Consistency Guided Diffusion Model with Neural Syntax for Perceptual Image Compression

Haowei Kuang
Wangxuan Institute of Computer Technology, State Key Laboratory of Multimedia Information Processing, Peking University
Beijing, China
kuanghw@stu.pku.edu.cn

Yiyang Ma
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
myy12769@pku.edu.cn

Wenhan Yang
Pengcheng Laboratory
Shenzhen, China
yangwh@pcl.ac.cn

Zongming Guo
Wangxuan Institute of Computer Technology, State Key Laboratory of Multimedia Information Processing, Peking University
Beijing, China
guozongming@pku.edu.cn

Jiaying Liu*
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
liujiaying@pku.edu.cn

ABSTRACT

Diffusion models show impressive performances in image generation with excellent perceptual quality. However, its tendency to introduce additional distortion prevents its direct application in image compression. To address the issue, this paper introduces a Consistency Guided Diffusion Model (CGDM) tailored for perceptual image compression, which integrates an end-to-end image compression model with a diffusion-based post-processing network, aiming to learn richer detail representations with less fidelity loss. In detail, the compression and post-processing networks are cascaded and a branch of consistency guided features is added to constrain the deviation in the diffusion process for better reconstruction quality. Furthermore, a Syntax driven Feature Fusion (SFF) module is constructed to take an extra ultra-low bitstream from the encoding end as input, guiding the adaptive fusion of information from the two branches. In addition, we design a globally uniform boundary control strategy with overlapped patches and adopt a continuous online optimization mode to improve both coding efficiency and global consistency. Extensive experiments validate the superiority of our method to existing perceptual compression techniques. Our project is publicly available at: https://ellisonkuang.github.io/CGDM.github.io/.

CCS CONCEPTS

• **Computing methodologies** → **Image compression**.

*Corresponding author.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681336

KEYWORDS

Image compression, generative model, denoising diffusion model, neural syntax

ACM Reference Format:
Haowei Kuang, Yiyang Ma, Wenhan Yang, Zongming Guo, and Jiaying Liu. 2024. Consistency Guided Diffusion Model with Neural Syntax for Perceptual Image Compression. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681336

1 INTRODUCTION

Image compression technology plays a pivotal role in diverse fields, *e.g.*, multimedia, communications, and computer vision. Its objective is to efficiently minimize the storage space and bandwidth requirements of digital images for their efficient storage and transmission, while preserving the main content of the original images and maintaining the visual quality. In today's digital age, with the ever-increasing demand for high-resolution and high-quality images from multimedia devices, the challenge of managing storage and bandwidth resources has become paramount, and the demand for more efficient and high-performance image compression methods is also growing rapidly.

Over the past decades, conventional image compression techniques like JPEG [58], BPG [5] and JPEG2000 [42] have become the common choice in image processing. These methods or standards exhibit excellent encoding capabilities, adopting the route of transform/hybrid coding framework for Rate-Distortion Optimization (RDO) with key modules such as transformation, quantization, and entropy coding, while being complemented by additional complex predictive modes. Moreover, numerous efforts have been made to improve the RDO of each input image by projecting the image signals into a manually designed specific subspace for more compact representations, *e.g.* intra-prediction based on various directions

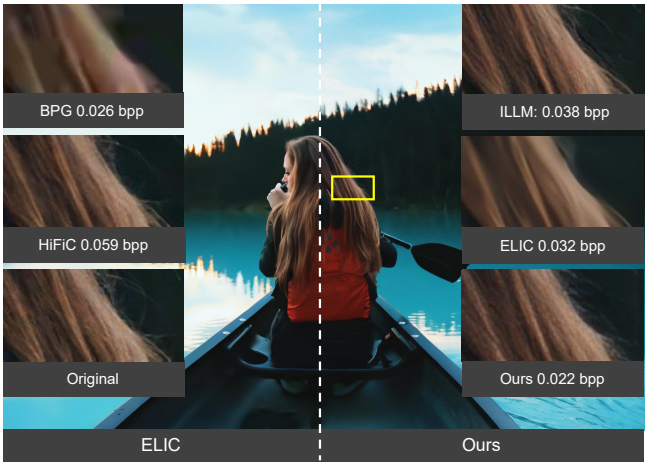

Figure 1: Visual comparisons of different methods. The patch is cropped from *roberto-nickson-48063.png* from CLIC professional dataset [56]. Compared to previous methods, our method achieves better or competitive perceptual quality at a lower bitrate. In particular, our method achieves similar perceptual quality using about half bitrate compared to the milestone HiFiC [43]. [*Zoom in for best view*]

[58]. Nonetheless, these optimizations rely on human design, lacking the capability for global optimization. As the number of manually designed strategies continues to grow, the framework of the compression model becomes increasingly complicated, gradually revealing its performance bottlenecks.

In recent years, deep learning technologies have made remarkable advancements, prompting many researchers to explore image compression methods with the immense capabilities of neural networks [4, 21, 27, 40, 59]. By leveraging vast datasets [1, 33, 38, 56] to train neural networks, these methods excel in discovering the latent relationships and underlying structures embedded within image data. As a result, they attain a remarkable compression efficiency, while ensuring minimal distortion, thereby surpassing traditional techniques in terms of performance. In addition, there are some implicit neural representation based efforts [7, 14, 36, 55] to store images in neural network parameters for image compression. While these methods have achieved a notable level of image compression and reconstruction, they gradually encounter performance bottlenecks: further improving the compression ratio resulting in significantly degraded quality.

Some researches [6, 61] indicate that the distortion of images does not align with human perception of subjective image quality. Due to the inherent trade-off between image quality and storage efficiency, perceptual image compression techniques [43, 47, 62] are proposed. These technologies strive to protect the quality in terms of human visual perception rather than focusing on fidelity measurement. To enhance the subjective visual quality of images, many generative models [20, 35, 57] skill at producing visually appealing details such as GANs are seamlessly integrated into image compression methods [43, 47, 64] resulting in a significant improvement in perception quality. However, the development of these methods is constrained by the inherent limitations of generative models, including the frequent lack of diversity in the images produced by GAN models and so on, thereby posing significant challenges in their further advancement.

In recent years, diffusion models [24, 39, 53] have emerged as a powerful tool in the field of image generation, exhibiting remarkable capabilities in producing images with exceptional perceptual quality. These models, developed according to the formulation of diffusion processes, have demonstrated their ability to capture intrinsic details and generate realistic images. However, despite their remarkable success in image generation, it has been established in many practices that vanilla diffusion models tend to reconstruct images with richer visual details at a cost of significantly impaired fidelity [50], due to the random nature of the process of progressive adding or removing noise. Applying the diffusion models directly to the image compression task may take on risks of a synchronous drop of both visual quality and fidelity. Thus, until now, how to apply diffusion models to image compression remains under-explored.

To address the issue of utilizing the power of diffusion models while avoiding their generated artifacts, this paper proposes to regularize the diffusion models with global consistency guidance. In detail, we propose a novel approach called the **Consistency Guided Diffusion Model (CGDM)**, which incorporates additional consistent guidance into the network structure of the diffusion model. This approach aims to constrain deviations in the diffusion process for improving the quality of the reconstructed image. Furthermore, we propose a Syntax driven Feature Fusion (SFF) strategy. This strategy encodes an additional ultra-low bitstream obtained from the encoding stage, providing semantic prior information about the image. By leveraging this prior information, we can reduce the ambiguity in the inference target during the post-processing phase, leading to more accurate and faithful reconstructions. To achieve the same objective of reducing randomness in the diffusion process and effectively leveraging image semantic information, we apply a globally consistent edge control strategy into our model's inference phase. Additionally, we adopt a continuous online optimization mode to further enhance the model's performance. These efforts not only reduce the stochasticity associated with diffusion processing but also entourage the model to capture rich semantics from the input images, thereby leading to improved overall performance.

Our contributions are summarized as follows:

- We develop a Consistency Guided Diffusion Model (CGDM) for perceptual image compression, which incorporates an additional consistent guidance with a diffusion-based network, aiming to constrain the deviation in the diffusion process for learning richer detail representations with less fidelity loss.
- We devise a Syntax driven Feature Fusion (SFF) module that takes an extra ultra-low bitstream from the encoding end as input, guiding the adaptive fusion of information from the two branch.
- We design a globally uniform boundary control strategy, and adopt continuous online optimization mode to further improve both coding efficiency and global consistency.

Experimental results show that our proposed method achieves a BD-rate [19] savings of 9.227% in perception and 6.251% in distortion compared to the current state-of-the-art perceptual image compression method ILLM [47].

2 RELATED WORKS

2.1 Generative Model

Generative models aim to learn the overall distribution of data and generate data within the same distribution, which have been a central focus of research in recent years, leading to significant advancements in multimedia generation and processing. With the rapid development of deep learning, one of the milestone and the most notable generative models is the Generative Adversarial Networks (GANs) [20], which consist of two competing networks, a generator and a discriminator. The adversarial training process results in highly realistic and diverse generations. Subsequent work built upon GANs such as Conditional GAN [46], StyleGAN [28, 29] further enhances its ability to generate high-quality images based on given conditions. Beyond GANs, there has also been a surge of interest in other types of generative models, including autoregressive models [17], variational autoencoders (VAEs) [32], normalizing flow models [12], energy-based models [13], score-based model [54], flow-based model [31] and so on.

Recently, diffusion models [24, 39, 53] have emerged as powerful generative models that define the forward and reverse diffusion processes for data noise addition and removal, respectively. Their generations often exhibit superior quality and diversity, and there have been many studies attempting to use diffusion-based models for image generation [10, 65], enhancement [15, 41, 48, 63] and so on. In our work, we apply the diffusion-based generation model to image compression, and obtain the image with higher quality through the guidance of a semantic stream.

2.2 Learned Image Compression

With the significant advancements in deep learning, recent years have witnessed deep learning based image compression methods outperforming classical methods in striking a balance between bit rate and reconstruction quality. Initially, Ballé *et al.* [2, 3] pioneered the utilization of neural network to establish lossy image compression autoencoders, sparking a surge in learning-based image compression methods [44, 45]. In addition to transformations, numerous studies have focused on entropy coding of latent representations based on learned probability models, including hyperpriors [4] and context models [8, 37]. Furthermore, the employment of Gaussian Mixture Models and attention-based modules in transformations has further enhanced image compression performance [9].

Facing the trade-off between image quality and storage efficiency, a range of perceptual image compression methods have been proposed, which aim to enhance the perceptual quality of compressed images and align them more closely with human perception. Agustsson *et al.* [1] introduced the concept of using GANs [20] as decoders for image compression. This approach allows for the generation of reconstructed images with rich details. Subsequently, He *et al.* [22] further enhanced these GAN-based methods by incorporating advanced perception models. Recently, with the great success of diffusion models, some efforts [18, 25, 62] have been made to study perceptual image compression. However, as we have stated previously, due to the lack of fidelity caused by the uncertainty of the diffusion process, this area needs to be further explored. In this paper, we propose a solution for this issue through an additional consistent guidance and a neural syntax driven strategy.

3 METHOD

In this part, we first describe general information of diffusion models while outlining our motivations in Section 3.1, followed by a detailed elaboration on our proposed consistency guided diffusion model in Section 3.2. Then, we further propose our syntax driven feature fusion module in Section 3.3. The globally uniform boundary control strategy and continuous online optimization during inference is introduced in Section 3.4. Finally, our training strategy is described in Section 3.5.

3.1 Preliminaries and Motivations

We start with the characteristic analysis of the diffusion model. As a generative model, diffusion models have been demonstrated to effectively create images with excellent perceptual quality by leveraging a conditional model that incorporates latent features.

Simultaneously, there are numerous works [15, 41, 48, 63] that employ the diffusion models as a post-processing or enhancement module. Generally, these works utilize the degraded image $\tilde{x}$ as a condition and construct a conditional model that aims to learn the data distribution $p(x|\tilde{x})$ through a fixed multi-step chain of length T. The diffusion process is defined by a forward process q through adding Gaussian noise. Formally, the distribution of the forward process can be expressed as:

$$\begin{aligned} q(x_t|x_0) &= \mathcal{N}(x_t; \alpha_t x_0, \sigma_t^2 \mathbf{I}), \\ q(x_T) &= \mathcal{N}(x_T; \mathbf{0}, \mathbf{I}), \end{aligned} \tag{1}$$

where α_t and σ_t^2 are hyper-parameter functions of t [39].

