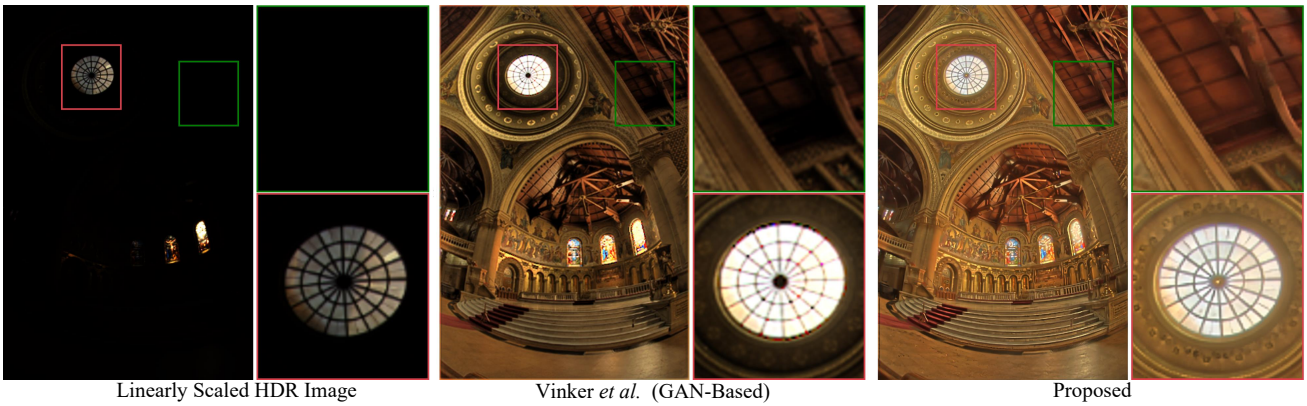
Zero-Shot Structure-Preserving Diffusion Model for High Dynamic Range Tone Mapping

Ruoxi Zhu[1,*], Shusong Xu[2,3,*], Peiye Liu[2,3], Sicheng Li[2,3], Yanheng Lu[2,3], Dimin Niu[2,3], Zihao Liu[2,3], Zihao Meng[2,3], