm{e}_{text}), \tag{4}$$

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{align} \cdot \mathcal{L}_{align}. \tag{5}$$

To intuitively illustrate that SAIPE enhances the parts with stronger semantic associations in image embeddings, we performed Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville 2018) dimensionality reduction on the embeddings. For comparison,



(a) w/o alignment (b) w/ alignment

Figure 3: Visualization of different semantic-aligned training strategies. Red dots represent text embeddings, while blue dots represent aligned embeddings under different $\mathcal{T}_{align}$ training strategies. Zoom in for better view.

we first train the reconstruction branch according to the same strategy, then train $\mathcal{T}_{align}$ together with the Diffusion model. Fig. 3 shows their comparison results. It can be observed that in the high-dimensional space distribution, $\mathcal{T}_{align}$ constrained by alignment loss is closer to the magnitude and spatial distribution of text embeddings. For a detailed comparison of model performance, please refer to the ablation experiments in Section 4.3.

3.2 One-Step Diffusion (OSD) Model

Model Architecture. Recent diffusion-based restoration methods (Wu et al. 2024a; Wang et al. 2025; Gong et al. 2025a) leverage pretrained text-to-image (T2I) models for strong generative priors. In our method, we adopt Stable Diffusion (SD) (Stability AI 2022) as the base model and retain the original architecture of both the variational autoencoder (VAE) (Kingma and Welling 2014) and the UNet (Ronneberger, Fischer, and Brox 2015) components. After obtaining the latent vector $z_L = E_\theta(I_L)$ through the VAE encoder E_θ, our model follows a denoising process similar to that in SD to predict the high-quality latent vector $\hat{z}_H$:

$$\hat{z}_H = \frac{z_L - \sqrt{1-\bar{\alpha}_\tau}\varepsilon_\theta(z_L; \mathrm{e}_{img}, \tau_{pred})}{\sqrt{\bar{\alpha}_\tau}}, \tag{6}$$

where ε_θ denotes the denoising UNet guided by both the prompt embedding e_{img} and the timestep τ_{pred}. Specifically, e_{img} is the output of the image guidance embedder of the semantic-aligned image prompt extractor (SAIPE), while τ_{pred} is predicted by the quality factor-aware time predictor. Finally, the restored high-quality image $\hat{I}_H$ is obtained by decoding $\hat{z}_H$ using the VAE decoder D_θ: $\hat{I}_H = D_\theta(\hat{z}_H)$.

Training Objective. During training, we employ three types of loss functions: reconstruction loss, adversarial loss, and QF-aware loss, which is mentioned in Sec. 3.3. The reconstruction loss consists of an MSE term and an edge-aware DISTS perceptual loss which enhances the sensitivity of the DISTS (Ding et al. 2020) metric to image edges using the Sobel operator $\mathcal{S}(\cdot)$. Therefore, the reconstruction loss $\mathcal{L}_{\text{recon}}$ can be formulated as:

$$\begin{aligned} \mathcal{L}_{\text{recon}} =& \mathcal{L}_{\text{MSE}}(\hat{I}_H, I_H) + \mathcal{L}_{\text{EA}} \quad \text{with} \\ \mathcal{L}_{\text{EA}} =& \mathcal{L}_{\text{DISTS}}(\mathcal{S}(\hat{I}_H), \mathcal{S}(I_H)) + \mathcal{L}_{\text{DISTS}}(\hat{I}_H, I_H). \end{aligned} \tag{7}$$

Meanwhile, previous work (Li et al. 2025) reveals a significant discrepancy between predicted and ground truth distributions in the latent space, which reconstruction losses in the RGB domain alone cannot effectively mitigate. Following this line of research, we adopt a pre-trained diffusion UNet as the discriminator $\mathcal{D}$. The loss functions for the generator $\mathcal{G}$ and discriminator $\mathcal{D}$ are defined as follows:

$$\mathcal{L}_{\mathcal{G}}(\hat{z}_H) = -\mathbb{E}_t\left[\log \mathcal{D}\left(F(\hat{z}_H, t)\right)\right], \tag{8}$$

$$\begin{aligned} \mathcal{L}_{\mathcal{D}}(\hat{z}_H, z_H) =& -\mathbb{E}_t\left[\log\left(1 - \mathcal{D}\left(F(\hat{z}_H, t)\right)\right)\right] \\ & -\mathbb{E}_t\left[\log \mathcal{D}\left(F(z_H, t)\right)\right], \end{aligned} \tag{9}$$

where $F(\cdot)$ represents the process of adding diffusion noise, which depends on a random timestep $t \in [0, T]$.

