

Compression-Aware One-Step Diffusion Model for JPEG Artifact Removal

Jinpei Guo^{1,2}, Zheng Chen², Wenbo Li³, Yong Guo⁴, Yulun Zhang^{2,*}

¹Carnegie Mellon University, ²Shanghai Jiao Tong University,

³Huawei Noah’s Ark Lab, ⁴South China University of Technology

Abstract

Diffusion models have demonstrated remarkable success in image restoration tasks. However, their multi-step denoising process introduces significant computational overhead, limiting their practical deployment. Furthermore, existing methods struggle to effectively remove severe JPEG artifact, especially in highly compressed images. To address these challenges, we propose CODiff, a compression-aware one-step diffusion model for JPEG artifact removal. The core of CODiff is the compression-aware visual embedder (CaVE), which extracts and leverages JPEG compression priors to guide the diffusion model. Moreover, We propose a dual learning strategy for CaVE, which combines explicit and implicit learning. Specifically, explicit learning enforces a quality prediction objective to differentiate low-quality images with different compression levels. Implicit learning employs a reconstruction objective that enhances the model’s generalization. This dual learning allows for a deeper understanding of JPEG compression. Experimental results demonstrate that CODiff surpasses recent leading methods in both quantitative and visual quality metrics. The code and models will be publicly available.

1. Introduction

JPEG [37] artifact removal task aims to remove the artifact caused by the compression algorithm and recover the lost information from the compressed images. Recent advancements have focused on leveraging CNN-based and Transformer-based methods [5, 15, 20, 24, 54] to remove JPEG artifact for compressed images with different quality factors (QFs), achieving remarkable performance. However, these methods (see Fig. 1) face significant challenges at high compression levels, due to substantial compression artifact and severe visual information loss.

Recent advances in diffusion models [2, 12, 33, 34], particularly large-scale pre-trained text-to-image (T2I) models [28, 29, 31], have strong image generation priors. These priors make diffusion models a promising solution for compression artifact removal, especially in cases of severe degradation. However, their multi-step denoising pro-



Figure 1. Visual comparison of JPEG artifact removal (QF=5). We provide multiply accumulate operations (MACs) and time during inference. DiffBIR* [25], OSEDiff* [43] are retrained on the same dataset as our method. All models are tested with an input image size of $1,024 \times 1,024$. Our CODiff reconstructs a more realistic and faithful visual result than competing methods.

cess introduces substantial computational overhead. To address this, one-step diffusion (OSD) models [22, 43] have emerged as efficient alternatives. By leveraging large-scale pre-trained multi-step T2I diffusion models [28, 29], these OSD models strike a balance between strong restoration capabilities and significantly faster inference, making them a compelling choice for JPEG artifact removal.

As illustrated in Fig. 1, for highly compressed images (e.g., QF=5), diffusion-based models generally outperform traditional CNN-based and Transformer-based methods in reducing JPEG artifact such as blocky patterns and color banding. However, the restored images remain inharmonious, with certain artifact still visible. The primary issue lies in the insufficient integration of JPEG compression priors. Without this information, models struggle to differentiate between compression-induced distortions and natural image features, leading to suboptimal restoration quality. Notably, most existing diffusion-based methods [25, 43, 44, 49] for image restoration tasks neglect such compression-related information.

Thus, it is worthwhile to investigate how to extract JPEG compression priors and integrate them into the diffusion model. A major challenge is how to design effective JPEG compression representations. Previous works [6, 11] use the quantization matrix [37] to represent the compression priors. However, since it only depends on the quality factor

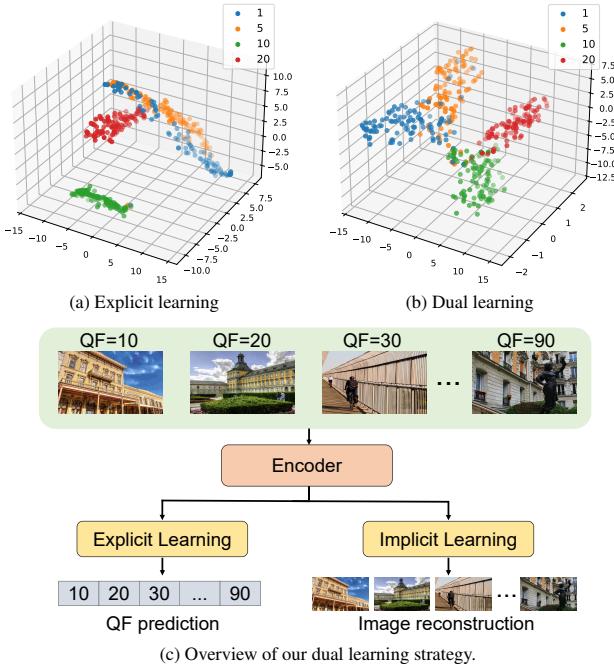


Figure 2. An illustration of the explicit learning approach with our dual learning strategy. We employ the t-SNE [36] method to visualize the clustering of output compression quality representations on the Urban100 [14] dataset, where QF=1 and QF=5 are unseen compression levels. Notably, the explicit learning method struggles to distinguish between these unseen QFs, whereas our dual learning strategy effectively differentiates them.

(QF) with static numerical values, the information it provides is insufficient. To dynamically utilize the JPEG priors, QF learning approaches [15, 40] attempt to learn compression representations. They train models to predict QF from low-quality (LQ) images and use intermediate visual embeddings as priors. However, as QF is merely a single integer, taking it as the sole learning objective constrains the model’s ability to capture the comprehensive JPEG priors. Furthermore, the models may encounter generalization challenges when exposed to unseen compression levels.

To alleviate these challenges, we propose CODiff, a compression-aware **one-step diffusion** model for JPEG artifact removal. CODiff leverages the powerful pre-trained knowledge of multi-step diffusion (MSD) models while significantly improving inference speed. A key component of CODiff is the compression-aware visual embedder (CaVE), which extracts JPEG compression priors to guide the denoising process. To provide more informative JPEG compression priors and improve generalization, CaVE learns the JPEG compression process through a dual-learning strategy (see Fig. 2). In the explicit learning phase, it is trained to predict the quality factor (QF) from a low-quality image. In the implicit learning phase, CaVE learns to restore high-quality (HQ) images from compressed inputs. This joint learning approach enhances CaVE’s ability

to model the JPEG compression process, thereby enhancing its generalization ability to differentiate unseen compression levels. Comparison in Fig. 1 highlights the superior restoration quality and efficiency of our CODiff. Our main contributions are summarized as follows:

- We propose CODiff, a novel and effective OSD model for JPEG artifact removal. To the best of our knowledge, this is the first attempt to design an OSD model specifically for this task with compression priors.
- We design compression-aware visual embedder (CaVE), which generates embeddings that capture rich JPEG compression information. CaVE provides effective compression priors for the OSD model to restore LQ images.
- We propose dual learning strategy for CaVE. Specifically, explicit learning allows CaVE to distinguish LQ images across various compression levels. While implicit learning enhances CaVE’s generalization capabilities.
- Our CODiff achieves SOTA performance in JPEG artifact removal, excelling in both quantitative metrics and visual quality. CODiff reduces computational costs compared with both MSD and OSD models.