Meanwhile, the inference process can be conducted as a reverse process from Gaussian noise $q(x_T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to a target x_0, which can be expressed as:

$$\begin{aligned} p(x_T) &= \mathcal{N}(x_T|\mathbf{0}, \mathbf{I}), \\ p(x_{t-1}|x_t, \tilde{x}) &= \mathcal{N}(x_{t-1}|\mu_\theta(\tilde{x}, x_t, t), \sigma_t^2 \mathbf{I}), \end{aligned} \tag{2}$$

where the $\mu_\theta(\tilde{x}, x_t, t)$ denotes the mean value of the conditional distribution $p_\theta(x_{t-1}|x_t, \tilde{x})$, and the diffusion model is trained to learn the conditional distributions by parametric approximation to the distribution. For the neural network, mostly existing diffusion post-processing frameworks directly feed the condition $\tilde{x}$ and noise x_t, along with the timestamp t, into the U-Net backbone, similar to the vanilla DDPM and output the predicted noise $\hat{\epsilon}_t$ at each step.

However, there are two notable issues with this paradigm:

- This approach often leads to the final reconstructed image x_0 deviating from the condition $\tilde{x}$, amplifying the distortion of $\tilde{x}$ and results in simultaneous degradation of both fidelity and perceived quality.
- As the condition $\tilde{x}$ represents a degraded image with some information loss, there can exist multiple images that lead to the same $\tilde{x}$. That means, the optimized probability distribution target $p(x|\tilde{x})$ is ambiguous, making it challenge to ensure that the reconstructed image is more similar to the original image.

In our work, we aim to adopt a novel framework to address both of these issues:

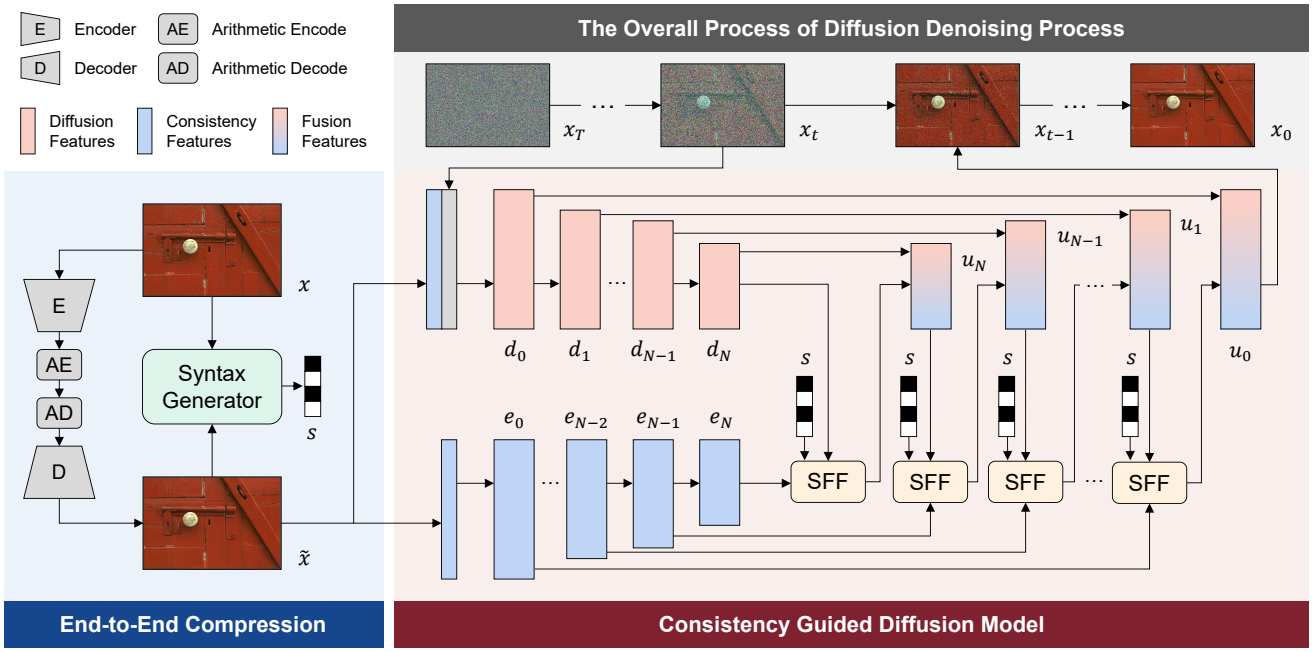

Figure 2: The entire framework of our proposed method. For an image x to be encoded, we first perform lossy compression using a standard end-to-end image compression network, resulting in an output degraded image $\tilde{x}$. Then, we extract a syntax vector from the original image x using a syntax generator. This syntax vector is then used to guide the fusion of consistency features e and diffusion features d in a *Consistency Guided Diffusion Model with Neural Syntax*. After a complete diffusion process, we obtain a higher-quality reconstructed image x_0. The consistent guidance architecture, neural syntax driven mechanism lead the diffusion model to stably reconstruct high-quality images, making the final output excellent in terms of perception and fidelity.

- For the first issue, we propose to incorporate an additional consistent guidance into the network structure of the diffusion model, called consistency guided diffusion model, constraining the deviation in the diffusion process and improving the quality of the reconstructed image.
- For the second issue, we propose a syntax driven feature fusion strategy to encode an additional ultra-low bitstream s from the encoding stage to provide semantic prior information of the image, thereby alleviating the ambiguity in the inference target of post-processing.

In the following sections, we describe our method in detail.

3.2 Consistency Guided Diffusion Model

On a high level, our compression process consists of two parts, an end-to-end image compression model and a diffusion-based post-processing model called consistency guided diffusion model, with structure shown in Fig. 2.

End-to-end image compression model. Firstly, we utilize a standard end-to-end image compression network to perform lossy compression on the original image x, obtaining an image $\tilde{x}$ with some detailed information lost:

$$\tilde{x} = \mathrm{D}(\mathrm{Q}(\mathrm{E}(\mathrm{x}))), \tag{3}$$

where $\mathrm{Q}(\cdot)$ means quantizer and $\mathrm{E}(\cdot)$, $\mathrm{D}(\cdot)$ represents the pre-trained autoencoder. Here, we utilize the recently proposed ILLM [47] as the end-to-end autoencoder, which is the current state-of-the-art perception-oriented end-to-end compression method.