3.3 Quality Factor-Aware Time Predictor

Although one-step denoising reduces inference time, pre-trained T2I models achieve optimal performance with multi-step progressive refinement (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021). For a powerful pre-trained SD-UNet, the predicted noise levels at different timesteps in the latent space are not uniform (Luo et al. 2023). Previous study (Wang et al. 2024a) demonstrates that introducing a suitable timestep predictor for the OSD model can yield better performance compared to inference with a fixed timestep.

Before compression, a specific quality factor (QF) is typically specified , where lower QF generally leads to more severe detail loss. To leverage the prior knowledge embedded in the pre-trained model's noise schedule, we propose a quality factor-aware time predictor. It predicts the denoising timestep from LQ image I_L using four residual blocks. To further inject JPEG compression priors into the decision process, an additional linear branch following a shared average pooling layer is employed to predict the QF of I_L. The predicted quality factor QF_{pred} is supervised using an L1 loss against the ground truth QF_{gt}, formulated as:

$$\mathcal{L}_{qf} = ||\text{QF}_{pred} - \text{QF}_{gt}||_1. \tag{10}$$

For the dedicated timestep prediction branch, ensuring the differentiable property of input timestep is crucial for end-to-end training of the overall Diffusion framework. Since discrete timestep selection would break the gradient flow, we employ a continuous relaxation approach. The predicted logits $l_i, i \in \{0, 1, 2, \ldots, T_{\max} - 1\}$ corresponding to different timesteps need to be processed through the Gumbel-Softmax technique (Jang, Gu, and Poole 2017), which provides a differentiable approximation to categorical sampling. This approach introduces Gumbel noise $g_i \sim \text{Gumbel}(0, 1)$ to the logits before applying softmax normalization, enabling smooth gradient computation while maintaining the stochastic nature of timestep selection. Therefore, the predicted timestep τ_{pred} can be expressed as a learnable weighted combination of all possible timesteps:

$$\tau_{pred} = \sum_{i=0}^{T_{max}-1} i \cdot \frac{\exp(l_i + g_i)}{\sum_{j=0}^{T_{max}-1} \exp(l_j + g_j)}. \tag{11}$$

Thus, the overall training loss $\mathcal{L}_{\text{total}}$ is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}}(\hat{I}_H, I_H) + \alpha \cdot \mathcal{L}_{\mathcal{G}}(\hat{z}_H) + \beta \cdot \mathcal{L}_{qf}. \tag{12}$$

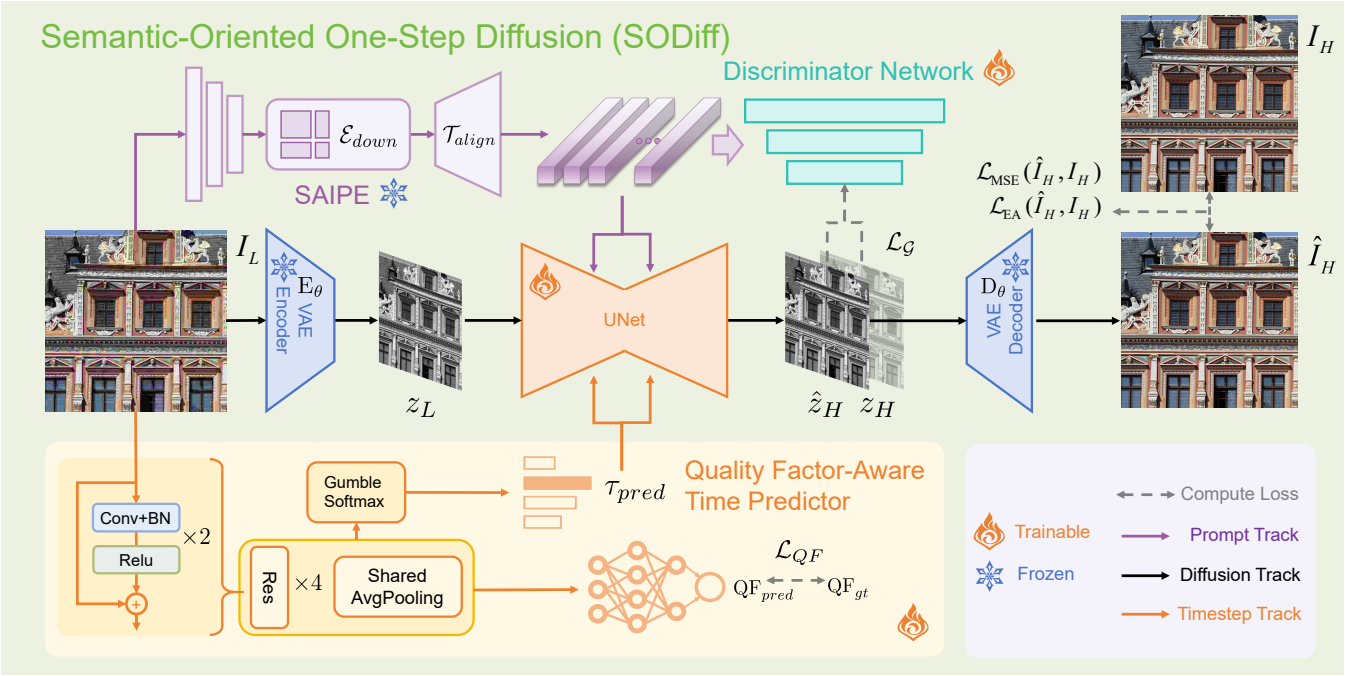


Figure 4: Training Framework of SODiff. First, the Prompt Track uses the first-stage trained SAIPE to extract semantic-oriented embedded image guidance e_{img} from I_L. Second, the Timestep Track performs OSD timestep and QF prediction, obtaining τ_{pred} and computing $\mathcal{L}_{qf}$. Next, I_L is encoded into latent space through the frozen VAE encoder to obtain z_L, with e_{img} and τ_{pred} providing conditional control for the noise prediction UNet, which performs one-step noise prediction to obtain $\hat{z}_H$. The SDXL Discriminator receives the transformed e_{img}, $\hat{z}_H$, and the latent space representation z_H of I_H to compute $\mathcal{L}_{\mathcal{G}}$. Last, $\hat{z}_H$ passes through the frozen VAE Decoder to obtain the reconstructed image $\hat{I}_H$, which computes $\mathcal{L}_{recon}$ with I_H.

4 Experiments

4.1 Experimental Settings

Training and Testing Datasets. Our model is trained on the DF2K (Agustsson and Timofte 2017; Timofte, Agustsson, and Gool 2017), LSDIR (Li et al. 2023) datasets. For preprocessing, we randomly crop DF2K and LSDIR to a size of 512×512. We randomly select JPEG compression with quality factors in the range [5,95] as the degradation pipeline. For the test set, we follow previous work (Guo et al. 2025) and select LIVE-1 (Sheikh et al. 2005), Urban100 (Huang, Singh, and Ahuja 2015), and DIV2K-val (Agustsson and Timofte 2017), conducting evaluations at QF values of 5, 10, and 20, respectively.

Evaluation Metrics. To evaluate the model's performance, we selected both full-reference and no-reference image quality assessment (IQA) to assess the test results. The full-reference IQA includes LPIPS (Zhang et al. 2018a) and DISTS (Ding et al. 2020). The no-reference IQA are MUSIQ (Ke et al. 2021), MANIQA (abbreviated as M-IQA) (Yang et al. 2022), and CLIPIQA (abbreviated as C-IQA) (Wang, Chan, and Loy 2023), which provide evaluation that aligns with natural human perceptual judgment.

Implementation Details. The overall training process consists of two stages: the training of SAIPE and SODiff. In the first stage, to improve training efficiency, all text descriptions are pre-generated before the training stage, and ensure they are generated from image content within the cropped regions. With λ_{align} of alignment loss in Eq. (5) set to 0.5, we use the Adam (Kingma and Ba 2017) optimizer with a learning rate of 2×10^{-4}. In the second stage, we select SD2.1-base (Stability AI 2022) as the base model. SODiff training freezes SAIPE and the VAE encoder and decoder, sets the LoRA rank to 16 to finetune SD, and uses AdamW as the optimizer with learning rate of 1×10^{-5}. Following D^{3}SR (Li et al. 2025), we use SDXL as our discriminator network and set α in Eq. (12) to 1×10^{-2}. β in Eq. (12) is set to 1×10^{-3}. With batch sizes of 16 and 2 for the two stages, training is performed on 4 and 2 NVIDIA RTX A6000 GPUs for 50k and 150k iterations, respectively.

Compared State-of-the-Art (SOTA) Methods. We compare SODiff with two categories of models. First, we compare with non-diffusion models, including FBCNN (Jiang, Zhang, and Timofte 2021), JDEC (Han et al. 2024), and PromptCIR (Li et al. 2024). Second, we compare with diffusion-based models, including DiffBIR (Lin et al. 2024) and SUPIR (Yu et al. 2024). Note that DiffBIR is retrained for JPEG compression artifacts removal.

4.2 Main Results

Quantitative Comparisons. Table. 1 presents the quantitative comparison results across three benchmark datasets (LIVE-1, Urban100, and DIV2K-val) under different qual-

LIVE-1

Method	LPIPS ↓			DISTS ↓			MUSIQ ↑			MANIQA ↑			CLIPIQA ↑		
	QF5	QF10	QF20	QF5	QF10	QF20	QF5	QF10	QF20	QF5	QF10	QF20	QF5	QF10	QF20
JPEG	0.4384	0.3013	0.1799	0.3242	0.2387	0.1653	40.33	53.88	64.12	0.2294	0.3509	0.4411	0.1716	0.2737	0.5542
FBCNN	0.3736	0.2503	0.1583	0.2353	0.1785	0.1319	63.56	71.00	73.96	0.3425	0.4207	0.4551	0.2763	0.4767	0.5535
JDEC	0.4113	0.2450	0.1555	0.2364	0.1740	0.1282	55.66	70.80	73.81	0.2002	0.4065	0.4433	0.1539	0.4811	0.5512
PromptCIR	0.3797	0.2290	0.1450	0.2334	0.1658	0.1223	60.34	72.39	**74.12**	0.2790	0.4500	0.4713	0.2655	0.5176	0.5847
DiffBIR*	0.3509	0.2160	0.1500	0.2035	0.1319	0.0988	58.09	67.38	71.08	0.2812	0.3789	0.4371	0.3776	0.5789	0.6814
SUPIR	0.4856	0.2770	0.1683	0.2720	0.1558	0.1121	52.69	68.77	73.02	0.3229	0.5183	**0.6237**	0.3149	0.6115	0.7364
SODiff (ours)	**0.2229**	**0.1605**	**0.1237**	**0.1173**	**0.0938**	**0.0763**	**72.88**	**73.84**	74.11	**0.4957**	**0.5192**	0.5272	**0.7087**	**0.7323**	**0.7587**

Urban100

Method	LPIPS ↓			DISTS ↓			MUSIQ ↑			MANIQA ↑			CLIPIQA ↑		
	QF5	QF10	QF20	QF5	QF10	QF20	QF5	QF10	QF20	QF5	QF10	QF20	QF5	QF10	QF20
JPEG	0.3481	0.2254	0.1244	0.2834	0.2145	0.1521	50.46	60.87	67.60	0.3656	0.4401	0.4967	0.2806	0.3517	0.5343
FBCNN	0.2341	0.1462	0.0896	0.2162	0.1648	0.1249	69.03	72.55	73.39	0.4263	0.5033	0.5288	0.3800	0.5014	0.5437
JDEC	0.2794	0.1382	0.0846	0.2309	0.1570	0.1175	62.97	72.52	73.30	0.3386	0.5001	0.5230	0.2518	0.4959	0.5369
PromptCIR	0.2389	0.1183	**0.0739**	0.2037	0.1431	0.1083	66.08	**73.01**	**73.47**	0.3946	0.5380	0.5489	0.3619	0.5337	0.5662
DiffBIR*	0.2018	0.1344	0.1005	0.1657	0.1207	0.0939	69.63	71.77	72.51	0.4285	0.4813	0.5105	0.5470	0.5966	0.6306
SUPIR	0.3279	0.2489	0.2125	0.2018	0.1659	0.1518	69.94	72.37	73.01	0.5546	**0.5995**	**0.6105**	0.5536	0.6178	0.6397
SODiff (ours)	**0.1579**	**0.1098**	0.0846	**0.1196**	**0.0914**	**0.0734**	**71.33**	72.23	72.63	**0.5598**	0.5451	0.5561	**0.6392**	**0.6558**	**0.6733**