2. Related Work

2.1. JPEG Artifact Removal

In recent years, learning-based methods have significantly advanced JPEG artifact removal. The pioneering work AR-CNN [5] first introduces deep learning into this task, leveraging a super-resolution network [4] to reduce compression artifact. Transformer-based methods [24, 54] incorporate attention mechanisms to enhance feature representation, achieving strong performance across image artifact removal tasks. Inspired by the success of GANs in image restoration, several GAN-based methods [7, 8] have been proposed to enhance the perceptual quality of compressed images. Meanwhile, dual-domain convolutional network approaches [10, 11, 18, 53, 55] have been developed to exploit redundancies in both the pixel and frequency domains.

To further integrate the JPEG compression priors, the ranker-guided framework [40] utilizes compression quality ranking as auxiliary information. Quantization tables [6, 21] are also utilized as prior knowledge, enabling a single network to handle artifact across different quality factors (QFs). To enable blind JPEG artifact removal, FBCNN [15] predicts an adjustable QF to balance artifact removal and detail preservation. Additionally, Prompt-CIR [20] explores prompt learning for blind compressed image restoration. However, most existing methods struggle to recover highly compressed images due to severe information loss. With the advancement of diffusion models, incorporating JPEG compression priors into the pre-trained large scale T2I diffusion models offers a promising solution to mitigate this information loss.

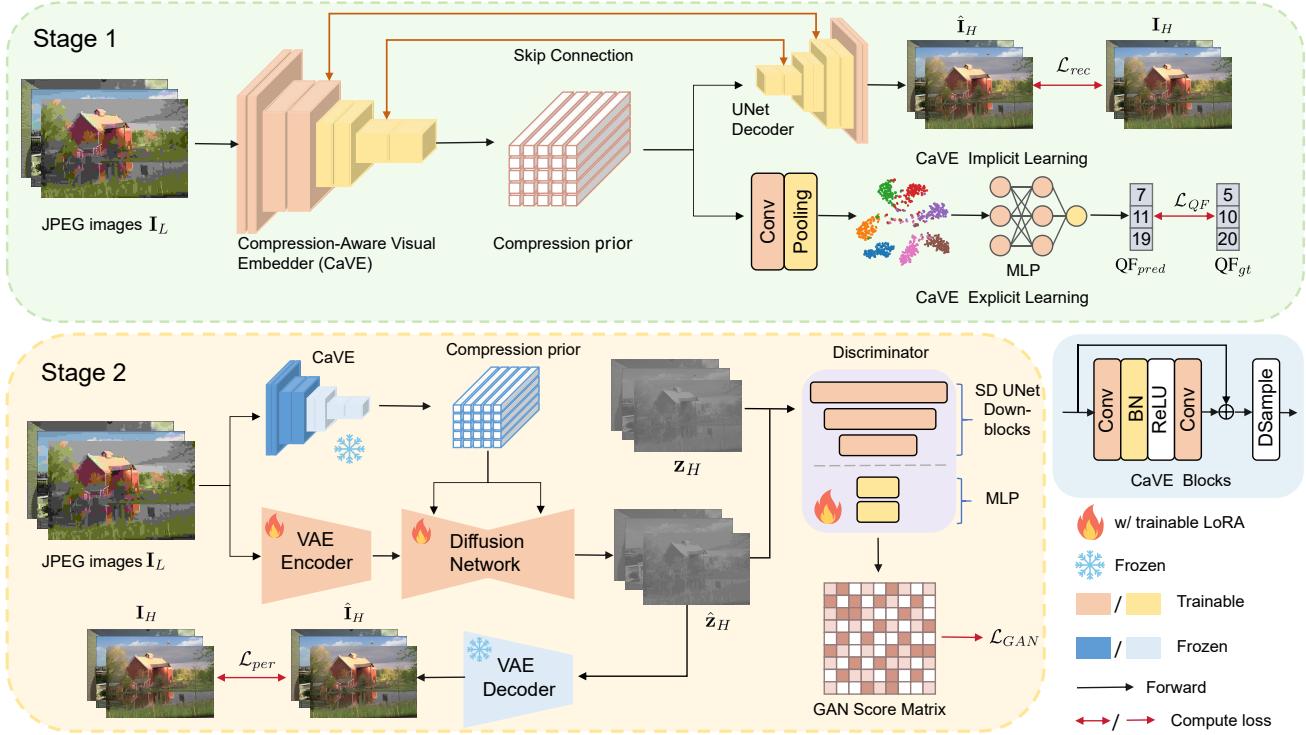


Figure 3. Overview of our proposed CODiff. In the first stage, we train our compression-aware visual embedder (CaVE) via a dual learning strategy. In implicit learning, compression prior embeddings are fed into a UNet decoder to reconstruct high-quality (HQ) images. In explicit learning, they are input into a lightweight quality factor (QF) predictor. In the second stage, the priors from CaVE are then used by the generator \mathcal{G}_θ to restore the HQ images: $\hat{\mathbf{I}}_H = \mathcal{G}_\theta(\mathbf{I}_L; \mathbf{c}_L)$. The generator \mathcal{G}_θ integrates a pre-trained VAE and UNet from StableDiffusion [29], with the VAE encoder and UNet fine-tuned via LoRA [13].

2.2. Diffusion Models

Diffusion models, known for their powerful generative capabilities, progressively transform random noise into structured data through iterative denoising. Recently, they have shown strong performance in image-to-image tasks, particularly in image restoration [16, 25, 39, 44, 46, 49, 50]. By leveraging their ability to capture fine-grained details and produce high-quality outputs, diffusion models have outperformed traditional image restoration methods. However, their complex architectures [25, 46, 49] and reliance on numerous iterative steps hinder their real-world deployment due to high computational costs.

Accelerating diffusion models by reducing inference steps is crucial for practical applications. However, excessive reduction often degrades performance. Therefore, most one-step diffusion (OSD) methods use distillation to learn from a teacher model, preserving image fidelity [42, 47, 48]. Notably, OSEDiff [43] has obtained promising results using variational score distillation. Despite these advances, existing diffusion-based restoration models often overlook degradation-related priors. Consequently, it is difficult to distinguish compression artifact from natural image features. Thus, exploring how to extract prior knowledge specific to JPEG compression is essential to guide diffusion models for JPEG artifact removal tasks.

3. Methodology

3.1. Preliminaries: One-Step Diffusion Model

Latent diffusion models [29] operate through a two-stage process: forward diffusion and reverse denoising. During the forward phase, Gaussian noise with variance $\beta_t \in (0, 1)$ is progressively added to the latent representation \mathbf{z} : $\mathbf{z}_t = \sqrt{\bar{\alpha}_t} \mathbf{z} + \sqrt{1 - \bar{\alpha}_t} \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$. Here, the term $\bar{\alpha}_t$ is defined as $\bar{\alpha}_t = \prod_{i=1}^t (1 - \beta_i)$. In the reverse process, the clean latent vector $\hat{\mathbf{z}}_0$ is estimated from predicted noise $\hat{\varepsilon}$: $\hat{\mathbf{z}}_0 = \frac{\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t} \hat{\varepsilon}}{\sqrt{\bar{\alpha}_t}}$, where $\hat{\varepsilon}$ is the prediction of the model ε_θ given \mathbf{z}_t , t , and the prior \mathbf{c} : $\hat{\varepsilon} = \varepsilon_\theta(\mathbf{z}_t; \mathbf{c}, t)$.

Unlike conventional text-to-image (T2I) diffusion models [28, 29, 31], CODiff feeds the latent representation of the low-quality (LQ) image to the UNet instead of the Gaussian noise. Specifically, our CODiff first encodes the LQ image \mathbf{I}_L into the latent space via an encoder E_θ , resulting in $\mathbf{z}_L = E_\theta(\mathbf{I}_L)$. A single-step denoising operation is then performed to predict the noise $\hat{\varepsilon}$, which allows us to derive the estimated high-quality (HQ) latent representation $\hat{\mathbf{z}}_H$:

$$\hat{\mathbf{z}}_H = \frac{\mathbf{z}_L - \sqrt{1 - \bar{\alpha}_{T_L}} \varepsilon_\theta(\mathbf{z}_L; \mathbf{c}_L, T_L)}{\sqrt{\bar{\alpha}_{T_L}}}, \quad (1)$$

where ε_θ is the denoising network with learnable parameters θ , \mathbf{c}_L is the compression prior, and $T_L \in [0, T]$ is the predefined diffusion timestep.

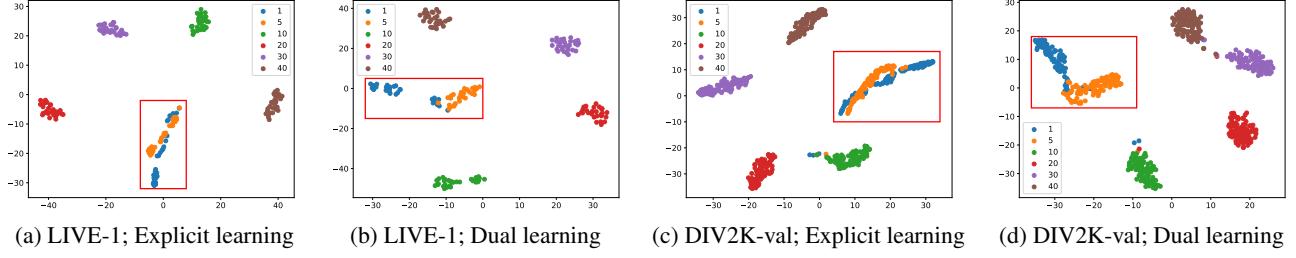


Figure 4. Visualization of JPEG prior embeddings from CaVE under different training objectives. In (a) and (c), CaVE is trained using only explicit learning, whereas in (b) and (d), it incorporates both explicit and implicit learning. The clusters enclosed in the red box correspond to unseen compression levels (QF=1,5). Notably, clusters from CaVE with dual learning separate from each other more clearly.

Finally, the decoder D_θ reconstructs HQ image $\hat{\mathbf{I}}_H$ from the latent representation $\hat{\mathbf{z}}_H$: $\hat{\mathbf{I}}_H = D_\theta(\hat{\mathbf{z}}_H)$. Defining the generator G_θ and the compression prior \mathbf{c}_L , the overall reconstruction from \mathbf{I}_L to $\hat{\mathbf{I}}_H$ can be compactly written as:

$$\hat{\mathbf{I}}_H = G_\theta(\mathbf{I}_L; \mathbf{c}_L). \quad (2)$$

3.2. Stage 1: Compression-aware Visual Embedder

Previous studies have shown that incorporating compression information into the reconstruction process can enhance performance. To incorporate JPEG compression priors into diffusion models, we first train a compression-aware visual embedder (CaVE) to capture these priors effectively. In the second stage, the pre-trained CaVE is leveraged to guide the diffusion model, enabling it to utilize compression-related features from low-quality (LQ) images to improve reconstruction.

3.2.1. Overall Architecture

Our CaVE consists of a UNet [30] architecture equipped with a lightweight quality factor (QF) predictor designed specifically for extracting compression priors. The UNet decoder and the QF predictor are utilized exclusively for the CaVE training. As illustrated in Fig. 3, in explicit learning, the QF predictor guides CaVE to explicitly learn the JPEG compression priors. Simultaneously, implicit learning helps CaVE further capture the relationship between compression and reconstruction, enhancing its ability to model the JPEG compression process.

As illustrated in Fig. 3, given a low-quality image $\mathbf{I}_L \in \mathbb{R}^{H \times W \times 3}$, the CaVE encodes it into a set of feature vectors:

$$\mathbf{c}_L = \{\mathbf{c}_{L_k} \in \mathbb{R}^d\}_{k=1}^K. \quad (3)$$

The QF predictor will then utilize the average feature representation, $\bar{\mathbf{c}}_L = \frac{1}{K} \sum_{k=1}^K \mathbf{c}_{L_k}$, to estimate QF. Meanwhile, the UNet decoder leverages the extracted feature vectors \mathbf{c}_L to reconstruct the corresponding high-quality image $\hat{\mathbf{I}}_H$.