Diffusion-based post-processing model. Our diffusion-based post-processing model follows the encoder-decoder architecture with skip connections [49] as the denoising model similar to [16], which includes two encoders and one decoder.

The upper branch in Fig. 2 encodes the noisy image into N multi-resolution diffusion feature maps d_i with different scales, where N is the depth of the U-Net backbone and $i \in \{0, ..., N\}$, while the lower branch extracts feature maps e_i of corresponding scales from the image $\tilde{x}$. Then, we introduce a syntax driven feature fusion module guided by an ultra-low semantic information bitstream s in the decoder part of the U-Net, which is expounded in the following section. Initially, the feature d_N and e_N are fused to obtain the feature u_N. Then, at each layer, the corresponding layer's feature u_{i+1} and e_i are adaptively fused to produce u_i, which can be expressed as:

$$u_i = \mathrm{SFF_i}(\mathrm{u_{i+1}}, \mathrm{e_i}, \mathrm{s}), \tag{4}$$

where $\mathrm{SFF_i}$ means the syntax driven feature fusion module and s is a compact syntax vector. This process is repeated layer by layer, with adaptive fusion and upsampling of features performed at each step, ultimately generating the predicted output noise.

By taking this approach, during the denoising diffusion process of the diffusion post-processing model, we consistently inject a constant guidance feature derived from the degraded image, guiding its inference process to stay close to the conditional distribution $\tilde{x}$. This approach enables the final output to enhance the perceptual quality while maintaining the similarity to the original image, achieving a better trade-off between fidelity and perceptual quality.

3.3 Syntax Driven Feature Fusion

As we mentioned in section 3.1, the target optimized probability distribution of the post-processing module $p(x|\tilde{x})$ is ambiguous. To mitigate the ambiguity to ensure the reconstructed image x_0 similar to original image x, we proposed to send a compact syntax vector extracted from the original image x and decoded image $\tilde{x}$ which costs ultra-low bitstream to provide the syntax information of the original image. Based on this idea, inspired by [59], we encode the syntax information of the original image into a compact and discrete one-dimensional vector by a syntax generator which serve as a dynamic convolution kernel in the syntax driven feature fusion module. By decoding the syntax vector into dynamic convolution kernels and performing convolution operations on the features to be fused, syntax information is transmitted in a neural representation-like manner. The structure of the syntax generator module and syntax driven feature fusion module is illustrated in Fig. 3.

Syntax Generator. The syntax generator's structure follows the design of [59], which contains a multi-scale network on the basis of hyper-priors entropy model [4, 26]. During syntax extraction, the features at each scale are globally average pooled and concatenate to a compact one-dimensional vector. This approach effectively utilizes multi-scale information while ensuring global consistency of the semantic information.

Syntax Driven Feature Fusion. The syntax driven feature fusion module takes the features d_i, e_i, the obtained syntax vector s, and the timestamp t of the current step as inputs. After getting the syntax vector s and timestamp t, we concatenate them and utilize a fully connected network to map them to two convolutional kernels W_e^i, W_d^i. These two sample-adaptive dynamic convolutional kernels separately perform convolution on the two input features, achieving adaptive fusion of features at each layer:

$$u_i = W_e^i * e_i + W_d^i * d_i, \tag{5}$$

where $*$ denotes convolution. Since the semantic vector used for generating these two convolutions is highly dependent on the input original image, the fusion process can effectively capture the characteristics of the original image, enabling the more adaptive fusion of the two streams of features during the generation process to obtain a better reconstructed image.

3.4 Inference Time Optimization

Furthermore, we propose that due to the excellent sample adaptability of our method, a more refined design during the inference process can more fully tap into the performance potential of our proposed method and achieve better performance. Specifically, during the inference process, we employ a globally uniform boundary control strategy and a sample-adaptive continuous online optimization mechanism for different resolutions and styles of images to be compressed, respectively, to further enhance the performance.

Globally Uniform Boundary Control. During model training, we used patches of a fixed size. However, directly feeding images of different resolutions into the diffusion model during inference can cause performance degradation due to distribution discrepancies. Therefore, we chose to use a tiling approach to adapt to images of arbitrary resolutions. To mitigate the potential block artifacts that may arise from piece-by-piece tiling, we employ the following two strategies. Firstly, we overlap the patches with surrounding ones to a certain extent, so that the pixels at the edge position are predicted by multiple patches, which makes the transition of the edge position smoother.

In addition, we observe that the initial noise x_T of the diffusion model can be viewed as the boundary condition of the diffusion ordinary differential equation, which affects the stylistic characteristics of the sampled images [41]. Recognizing that a complete image should exhibit consistent stylistic features, we set the initial noise for all patches to a fixed distribution. This ensures consistent image style, further mitigates block artifacts, and enhances the performance of encoding and decoding images at arbitrary resolutions. Specifically, for a given patch size, we commence by randomly selecting a boundary condition (just a Gaussian noise). Subsequently, we maintain this noise as a uniform boundary condition and tile it across the entire image, resulting in a boundary condition x_T, that spans the entire image. Using x_T as the starting, we then employ our CGDM to initiate the diffusion process. Through this process, we ultimately reconstruct a high-quality image, X0, that exhibits a stable and consistent style.

Sample-Adaptive Continuous Online Optimization. Similar to [40, 59], the approach of using a syntax generator to extract global syntax information from images inherently brings the potential for online optimization in the encoding phase during inference. This process is analogous to the mode decision process in traditional hybrid coding frameworks, where the best mode is selected from a discrete set of candidates. However, employing iterative optimization allows for continuous selection of the best option from an infinite set, greatly enhancing the flexibility of this online optimization strategy.

Specifically, during the inference process for each image, we iteratively optimize the encoder parameters of the syntax generator on randomly selected patches. This enables the generator to produce syntax vectors that more closely align with the image's semantics. For the iterative optimization process during inference, we set an optimization objective consistent with the fine-training process, which is detailed in the next section.

3.5 Training Strategy

Following the training strategy proposed by [60], we use a coarse-to-fine two-stage training strategy. Training at the coarse level aims to train the diffusion model to constrain noise, while training at the fine level further optimizes the diffusion model to further enhance the model performance by constraining the sampled clean images with fixed steps and corresponding ground truth ones.