DIV2K-val

Method	LPIPS ↓			DISTS ↓			MUSIQ ↑			MANIQA ↑			CLIPIQA ↑		
	QF5	QF10	QF20	QF5	QF10	QF20	QF5	QF10	QF20	QF5	QF10	QF20	QF5	QF10	QF20
JPEG	0.3459	0.3234	0.2072	0.2570	0.2255	0.1465	25.95	47.53	57.45	0.2570	0.3120	0.3557	0.2595	0.3303	0.5072
FBCNN	0.3445	0.2448	0.1733	0.2078	0.1581	0.1168	56.52	61.79	65.20	0.3025	0.3593	0.3775	0.3004	0.4561	0.5221
JDEC	0.3811	0.2313	0.1565	0.2234	0.1574	0.1152	53.88	**67.48**	**69.90**	0.2118	0.3689	0.3927	0.1841	0.4675	0.5319
PromptCIR	0.3549	0.2240	0.1581	0.2067	0.1459	0.1061	52.21	62.63	65.62	0.2705	0.3758	0.3871	0.3041	0.4956	0.5483
DiffBIR*	0.2788	0.1953	0.1542	0.1533	0.1072	0.0856	60.21	65.22	67.06	0.3220	0.3754	**0.4033**	0.4975	0.5912	0.6355
SUPIR	0.4372	0.3121	0.2295	0.2148	0.1410	0.1161	54.07	61.93	64.87	0.3438	0.3570	0.3723	0.4219	0.5186	0.5535
SODiff (ours)	**0.2425**	**0.1732**	**0.1295**	**0.1126**	**0.0816**	**0.0622**	**64.42**	65.90	66.49	**0.3733**	**0.3924**	0.3984	**0.5851**	**0.6193**	**0.6398**

Table 1: Performance comparisons across datasets and different QF. The best and second-best results are shown in bold and underlined, respectively. Model with asterisk (*) denotes that it is retrained for JPEG compression artifacts removal.

Method	Urban100			DIV2K-Val		
	DISTS↓	MUSIQ↑	M-IQA↑	DISTS↓	MUSIQ↑	M-IQA↑
No Align	0.1261	64.41	0.4609	0.1071	62.83	0.3250
DAPE	0.0877	71.53	0.4953	**0.0697**	64.39	0.3652
SAIPE	**0.0862**	**72.51**	**0.5531**	0.0731	**66.72**	**0.4013**