3.2.2. Explicit Learning

To explicitly incorporate JPEG compression priors, we train CaVE to produce embeddings that directly facilitate QF prediction [15, 41]. Specifically, given a low-quality image, we employ the UNet encoder to encode the image into latent representations $\mathbf{c}_L = E(\mathbf{I}_L; \phi)$. The UNet encoder

contains four scales, each of which involves residual blocks. Each residual block consists of two 3×3 convolutional layers with Batch Normalization layer and ReLU activation function. 4×4 stride convolutions are employed for down-sampling operations. The output dimensions of each scale are set to 64, 128, 256, and 512, respectively.

We then use the embeddings \mathbf{z}_L from the UNet encoder to explicitly predict the JPEG compression level QF as $QF_{pred} = P(\bar{\mathbf{c}}_L; \phi)$, where $\bar{\mathbf{c}}_L$ is the average of latent embeddings \mathbf{z}_L . The QF predictor $P(\cdot; \phi)$ is a lightweight convolutional neural network (CNN) cascaded to a 3-layer multilayer perceptron (MLP). We set the channels in CNN and MLP hidden layers as 512 for a more accurate prediction.

During training, small patches may contain limited compression-related information (e.g., background regions in compressed images could correspond to multiple QFs). This ambiguity can introduce instability in the training process and hinder CaVE’s ability to effectively capture the underlying JPEG compression patterns [15]. Therefore, we use the \mathcal{L}_1 loss function to mitigate the negative impact of such outliers. Let B be the batch size during training, the QF prediction loss can be expressed as:

$$\mathcal{L}_{QF} = \frac{1}{B} \sum_{i=1}^B \|QF_{pred}^i - QF_{gt}^i\|_1. \quad (4)$$

To illustrate that our CaVE learns discriminative representations for images with different QFs, we visualize \mathbf{z}_L from the UNet encoder using the t-SNE [36] method, as shown in Fig. 4. Our results show that CaVE, when trained with the QF loss, effectively distinguishes compression qualities present in the training set. However, it struggles to differentiate unseen compression levels (e.g., QF = 1, 5), indicating that relying solely on the QF loss fails to capture the full spectrum of compression-related information.

3.2.3. Implicit Learning

To more comprehensively capture JPEG compression priors, we train CaVE to produce embeddings that not only predict the quality factor (QF) (explicit learning) but also reconstruct high-quality images (implicit learning). By incorporating a reconstruction objective, CaVE implicitly learns the intricate details of the JPEG compression process.

Specifically, given a low-quality image \mathbf{I}_L , we employ the UNet-based architecture [30] to restore the corresponding high-quality image: $\hat{\mathbf{I}}_H = D(\mathbf{z}_L; \phi)$. We use the same encoder in Sec. 3.2.2. The decoder, on the other hand, includes three scales, each of which contains residual blocks that consist of two 3×3 convolutional layers with Batch Normalization layer and ReLU activation function. Each scale receives the corresponding intermediate image features from the encoder to generate the restored image.

Given a batch of B training samples, the objective of implicit learning is to minimize the \mathcal{L}_1 loss between the reconstructed image $\hat{\mathbf{I}}_H$ and the ground-truth \mathbf{I}_H :

$$\mathcal{L}_{rec} = \frac{1}{B} \sum_{i=1}^B \|\hat{\mathbf{I}}_H^i - \mathbf{I}_H^i\|_1. \quad (5)$$

The overall CaVE training objective is formulated as:

$$\mathcal{L}_{CaVE} = \mathcal{L}_{QF} + \lambda \cdot \mathcal{L}_{rec}. \quad (6)$$

To demonstrate how implicit learning enhances CaVE’s understanding of JPEG compression, we visualize its learned embeddings in Fig. 4. Our results reveal that the dual learning allows CaVE to capture more nuanced variations in compression artifact, enabling it to differentiate even previously unseen compression levels.

3.3. Stage 2: JPEG Artifact Removal

After extracting compression prior embeddings, we integrate them into our OSD model and fine-tune the pre-trained diffusion model using LoRA [13]. The overall training objective combines perceptual and GAN losses. Specifically, perceptual loss directly aligns with the reconstruction process to ensure high-fidelity restoration. The GAN loss [9] enhances the realism of the generated images by refining textures and preserving natural image characteristics. By jointly optimizing these loss functions, our model effectively achieve high-quality, photorealistic reconstructions.

Perceptual Loss. While \mathcal{L}_2 loss is widely used for image reconstruction due to its simplicity and pixel-wise precision, it often struggles to capture perceptual quality. Minimizing \mathcal{L}_2 loss tends to produce overly smooth outputs, as it treats all pixel differences equally, ignoring structural and textural information critical for human perception.

To address this issue, we additionally incorporate DISTs (Deep Image Structure and Texture Similarity) loss [3], which goes beyond pixel-wise comparisons by measuring perceptual similarity between images. Unlike \mathcal{L}_2 loss, DISTs is designed to align with human visual perception, capturing both structural details and textural characteristics. The perceptual loss can be written as:

$$\mathcal{L}_{per} = \mathcal{L}_2(\hat{\mathbf{I}}_H, \mathbf{I}_H) + \lambda_D \mathcal{L}_{DISTs}(\hat{\mathbf{I}}_H, \mathbf{I}_H). \quad (7)$$

GAN Loss. Generating stable images remains a significant challenge for OSD models, primarily due to their constrained computational resources. Prior research [42, 43, 47, 48] has largely relied on distillation techniques to trans-

fer knowledge from multi-step diffusion (MSD) models. However, these approaches are inherently limited by the performance of their teacher models, restricting their potential for further improvement.

To overcome these limitations, we adopt an alternative strategy by integrating a discriminative network to enhance the realism of restoration. The GAN losses [9] used to train the generator G_θ and discriminator D_θ are defined as:

$$\mathcal{L}_G = -\mathbb{E}_t [\log \mathcal{D}_\theta(\hat{\mathbf{z}}_H)], \quad (8)$$

$$\begin{aligned} \mathcal{L}_D = & -\mathbb{E}_t [\log (1 - \mathcal{D}_\theta(\hat{\mathbf{z}}_H))] \\ & - \mathbb{E}_t [\log \mathcal{D}_\theta(\mathbf{z}_H)], \end{aligned} \quad (9)$$

where $\hat{\mathbf{z}}_H$ and \mathbf{z}_H are the latent embeddings of the reconstructed image and the ground-truth image, respectively.