Our coarse training is analogous to existing conditional diffusion models, which aim to estimate noise. The difference is, that we introduce an additional bitrate control term to constrain the bit cost of syntax vectors. Consequently, the loss function for the coarse training is as follows:

$$\mathcal{L}_{coarse} = \mathbb{E}_{t\sim[1,T],x_0,\epsilon_t} \|\epsilon_t - \mu_\theta(x_t, \tilde{x}, s, t)\|_2 + \lambda_c R, \tag{6}$$

where μ_θ denotes our noise prediction network, ϵ_t means the adding noise at step t in the forward process, R means the bitrate and λ_c is the hyper-parameter to trade-off between rate and distortion.

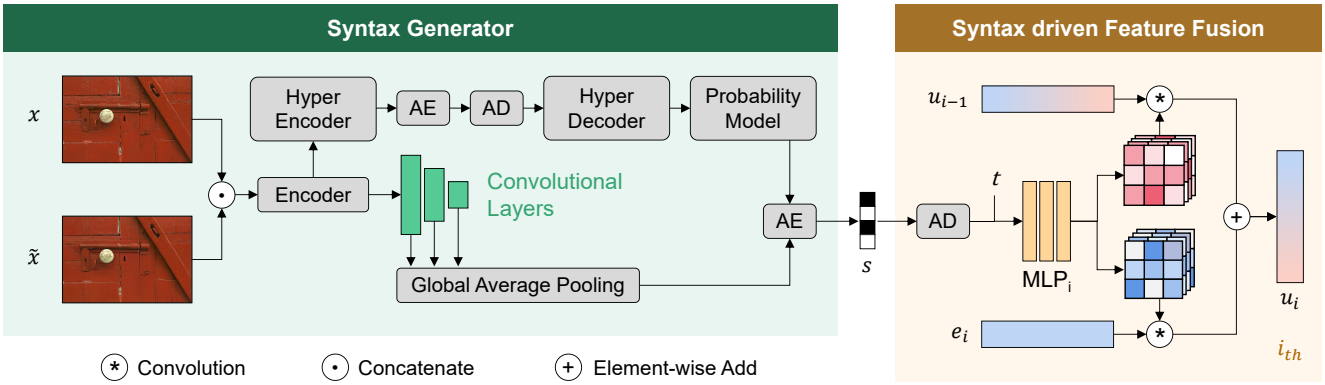

Figure 3: The structure of Syntax Generator and Syntax driven Feature Fusion module (SFF). The Syntax Generator module is responsible for extracting global syntax vectors from images during the encoding process, while the syntax driven feature fusion module adaptively integrates consistency features with diffusion features within the diffusion framework.

During the fine training stage, we fixed the sampling strategy to the 9-step DDIM sampling and imposed constraints on the generated sampled images $\hat{x}_0$ to compensate for the unsatisfactory results from the coarse training. The loss function we used for the constraints is as follows:

$$\mathcal{L}_{fine} = \lambda_d \mathcal{L}_d(\hat{x}_0, x) + \lambda_p \mathcal{L}_p(\hat{x}_0, x) + \lambda_f R, \tag{7}$$

where $\mathcal{L}_d$ denotes the distortion loss and we use L1 loss during training, $\mathcal{L}_p$ denotes the perception loss and we utilize three kinds of perceptual objective functions DISTS [11], Alex-based [34] and VGG-based [52] LPIPS [66] in the experiments. λ_d, λ_p and λ_f are the hyper-parameters.

4 EXPERIMENTS

4.1 Implementation

Network Implementation. We implement our diffusion model based on the architecture of [16] with fewer parameters. In addition, for further reduce the video memory consumption, we also remove the self-attention module. The detailed structure and hyperparameters are shown in the supplementary material.

Training Details. We utilize the DIV2K [1] dataset as our training dataset, which comprises 800 natural images with an average resolution of 2K. To enable our model to adapt to images of various resolutions, we perform downsampling on the images to half their original resolution, serving as an augmentation of the training data. During the training process, we randomly crop 256×256 patches from each image.

Our training process uses the Adam optimizer [30] and the learning rate is set to 1×10^{-4}. We train 6 models with different compression rates using different bit rates end-to-end compression model parameters. The hyper-parameter λ_c on the coarse training stage is set to 100 and the hyper-parameter λ_d, λ_p and λ_f are set to 1, 0.3 and 20 separately on the fine training stage. Each model is trained for 38k iterations on the coarse training stage and 32k iterations on the fine training stage.

Inference Details. During inference, the patch size used is 256×256, and the overlap range is 8 pixels close to the edge. Our method applies the continuous mode decision on inference. For each image, based on the pre-trained network weights, we additionally employ the Adam optimizer with a learning rate 5×10^{-5} to finetune the encoder for 250 iterations, and the optimization target is the same as $\mathcal{L}_{fine}$.

Evaluation Protocol. We evaluate our method on the Kodak image dataset [33] and the professional subset of the CLIC validation dataset [56]. The Kodak image dataset consists of 24 images, each with a resolution of 768×512. The CLIC professional validation dataset comprises 41 images with higher resolutions of about 1800×1200. Evaluating on it demonstrates the performance of our method on images with higher resolutions.

To demonstrate the superiority of our method in terms of distortion and perceptual quality, we utilize a set of diverse metrics. For distortion, we employ PSNR, VIF [51], and MS-SSIM [61]. And for perceptual quality, we used VGG-based [52] LPIPS [66], DISTS [11], and FID [23]. The R-D curves and BD-rate [19] on different evaluation metrics are illustrated to compare different methods.

4.2 Quantitative Comparison

We compare our method with existing conventional transform-based methods BPG [5], end-to-end learning-based image compression methods optimized for MSE like ELIC [21], and image compression methods optimized for perceptual quality including HiFiC [43], ILLM [47], and CDC [62]. Fig. 4 presents the R-D curves of various metrics on the CLIC and Kodak datasets for our proposed method and comparison methods. Evidently, our approach demonstrates superior performance across different perceptual measures compared to other perceptual image compression methods, meanwhile achieving a more favorable distortion effect. Among the comparison methods, only CDC surpasses our approach in terms of the perception metrics. However, its performance in the distortion metrics is significantly inferior to other methods, resulting in an overall performance that remain below ours. This underscores the excellent balance our method achieves between fidelity and perceptual quality, highlighting its superiority.