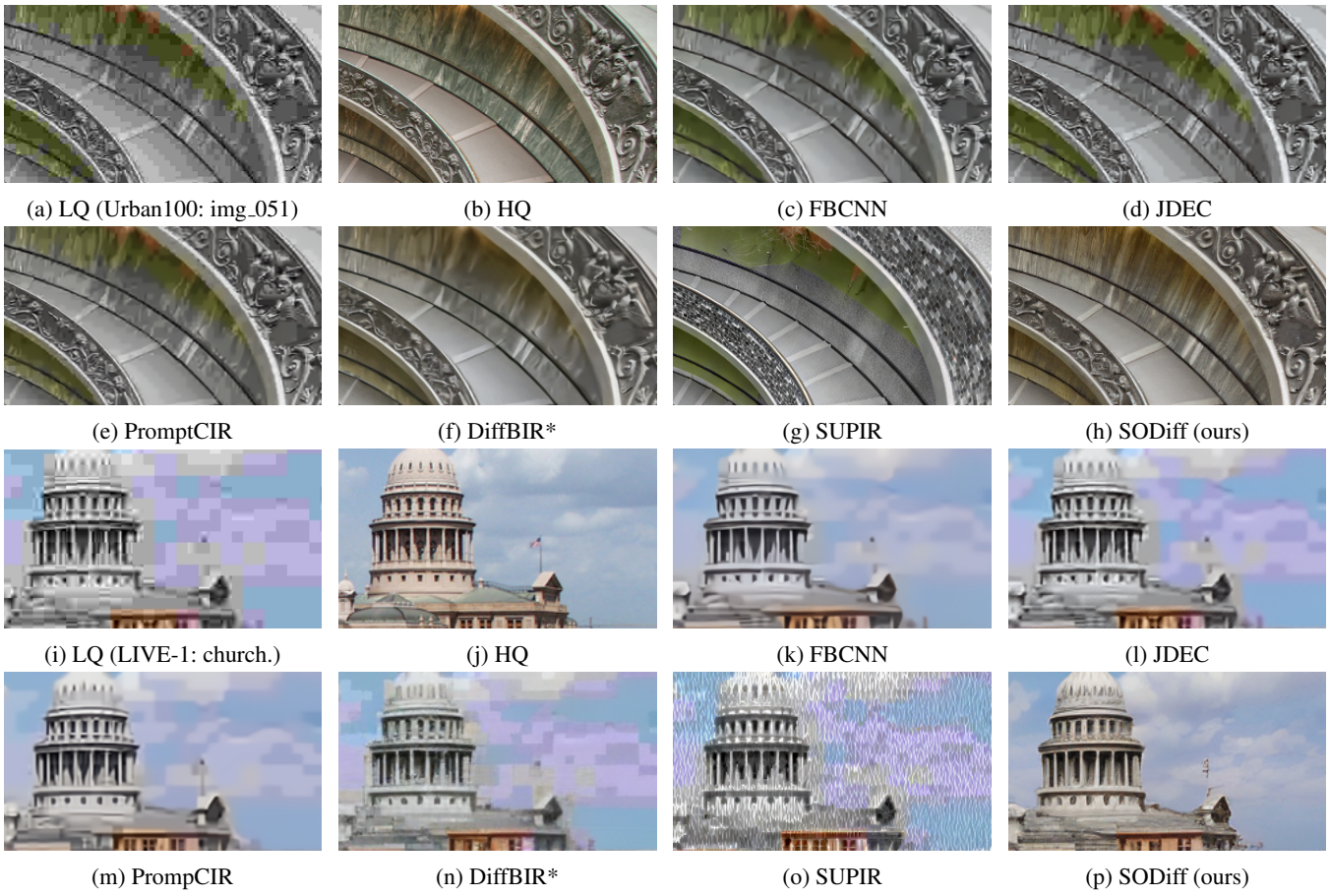
Table 2: Ablation study of prompt methods on Urban100 and DIV2K-val datasets. The best results are shown in bold.

ity factors (QF). Our method achieves the best results in perceptual quality metrics and competitive performance in no-reference quality measures across all evaluation scenarios. The performance advantage becomes more pronounced at lower quality factors, indicating SODiff's robustness in handling severe compression artifacts and validating the effectiveness of our semantic-oriented approach.

Qualitative Comparisons. Figure. 5 shows SODiff excels under severe compression. Non-diffusion methods (FBCNN, JDEC, PromptCIR) fail to handle severe artifacts, producing blurred textures and color blocking. DiffBIR and SUPIR remain affected: compression artifacts misguide diffusion, causing texture distortion (Fig. 5g). Despite retraining, DiffBIR still suffers from severe color blocking and inferior texture details compared to SODiff. Our model leverages semantic and texture cues to restore structures from degraded color blocks (e.g., clouds in Fig. 5p). More visual comparisons can be found in the supplementary materials.

4.3 Ablation Studies

The Effectiveness of SAIPE. To verify SAIPE's effectiveness, we compare different prompt methods. We employ the degradation-aware prompt extractor (DAPE) (Wu et al. 2024b) for text extraction comparison rather than LLaVA, which generates more detailed descriptions but has impractical inference time. We also test an alternative strategy where only the reconstruction branch was trained first, followed by joint training of $\mathcal{T}_{align}$ with the diffusion model. Results in Tab. 2 demonstrate that SAIPE achieves better restoration guidance than pure text prompts and outperforms image guidance under pure reconstruction guidance.



(a) LQ (Urban100: img_051) (b) HQ (c) FBCNN (d) JDEC

(e) PromptCIR (f) DiffBIR* (g) SUPIR (h) SODiff (ours)

(i) LQ (LIVE-1: church.) (j) HQ (k) FBCNN (l) JDEC

(m) PrompCIR (n) DiffBIR* (o) SUPIR (p) SODiff (ours)

Figure 5: Visual comparison (QF=5) on Urban100 and LIVE-1 datasets. Please zoom in for a better view.