In our JPEG artifact removal pipeline, we first train CaVE in stage 1, then we incorporate the perceptual and GAN losses to fine-tune our OSD with JPEG compression priors from CaVE. The overall training objective for the generator is formulated as:

$$\mathcal{L} = \mathcal{L}_{per} + \lambda_G \mathcal{L}_G. \quad (10)$$

4. Experiments

4.1. Experimental Settings

Training and Testing Datasets. Following the previous work [20, 43], We use DF2K which comprises 800 images from DIV2K [1] and 2,650 images from Flickr2K [35], and LSDIR [23] as our training datasets, totaling 88,441 different high-quality images. During training, we randomly crop 256×256 patches and dynamically synthesize low-quality and high-quality image pairs by sampling the quality factor from a range of 8 to 95. We evaluate CODiff and compare its performance against competing methods using LIVE-1 [32], Urban100 [14], and DIV2K-Val [1] datasets.

Evaluation Metrics. To ensure a comprehensive and holistic evaluation of different methods, we employ a diverse set of both full-reference and no-reference image quality metrics. For perceptual quality assessment with a reference image, we utilize LPIPS [52] and DISTs [3], both of which measure structural and perceptual similarity by leveraging deep feature representations. In addition, to assess image quality without requiring a reference, we adopt MANIQA [45], MUSIQ [17], and CLIPQA [38], which are designed to predict human-perceived image quality.

Implementation Details. We use PyTorch [27] to implement our model. In the first stage (Sec. 3.2), we optimize CaVE using the Adam optimizer [19] with a learning rate of 2×10^{-5} and a batch size of $B=128$. The weight for implicit compression prior learning, λ , is set to 1,000. We train CaVE for 200K iterations with 4 NVIDIA A6000 GPUs.

In the second stage (Sec. 3.3), we train CODiff using the AdamW optimizer [26] with a learning rate of 5×10^{-5} , weight decay of 1×10^{-5} , and a batch size of 32 for both the

Methods	QF=5					QF=10					QF=20				
	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑
JPEG [37]	0.4384	0.3242	40.33	0.2294	0.1716	0.3013	0.2387	53.88	0.3509	0.2737	0.1799	0.1653	64.12	0.4411	0.5542
FBCNN [15]	0.3736	0.2353	63.56	0.3425	0.2763	0.2503	0.1785	71.00	0.4207	0.4767	0.1583	0.1319	73.96	0.4551	0.5535
JDEC [11]	0.4113	0.2364	55.66	0.2002	0.1539	0.2450	0.1740	70.80	0.4065	0.4811	0.1555	0.1282	73.81	0.4433	0.5512
PromptCIR [20]	0.3797	0.2334	60.34	0.2790	0.2655	0.2290	0.1658	72.39	0.4500	0.5176	0.1450	0.1223	74.12	0.4713	0.5847
DiffBIR* [25] (s=50)	0.3509	0.2035	58.09	0.2812	0.3776	0.2160	0.1319	67.38	0.3789	0.5789	0.1500	0.0988	71.08	0.4371	0.6814
SUPIR [49] (s=50)	0.4856	0.2720	52.69	0.3229	0.3149	0.2770	0.1558	68.77	0.5183	0.6115	0.1683	0.1121	73.02	0.6237	0.7364
OSEDiff* [43] (s=1)	0.2675	0.1653	65.51	0.3417	0.5623	0.1749	0.1164	71.23	0.3963	0.7022	0.1270	0.0856	72.70	0.4219	0.7260
CODiff (ours, s=1)	0.2062	0.1121	73.16	0.5321	0.7212	0.1428	0.0867	74.39	0.5438	0.7559	0.1101	0.0692	74.34	0.5323	0.7565

(a) Quantitative comparison on the LIVE-1 dataset

Methods	QF=5					QF=10					QF=20				
	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑
JPEG [37]	0.3481	0.2834	50.46	0.3656	0.2806	0.2254	0.2145	60.87	0.4401	0.3517	0.1244	0.1521	67.60	0.4967	0.5343
FBCNN [15]	0.2341	0.2162	69.03	0.4263	0.3800	0.1462	0.1648	72.55	0.5033	0.5014	0.0896	0.1249	73.39	0.5288	0.5437
JDEC [11]	0.2794	0.2309	62.97	0.3386	0.2518	0.1382	0.1570	72.52	0.5001	0.4959	0.0846	0.1175	73.30	0.5230	0.5369
PromptCIR [20]	0.2389	0.2037	66.08	0.3946	0.3619	0.1183	0.1431	73.01	0.5380	0.5337	0.0739	0.1083	73.47	0.5489	0.5662
DiffBIR* [25] (s=50)	0.2018	0.1657	69.63	0.4285	0.5470	0.1344	0.1207	71.77	0.4813	0.5966	0.1005	0.0939	72.51	0.5105	0.6306
SUPIR [49] (s=50)	0.3279	0.2018	69.94	0.5546	0.5536	0.2489	0.1659	72.57	0.5995	0.6178	0.2125	0.1518	73.01	0.6105	0.6397
OSEDiff* [43] (s=1)	0.1959	0.1690	68.60	0.4491	0.5710	0.1262	0.1168	71.55	0.4927	0.6364	0.0911	0.0860	72.43	0.5082	0.6591
CODiff (ours, s=1)	0.1407	0.1101	72.16	0.5693	0.6741	0.0974	0.0842	72.61	0.5725	0.6824	0.0753	0.0667	72.63	0.5694	0.6830

(b) Quantitative comparison on the Urban100 dataset

Methods	QF=5					QF=10					QF=20				
	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑
JPEG [37]	0.4466	0.3183	34.59	0.2570	0.2595	0.3234	0.2255	47.53	0.3120	0.3303	0.2072	0.1465	57.45	0.3557	0.5072
FBCNN [15]	0.3445	0.2078	56.52	0.3025	0.3004	0.2448	0.1581	61.79	0.3593	0.4561	0.1733	0.1168	65.20	0.3775	0.5221
JDEC [11]	0.3811	0.2234	53.88	0.2118	0.1841	0.2313	0.1574	67.48	0.3689	0.4675	0.1565	0.1152	69.90	0.3927	0.5319
PromptCIR [20]	0.3549	0.2067	52.21	0.2705	0.3041	0.2240	0.1459	62.63	0.3758	0.4956	0.1581	0.1061	65.62	0.3871	0.5483
DiffBIR* [25] (s=50)	0.2788	0.1533	60.21	0.3220	0.4975	0.1953	0.1072	65.22	0.3754	0.5912	0.1542	0.0856	67.06	0.4033	0.6355
SUPIR [49] (s=50)	0.4372	0.2148	54.07	0.3438	0.4219	0.3121	0.1410	61.93	0.3570	0.5186	0.2295	0.1161	64.87	0.3723	0.5535
OSEDiff* [43] (s=1)	0.2624	0.1474	60.83	0.3252	0.4974	0.1823	0.0996	64.88	0.3640	0.6208	0.1341	0.0721	66.10	0.3689	0.6323
CODiff (ours, s=1)	0.2086	0.0994	66.28	0.4069	0.6413	0.1436	0.0714	66.97	0.4113	0.6498	0.1020	0.0511	66.66	0.4042	0.6443