To provide a more intuitive comparison of the overall performance of our method with other benchmark methods across all evaluation metrics, we compute the BD-rate [19] for each indicator. Using our method as the anchor, Table 1 presents the average BD-rate achieved by each method across all distortion and perception metrics on both CLIC and Kodak datasets anchored on our method. It is evident from the table that the overall performance of

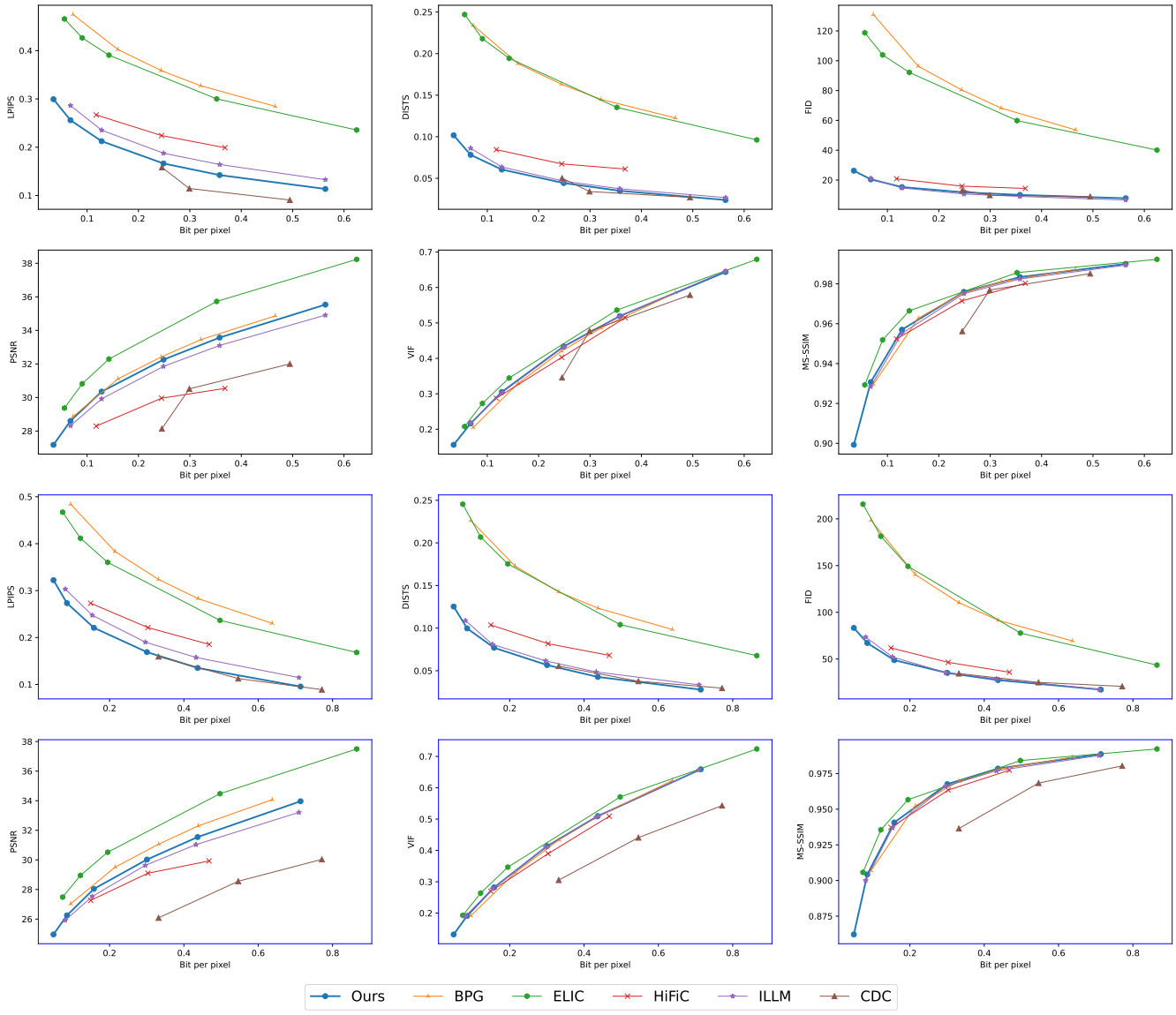

Figure 4: Tradeoffs between bitrate (x-axes, in bpp) and different metrics (y-axes) for various models tested on Kodak and CLIC. We consider both perceptual (LPIPS, DISTS, FID) and distortion metrics (PSNR, VIF, MS-SSIM). The upper 2 rows (black frame) are the performance on Kodak datasets and the lower 2 rows (blue frame) are on CLIC professional dataset.

Table 1: Average BD-rate for different methods on both CLIC and Kodak datasets *anchored on our method.*

Datasets	Kodak			CLIC		
Methods	Distortion	Perception	Average	Distortion	Perception	Average
HiFiC [43]	+14.6366	+45.1822	+29.9094	+43.4935	+113.7039	+78.5987
ILLM [47]	+5.5011	+11.4812	+8.4912	+6.9765	+14.1819	+10.5792
CDC [62]	+52.0296	+0.5460	+26.2878	+65.2966	-21.5621	+21.8673
ELIC [21]	-31.6699	+77.8194	+23.0748	-22.2319	+2288.3859	+1133.0770
BPG [5]	-1.5526	+84.3209	+41.3842	+3.0902	+4872.2267	+2437.6585

Figure 5: Visual comparisons with state-of-the-art methods on Kodak dataset. We provide further analysis that focuses on subjective results in the main text. As can be seen, compared to the baseline used in our method (ILLM), we achieve a significant improvement in subjective performance at the cost of extremely low additional bitstreams. [*Zoom in for best view*]