Method	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑
w/o TP	0.3779	0.2577	56.1229	0.3236	0.4105
w/o $\mathcal{L}_{qf}$	0.3270	0.1611	58.7188	0.3232	0.4616
Full model	**0.3124**	**0.1469**	**60.3954**	**0.3292**	**0.5181**

Table 3: Ablation studies on different components. The best results are shown in bold. TP represents timestep predictor.

$\mathcal{L}_{\text{EA}}$	$\mathcal{L}_{\mathcal{G}}$	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑
	✓	0.3369	0.2475	56.1691	0.3117	0.2745
✓		0.3602	0.1776	58.1101	0.3250	0.5022
✓	✓	**0.2102**	**0.1558**	**69.5945**	**0.4939**	**0.5901**

Table 4: Ablation studies on different loss functions. The best results are shown in bold.

Timestep Predictor. We verify the impact of adopting quality factor (QF) aware timestep predictor for timestep prediction on SODiff, and test the results without implicitly learning QF during training. Table. 3 demonstrates that with the assistance of QF-aware timestep predictor, the model achieves optimal performance. Moreover, it can be observed that implicit QF learning helps improve all metrics.

Training Loss Functions. To evaluate the individual contributions of training loss functions, we conduct experiments by excluding each component. As shown in Tab. 4, excluding $\mathcal{L}_{\text{EA}}$ leads to constrained performance across various metrics, indicating the importance of perceptual guidance. Conversely, omitting $\mathcal{L}_{\mathcal{G}}$ yields superior performance on perceptual quality but suboptimal on other criteria.

5 Conclusion and Limitation

We propose SODiff, a novel one-step diffusion model that extracts semantically rich features with generative priors from over-compressed images to effectively guide the restoration process. It identifies compression severity and selects appropriate timesteps, recovering missing structures and blurred details under extreme compression conditions.

However, when confronting severe chroma subsampling in extreme cases, SODiff remains susceptible to color shifts despite recovering textural details. This issue will be discussed in the supplementary material. Additionally, separate training of SAIPE (trained from scratch) and diffusion components (fine-tuned from pre-trained weights) is required due to convergence instability, reducing training efficiency.

Acknowledgements

This work was supported by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102) and the Fundamental Research Funds for the Central Universities.

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *CVPRW*.

Choromanski, K. M.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J. Q.; Mohiuddin, A.; Kaiser, L.; Belanger, D. B.; Colwell, L. J.; and Weller, A. 2021. Rethinking Attention with Performers. In *ICLR*.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2020. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE TPAMI*.

Dong, C.; Deng, Y.; Loy, C. C.; and Tang, X. 2015. Compression artifacts reduction by a deep convolutional network. In *ICCV*.

Galteri, L.; Seidenari, L.; Bertini, M.; and Del Bimbo, A. 2017. Deep universal generative adversarial compression artifact removal. In *ICCV*.

Gong, J.; Wang, J.; Chen, Z.; Liu, X.; Gu, H.; Zhang, Y.; and Yang, X. 2025a. Human Body Restoration with One-Step Diffusion Model and A New Benchmark. In *ICML*.

Gong, J.; Yang, T.; Wang, J.; Chen, Z.; Liu, X.; Gu, H.; Liu, Y.; Zhang, Y.; and Yang, X. 2025b. HAODiff: Human-Aware One-Step Diffusion via Dual-Prompt Guidance. *arXiv preprint 2505.19742*.

Guo, J.; and Chao, H. 2016. Building dual-domain representations for compression artifacts reduction. In *ECCV*.

Guo, J.; Chen, Z.; Li, W.; Guo, Y.; and Zhang, Y. 2025. Compression-Aware One-Step Diffusion Model for JPEG Artifact Removal. *arXiv preprint arXiv:2502.09873*.

Han, W. K.; Im, S.; Kim, J.; and Jin, K. H. 2024. Jdec: Jpeg decoding via enhanced continuous cosine coefficients. In *CVPR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*.

Huang, J.-B.; Singh, A.; and Ahuja, N. 2015. Single Image Super-Resolution From Transformed Self-Exemplars. In *CVPR*.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *ICLR*.