(c) Quantitative comparison on the DIV2K-val dataset

Table 1. Quantitative comparison on LIVE-1, Urban100, and DIV2K-Val datasets with non-diffusion, MSD, and OSD methods. M-IQA stands for MANIQA, and C-IQA stands for CLIPQA. The best and second best results are colored with red and blue. DiffBIR* and OSEDiff* are retrained on the same dataset as our method for reference.

generator and discriminator. The discriminator employs a pre-trained StableDiffusion UNet encoder and a lightweight MLP. We set the LoRA [13] rank to 16 for fine-tuning. We utilize CaVE to construct the prompt embeddings. The discriminator follows the same training setup as CODiff. The weight for DISTS loss λ_D and GAN loss λ_G is set to 1 and 5×10^{-3} , respectively. This phase is trained for 100K iterations on 4 NVIDIA A6000 GPUs.

4.2. Comparison with SOTA Methods

Compared Methods. We compare CODiff with state-of-the-art CNN and Transformer-based methods specifically designed for JPEG artifact removal, including FBCNN [15], JDEC [11], and PromptCIR [20]. Additionally, we evaluate CODiff against recent leading diffusion-based image restoration methods, including DiffBIR [25], SUPIR [49], and OSEDiff [43]. Among them, DiffBIR and OSEDiff are built upon the pre-trained StableDiffusion [29] model. While SUPIR leverages StableDiffusion-XL [28], a larger model with 2.6 billion parameters. We re-train DiffBIR and OSEDiff under the same experimental settings as ours and denote them as DiffBIR* and OSEDiff* respectively.

Quantitative Results. The quantitative comparisons on LIVE-1, Urban100, and DIV2K-Val datasets are summarized in Tab. 1. CODiff consistently outperforms competing methods across a diverse range of evaluation metrics. Specifically, it achieves notable improvements in both full-reference metrics (LPIPS and DISTS) and no-reference metrics (MUSIQ, MANIQA, and CLIPQA). Those comparisons demonstrate its superior ability to recover visually pleasing and perceptually accurate high-quality images.

Compared to existing OSD models, CODiff significantly surpasses OSEDiff across all evaluation criteria, highlighting the effectiveness of incorporating compression-aware visual embeddings into the diffusion process, as well as the elaborately designed training objectives. Additionally, CODiff outperforms MSD models, such as DiffBIR and SUPIR, in most metrics, despite requiring much fewer sampling steps. This suggests that CODiff can achieve high-quality restoration with greater efficiency.

Moreover, diffusion-based methods generally outperform CNN and Transformer-based approaches, particularly when handling highly compressed images (*e.g.*, QF=5). This advantage primarily stems from the rich image gen-

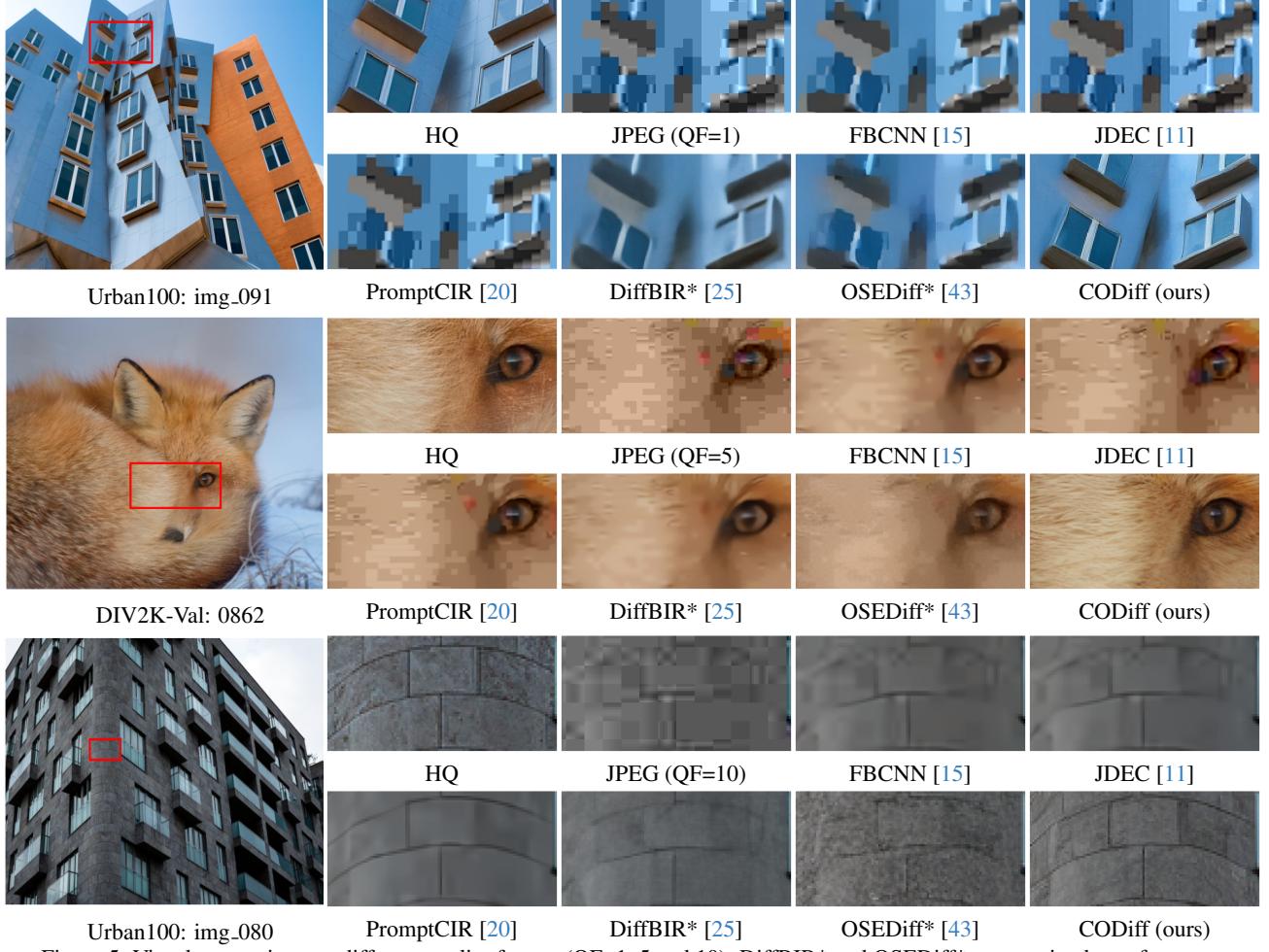


Figure 5. Visual comparison on different quality factors (QF=1, 5 and 10). DiffBIR* and OSEDiff* are retrained as references.

eration priors of large-scale T2I diffusion models (*e.g.*, StableDiffusion [29]). These priors enable the models to compensate for the severe loss of visual information in highly compressed images, allowing for more faithful and perceptually realistic reconstructions.