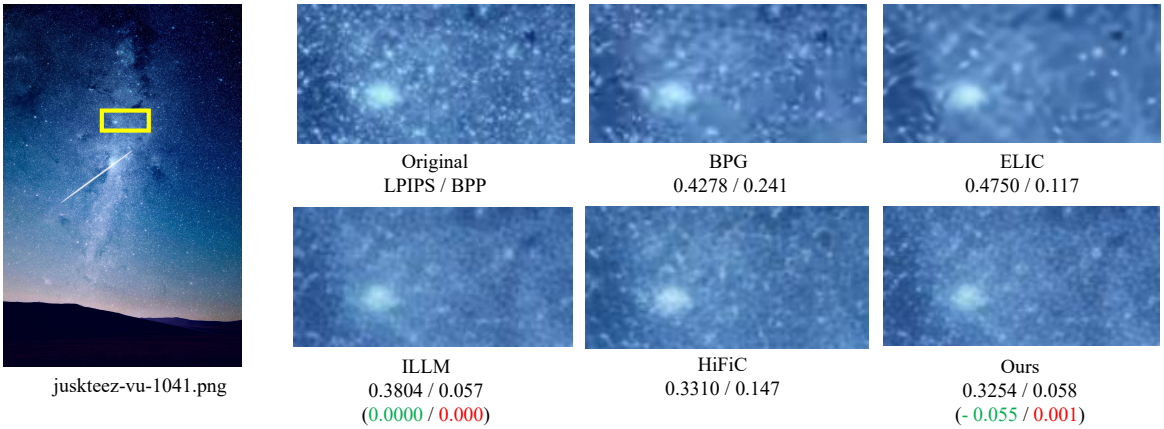

Figure 6: Visual comparisons with state-of-the-art methods on CLIC dataset. [*Zoom in for best view*]

all other methods under all evaluation indicators falls below that of our proposed approach.

4.3 Qualitative Comparison

To further underscore the perceptual quality of our results, we present several illustrative cases comparing different image compression methods in Fig. 1, 5 and 6. In Fig. 5, it is clearly observed that traditional compression methods such as BPG [5] and MSE-optimized compression methods like ELIC [21] produce overly smooth images with significant loss of detail information. Among the perception-oriented optimization methods, ILLM [47] introduces numerous continuous and repetitive artifacts in the decoded images, and the details in HiFiC [43] are not as clear as those in our method though its bitrate is much higher than ours. Fig. 6 shows the results on high-resolution images, whose subjective performance is generally consistent with that on the Kodak. Evidently, the reconstructed images using our method exhibit richer visual details, and less artifact while utilizing fewer or comparable bits.

4.4 Ablation Studies

We conduct extensive ablation studies for our proposed network architecture on the Kodak dataset. By replacing the syntax driven feature fusion module with direct element-wise addition, the model fuses information directly without syntax guided adaptive fusion (*w/o* SFF). By replacing the consistent boundary with random noise and inferring by pre-trained model parameters without online fine-tuning, the optimization during inference is moved (*w/o* Infer. Optim.). We do the above substitutions in turn and observe a performance drop as Table 2 shows, though the model sizes are kept almost the same. Hence, all of the components in our design contribute to performance improvement.

Table 2: Average BD-rates of the ablation studies.

w/ SFF	*w/* Infer. Optim.	Distortion	Perception	Average
✓	✓	–	–	–
✓	×	-0.6501	+1.7797	+0.5648
×	×	+0.8688	+2.4148	+1.6418

5 CONCLUSION

In this work, a novel consistency guided diffusion model with neural syntax is proposed, introducing a diffusion model for perceptual image compression. The consistency guidance architecture, neural syntax driven mechanism and inference time optimization strategy lead the diffusion model to reconstruct high-quality images, making the final output excellent in terms of perception and fidelity. Experimental evaluation shows the superiority of our methods.

6 ACKNOWLEDGMENTS

This work was partially supported by National R&D project of China under contract No. 2021YFC3340304, the National Natural Science Foundation of China under Grant 62332010, the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology), Guangdong Basic and Applied Basic Research Foundation (2024A1515010454).