Jiang, J.; Zhang, K.; and Timofte, R. 2021. Towards flexible blind jpeg artifacts removal. In *ICCV*.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-scale Image Quality Transformer . In *ICCV*.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Li, B.; Li, X.; Lu, Y.; Feng, R.; Guo, M.; Zhao, S.; Zhang, L.; and Chen, Z. 2024. PromptCIR: Blind Compressed Image Restoration with Prompt Learning. *arXiv preprint arXiv:2404.17433*.

Li, J.; Cao, J.; Zou, Z.; Su, X.; Yuan, X.; Zhang, Y.; Guo, Y.; and Yang, X. 2025. Unleashing the Power of One-Step Diffusion based Image Super-Resolution via a Large-Scale Diffusion Discriminator. *arXiv preprint arXiv:2410.04224*.

Li, Y.; Zhang, K.; Liang, J.; Cao, J.; Liu, C.; Gong, R.; Zhang, Y.; Tang, H.; Liu, Y.; Demandolx, D.; Ranjan, R.; Timofte, R.; and Van Gool, L. 2023. LSDIR: A Large Scale Dataset for Image Restoration. In *CVPRW*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. SwinIR: Image Restoration Using Swin Transformer. In *CVPRW*.

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Dai, B.; Yu, F.; Ouyang, W.; Qiao, Y.; and Dong, C. 2024. DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior. In *ECCV*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.

Luo, Z.; Gustafsson, F. K.; Zhao, Z.; Sjölund, J.; and Schön, T. B. 2023. Image Restoration with Mean-Reverting Stochastic Differential Equations. In *ICML*.

McInnes, L.; Healy, J.; and Melville, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*.

Sheikh, H.; Wang, Z.; Cormack, L.; and Bovik, A. 2005. LIVE Image Quality Assessment Database Release 2. http://live.ece.utexas.edu/research/quality.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *ICLR*.

Stability AI. 2022. stabilityai/stable-diffusion-2-1-base. [Online]. Available: https://huggingface.co/stabilityai/stable-diffusion-2-1-base.

Timofte, R.; Agustsson, E.; and Gool. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results. In *CVPRW*.

Wallace, G. K. 1991. The JPEG still picture compression standard. *Communications of the ACM*, 34(4): 30–44.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *AAAI*.

Wang, J.; Fan, Q.; Zhang, Q.; Liu, H.; Yu, Y.; Chen, J.; and Ren, W. 2024a. Hero-SR: One-Step Diffusion for Super-Resolution with Human Perception Priors. *arXiv preprint arXiv:2412.07152*.

Wang, J.; Gong, J.; Zhang, L.; Chen, Z.; Liu, X.; Gu, H.; Liu, Y.; Zhang, Y.; and Yang, X. 2025. One-Step Diffusion Model for Face Restoration. In *CVPR*.

Wang, M.; Fu, X.; Sun, Z.; and Zha, Z.-J. 2021. Jpeg artifacts removal via compression quality ranker-guided networks. In *IJCAI*.

Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.-P.; Liu, Z.; Qiao, Y.; Kot, A. C.; and Wen, B. 2024b. SinSR: diffusion-based image super-resolution in a single step. In *CVPR*.

Wu, R.; Sun, L.; Ma, Z.; and Zhang, L. 2024a. One-Step Effective Diffusion Network for Real-World Image Super-Resolution. In *NeurIPS*.

Wu, R.; Yang, T.; Sun, L.; Zhang, Z.; Li, S.; and Zhang, L. 2024b. SeeSR: Towards semantics-aware real-world image super-resolution. In *CVPR*.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment. In *CVPRW*.

Yu, F.; Gu, J.; Li, Z.; Hu, J.; Kong, X.; Wang, X.; He, J.; Qiao, Y.; and Dong, C. 2024. Scaling Up to Excellence: Practicing Model Scaling for Photo-Realistic Image Restoration In the Wild. In *CVPR*.

Zhang, A.; Yue, Z.; Pei, R.; Ren, W.; and Cao, X. 2024. Degradation-Guided One-Step Image Super-Resolution with Diffusion Priors. *arxiv*.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. In *ICCV*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Zhang, X.; Yang, W.; Hu, Y.; and Liu, J. 2018b. Dmcnn: Dual-domain multi-scale convolutional neural network for compression artifacts removal. In *ICIP*.

Zhang, Y.; Li, K.; Li, K.; Zhong, B.; and Fu, Y. 2019. Residual Non-Local Attention Networks for Image Restoration. *arXiv preprint arXiv:1903.10082*.