Qualitative Results. As shown in Fig. 5, previous methods, including FBCNN [15], JDEC [11], PromptCIR [20], and OSEDiff [43], can only partially mitigate JPEG artifact. However, noticeable compression artifact, such as blocky patterns, color banding, and grid-like distortions, still persist in their outputs. While multi-step DiffBIR is more effective in suppressing artifact, it tends to produce overly smooth results, sacrificing texture details in the process. In contrast, our CODiff effectively eliminates JPEG artifact while preserving fine-grained contents, such as fur patterns and architectural details. This highlights the model’s ability to recover complex details that are often lost due to heavy compression. Overall, our method strikes a balance between artifact removal and texture preservation, ensuring that the restored images retain both high perceptual quality and structural integrity.

4.3. Complexity Analyses

Table 2 presents a comparison of model complexity, considering key factors such as the number of sampling steps (#Step), parameter number (Params), multiply-accumulate operations (MACs), and inference time (Time). To ensure a fair assessment, all models are evaluated on a single NVIDIA A6000 GPU, except for SUPIR, which requires two A6000 GPUs due to its extensive model size. Notably, CODiff achieves a significant reduction in computational cost, outperforming MSD models by a substantial margin. This highlights its remarkable efficiency without compromising image quality. Moreover, CODiff requires fewer parameters compared to other diffusion based approaches, because we avoid complex auxiliary modules. Specifically, OSEDiff [43] employs DAPE [44] for textual prompt extraction. Meanwhile, DiffBIR [25] and SUPIR [49] utilize ControlNet [51] to integrate low-quality (LQ) image information as the denosing condition. These auxiliary modules substantially increasing parameter counts. In contrast, our CaVE is a lightweight module that efficiently extracts compression priors with low computational overhead.

Method	#Step	Params (G)	MACs (T)	Time (s)
DiffBIR [25]	50	1.52	188.24	50.81
SUPIR [49]	50	4.49	464.29	24.33
OSEDiff [43]	1	1.40	10.39	0.65
CODiff (ours)	1	1.00	9.46	0.57

Table 2. Complexity comparison among diffusion-based methods. Input image size is $1,024 \times 1,024$ for inference.

Method	LIVE-1			DIV2K-Val		
	LPIPS↓	MUSIQ↑	M-IQA↑	LPIPS↓	MUSIQ↑	M-IQA↑
Empty	0.3485	62.56	0.3793	0.3241	56.68	0.3142
Learnable	0.3471	63.39	0.3900	0.3235	56.92	0.3262
DAPE [44]	0.3463	62.54	0.3793	0.3230	57.29	0.3240
CaVE (ours)	0.3426	67.13	0.4584	0.3179	61.83	0.3709

Table 3. Ablation study of prompt embedding generation on LIVE-1 and DIV2K-val datasets. The best and second best results are colored with red and blue, respectively.

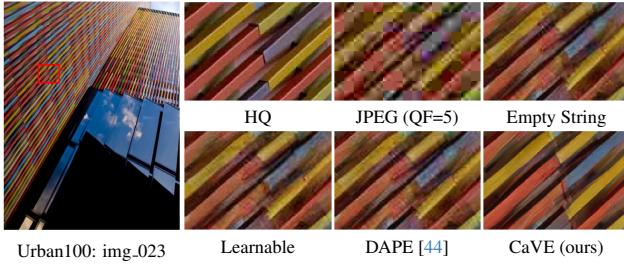


Figure 6. Visual comparison of prompt embedding generation. We use Urban100 img.023 as an example.

4.4. Ablation Studies

Compression-aware Visual Embedder (CaVE). We assess the effectiveness of different prompt methods, including empty strings, learnable embeddings, Degradation-Aware Prompt Extractor (DAPE) [44], and CaVE. Table 3 provides the results on LIVE-1 and DIV2K-Val test sets, focusing on challenging scenarios (*e.g.*, QF=1), where extreme compression leads to severe visual information loss. Our findings demonstrate that CaVE consistently outperforms other approaches across most evaluation metrics. In Fig. 6, utilizing CaVE significantly reduces JPEG artifact, yielding noticeably more clear image reconstructions. Such visualization further shows the effect of CaVE.

Dual Learning. To validate the effectiveness of our dual learning, we train CaVE with different learning strategies. Then we utilize CaVE to guide the reconstruction of CODiff. As shown in Tab. 4, the model trained with dual learning consistently achieves the best performance across all evaluation metrics. This suggests that dual learning paradigms enables CaVE to capture richer compression priors.

Moreover, we evaluate CaVE’s generalization ability by predicting QFs from low-quality images, specifically testing on QFs that are absent from the training set. As shown in Fig. 7, CaVE trained with explicit learning struggles to generalize to unseen QFs, whereas the dual learning approach significantly enhances its generalization capability.

Type	LIVE-1			DIV2K-Val		
	LPIPS↓	MUSIQ↑	M-IQA↑	LPIPS↓	MUSIQ↑	M-IQA↑
Explicit	0.3733	62.39	0.3828	0.3398	57.26	0.3202
Implicit	0.3436	64.74	0.4192	0.3208	58.74	0.3403
Dual	0.3426	67.13	0.4584	0.3179	61.83	0.3709

Table 4. Ablation study of the dual learning strategy. The best and second best results are colored with red and blue, respectively.

\mathcal{L}_2	\mathcal{L}_{DISTs}	\mathcal{L}_G	LPIPS↓	DISTS↓	MUSIQ↑	M-IQA↑	C-IQA↑
✓			0.3976	0.2509	59.79	0.3212	0.3085
✓	✓		0.2109	0.1229	72.69	0.5147	0.7119
✓		✓	0.3899	0.2411	59.98	0.3301	0.3130
✓	✓	✓	0.2113	0.1147	73.34	0.5221	0.7317

Table 5. Ablation studies on different loss functions. The best and second best results are colored with red and blue, respectively.