REFERENCES

[1] Eirikur Agustsson and Radu Timofte. 2017. NTIRE 2017 challenge on single image super-resolution: dataset and study. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*
[2] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. 2016. Density modeling of images using a generalized normalization transformation. In *Proc. Int'l Conf. Learn. Representations.*
[3] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. 2017. End-to-end optimized image compression. In *Proc. Int'l Conf. Learn. Representations.*
[4] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. 2019. Variational image compression with a scale hyperprior. In *Proc. Int'l Conf. Learn. Representations.*
[5] Fabrice Bellard. 2017. *BPG image format.* http://bellard.org/bpg/
[6] Yochai Blau and Tomer Michaeli. 2019. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *Proc. Int'l Conf. Mach. Learn.*
[7] Lorenzo Catania and Dario Allegra. 2023. NIF: A fast implicit image compression with bottleneck layers and modulated sinusoidal activations. In *Proc. ACM Int'l Conf. Multimedia.*
[8] Tong Chen, Haojie Liu, Zhan Ma, Qiu Shen, Xun Cao, and Yao Wang. 2021. End-to-end learnt image compression via non-local attention optimization and improved context modeling. *IEEE Trans. Image Process.* 30 (2021), 3179–3191.
[9] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. 2020. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*
[10] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. In *Proc. Annu. Conf. Neural Inf. Process. Systems.*
[11] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 5 (2020), 2567–2581.
[12] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2016. Density estimation using real NVP. In *Proc. Int'l Conf. Learn. Representations.*
[13] Yilun Du and Igor Mordatch. 2019. Implicit generation and modeling with energy based models. In *Proc. Annu. Conf. Neural Inf. Process. Systems.*
[14] Emilien Dupont, Hrushikesh Loya, Milad Alizadeh, Adam Goliński, Yee Whye Teh, and Arnaud Doucet. 2022. COIN++: Neural compression across modalities. *IEEE Trans. Mach. Learn. Res.* (2022).
[15] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. 2023. Generative diffusion prior for unified image restoration and enhancement. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*
[16] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. 2023. Implicit diffusion models for continuous super-resolution. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*
[17] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. 2015. MADE: Masked autoencoder for distribution estimation. In *Proc. Int'l Conf. Mach. Learn.*
[18] Noor Fathima Khanum Mohamed Ghouse, Jens Petersen, Auke J. Wiggers, Tianlin Xu, and Guillaume Sautiere. 2023. Neural Image Compression with a Diffusion-based Decoder. https://openreview.net/forum?id=4Jq0XWCZQel
[19] Bjøntegaard Gisle. 2001. Calculation of average PSNR differences between RD curves. In *VCEG-M33.* https://api.semanticscholar.org/CorpusID:61598325
[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.
[21] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. 2022. ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*
[22] Dailan He, Ziming Yang, Hongjiu Yu, Tongda Xu, Jixiang Luo, Yuan Chen, Chenjian Gao, Xinjie Shi, Hongwei Qin, and Yan Wang. 2022. PO-ELIC: Perception-oriented efficient learned image coding. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*
[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proc. Annu. Conf. Neural Inf. Process. Systems.*
[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proc. Annu. Conf. Neural Inf. Process. Systems.*
[25] Emiel Hoogeboom, Eirikur Agustsson, Fabian Mentzer, Luca Versari, George Toderici, and Lucas Theis. 2023. High-fidelity image compression with score-based generative models. *arXiv preprint arXiv:2305.18231* (2023).
[26] Yueyu Hu, Wenhan Yang, and Jiaying Liu. 2020. Coarse-to-fine hyper-prior modeling for learned image compression. In *Proc. AAAI Conf. on Artif. Intell.*
[27] Wei Jiang, Jiayu Yang, Yongqi Zhai, Peirong Ning, Feng Gao, and Ronggang Wang. 2023. MLIC: Multi-reference entropy model for learned image compression. In *Proc. ACM Int'l Conf. Multimedia.*
[28] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*
[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of StyleGAN. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*
[30] Diederik Pieter Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[31] Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. In *Proc. Annu. Conf. Neural Inf. Process. Systems.*
[32] Diederik P Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *Proc. Int'l Conf. Learn. Representations.*
[33] Eastman Kodak. 2024. *Kodak lossless true color image suite.* https://r0k.us/graphics/kodak/
[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proc. Annu. Conf. Neural Inf. Process. Systems.*
[35] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. 2017. Grammar variational autoencoder. In *Proc. Int'l Conf. Mach. Learn.*
[36] Théo Ladune, Pierrick Philippe, Félix Henry, Gordon Clare, and Thomas Leguay. 2023. COOL-CHIC: Coordinate-based low complexity hierarchical image codec. In *Proc. Int'l Conf. Comput. Vision.*
[37] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. 2018. Context-adaptive entropy model for end-to-end optimized image compression. In *Proc. Int'l Conf. Learn. Representations.*
[38] Jiaying Liu, Dong Liu, Wenhan Yang, Sifeng Xia, Xiaoshuai Zhang, and Yuanying Dai. 2020. A comprehensive benchmark for single image compression artifact reduction. *IEEE Trans. Image Process.* 29 (2020), 7845–7860.
[39] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Proc. Annu. Conf. Neural Inf. Process. Systems.*
[40] Guo Lu, Chunlei Cai, Xiaoyun Zhang, Li Chen, Wanli Ouyang, Dong Xu, and Zhiyong Gao. 2020. Content adaptive and error propagation aware deep video compression. In *Proc. Eur. Conf. Comput. Vision.*
[41] Yiyang Ma, Huan Yang, Wenhan Yang, Jianlong Fu, and Jiaying Liu. 2024. Solving diffusion ODEs with optimal boundary conditions for better image super-resolution. In *Proc. Int'l Conf. Learn. Representations.*
[42] Michael W Marcellin, Michael J Gormish, Ali Bilgin, and Martin P Boliek. 2000. An overview of JPEG-2000. In *Proc. Data Compression Conference.*
[43] Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson. 2020. High-fidelity generative image compression. In *Proc. Annu. Conf. Neural Inf. Process. Systems.*
[44] David Minnen, George Toderici, Michele Covell, Troy Chinen, Nick Johnston, Joel Shor, Sung Jin Hwang, Damien Vincent, and Saurabh Singh. 2017. Spatially adaptive image compression using a tiled deep network. In *Proc. IEEE Int'l Conf. Image Process.*
[45] David Minnen, George Toderici, Saurabh Singh, Sung Jin Hwang, and Michele Covell. 2018. Image-dependent local entropy models for learned image compression. In *Proc. IEEE Int'l Conf. Image Process.*
[46] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).
[47] Matthew Muckley, Alaaeldin El-Nouby, Karen Ullrich, Hervé Jégou, and Jakob Verbeek. 2023. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *Proc. Int'l Conf. Mach. Learn.*
[48] Savvas Panagiotou and Anna S Bosman. 2023. Denoising diffusion post-processing for low-light image enhancement. *arXiv preprint arXiv:2303.09627* (2023).
[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. Int'l Conf. Med. Image Comput. and Computer-Assisted Intervention.*
[50] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 4 (2022), 4713–4726.
[51] Hamid R Sheikh and Alan C Bovik. 2006. Image information and visual quality. *IEEE Trans. Image Process.* 15, 2 (2006), 430–444.
[52] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[53] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *Proc. Int'l Conf. Learn. Representations*.
[54] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. In *Proc. Int'l Conf. Learn. Representations*.
[55] Yannick Strümpler, Janis Postels, Ren Yang, Luc Van Gool, and Federico Tombari. 2022. Implicit neural representations for image compression. In *Proc. Eur. Conf. Comput. Vision*.
[56] George Toderici, Wenzhe Shi, Radu Timofte, Lucas Theis, Johannes Balle, Eirikur Agustsson, Nick Johnston, and Fabian Mentzer. 2020. Workshop and challenge on learned image compression. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit. Workshop*.
[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. Annu. Conf. Neural Inf. Process. Systems*.
[58] Gregory K Wallace. 1991. The JPEG still picture compression standard. *Commun. ACM* 34, 4 (1991), 30–44.
[59] Dezhao Wang, Wenhan Yang, Yueyu Hu, and Jiaying Liu. 2022. Neural data-dependent transform for learned image compression. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*
[60] Liyan Wang, Qinyu Yang, Cong Wang, Wei Wang, Jinshan Pan, and Zhixun Su. 2023. Learning a coarse-to-fine diffusion transformer for image restoration. *arXiv preprint arXiv:2308.08730* (2023).
[61] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *Proc. Asilomar Conf. Signals, Systems, Comp.*
[62] Ruihan Yang and Stephan Mandt. 2023. Lossy image compression with conditional diffusion models. In *Proc. Annu. Conf. Neural Inf. Process. Systems*.
[63] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. 2023. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proc. Int'l Conf. Comput. Vision*.
[64] Jeonghwan Yoon and Nam Ik Cho. 2023. JPEG artifact reduction based on deformable offset gating network controlled by a variational autoencoder. *IEEE Access* 11 (2023), 30282–30291.
[65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proc. Int'l Conf. Comput. Vision*.
[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*