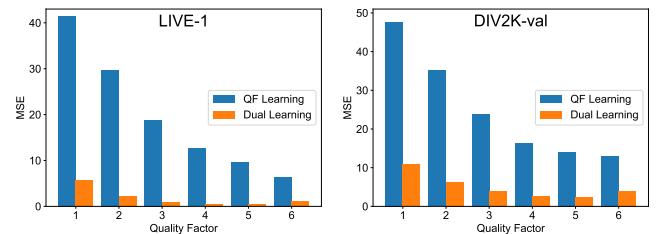


Figure 7. Comparison of CaVE’s QF prediction mean squared error (MSE) between explicit learning and dual learning on LIVE-1 and DIV2K-val datasets.

CODiff Training Loss Functions. CODiff incorporates multiple loss functions during training, including MSE loss, DISTs loss, and GAN loss, to enhance overall model performance. To systematically evaluate the contribution of each component, we conduct extensive ablation studies. As shown in Tab. 5, training CODiff solely with \mathcal{L}_2 results in suboptimal perceptual metrics. However, incorporating \mathcal{L}_{DISTs} significantly enhances CODiff’s performance. This highlights the pivotal role of DISTs loss in enhancing model’s ability to achieve high-quality reconstructions. Furthermore, adding the GAN loss provides an additional performance boost, indicating its complementary effect in refining the reconstruction quality.

5. Conclusion

We propose CODiff, a one-step diffusion model designed for efficient JPEG artifact removal. Our approach leverages the generative capabilities of text-to-image diffusion models to restore heavily JPEG-compressed images. By integrating a compression-aware visual embedder, we effectively utilize JPEG compression priors to guide the denoising process. This enables more accurate and visually coherent restorations. Beyond addressing the limitations of existing methods, CODiff establishes a novel pathway for using compression priors to guide diffusion models for JPEG artifact removal. Extensive experiments show the superiority of our CODiff over recent leading methods.

References

- [1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 5
- [2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 1
- [3] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 2020. 5
- [4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 2
- [5] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *ICCV*, 2015. 1, 2
- [6] Max Ehrlich, Larry Davis, Ser-Nam Lim, and Abhinav Shrivastava. Quantization guided jpeg artifact correction. In *ECCV*, 2020. 1, 2
- [7] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep generative adversarial compression artifact removal. In *ICCV*, 2017. 2
- [8] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep universal generative adversarial compression artifact removal. *IEEE TMM*, 2019. 2
- [9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 5
- [10] Jun Guo and Hongyang Chao. Building dual-domain representations for compression artifacts reduction. In *ECCV*, 2016. 2
- [11] Woo Kyoung Han, Sunghoon Im, Jaedeok Kim, and Kyong Hwan Jin. Jdec: Jpeg decoding via enhanced continuous cosine coefficients. In *CVPR*, 2024. 1, 2, 6, 7
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1
- [13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 5, 6
- [14] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 2, 5
- [15] Jiaxi Jiang, Kai Zhang, and Radu Timofte. Towards flexible blind jpeg artifacts removal. In *ICCV*, 2021. 1, 2, 4, 6, 7
- [16] Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. In *ECCV*, 2024. 3
- [17] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021. 5
- [18] Yoosik Kim, Jae Woong Soh, and Nam Ik Cho. Agarnet: Adaptively gated jpeg compression artifacts removal network for a wide range quality factor. *IEEE Access*, 2020. 2
- [19] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [20] Bingchen Li, Xin Li, Yiting Lu, Ruoyu Feng, Mengxi Guo, Shijie Zhao, Li Zhang, and Zhibo Chen. Promptcir: Blind compressed image restoration with prompt learning. *arXiv preprint arXiv:2404.17433*, 2024. 1, 2, 5, 6, 7
- [21] Jianwei Li, Yongtao Wang, Haihua Xie, and Kai-Kuang Ma. Learning a single model with a wide range of quality factors for jpeg image artifacts removal. *IEEE TIP*, 2020. 2
- [22] Jianze Li, Jiezhang Cao, Zichen Zou, Xiongfei Su, Xin Yuan, Yulun Zhang, Yong Guo, and Xiaokang Yang. Distillation-free one-step diffusion for real-world image super-resolution. *arXiv preprint arXiv:2410.04224*, 2024. 1
- [23] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *CVPR*, 2023. 5
- [24] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, 2021. 1, 2
- [25] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, 2024. 1, 3, 6, 7, 8
- [26] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5
- [28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 3, 6
- [29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3, 6, 7
- [30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4, 5
- [31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 1, 3
- [32] H Sheikh. Live image quality assessment database release 2. <http://live.ece.utexas.edu/research/quality>, 2005. 5
- [33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1
- [34] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1
- [35] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single

- image super-resolution: Methods and results. In *CVPRW*, 2017. 5
- [36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 2, 4
- [37] Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 1991. 1, 6
- [38] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 5
- [39] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024. 3
- [40] Menglu Wang, Xueyang Fu, Zepei Sun, and Zheng-Jun Zha. Jpeg artifacts removal via compression quality ranker-guided networks. In *IJCAI*, 2021. 2
- [41] Xi Wang, Xueyang Fu, Yurui Zhu, and Zheng-Jun Zha. Jpeg artifacts removal via contrastive representation learning. In *ECCV*, 2022. 4
- [42] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *CVPR*, 2024. 3, 5
- [43] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024. 1, 3, 5, 6, 7, 8
- [44] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024. 1, 3, 7, 8
- [45] 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 *CVPR*, 2022. 5
- [46] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*, 2024. 3
- [47] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024. 3, 5
- [48] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024. 3, 5
- [49] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, 2024. 1, 3, 6, 7, 8
- [50] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *NeurIPS*, 2024. 3
- [51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 7
- [52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5
- [53] Xiaoshuai Zhang, Wenhan Yang, Yueyu Hu, and Jiaying Liu. Dmcnn: Dual-domain multi-scale convolutional neural network for compression artifacts removal. In *ICIP*, 2018. 2
- [54] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 1, 2
- [55] Bolun Zheng, Yaowu Chen, Xiang Tian, Fan Zhou, and Xuesong Liu. Implicit dual-domain convolutional network for robust color image compression artifact reduction. *IEEE TCSVT*, 2019. 2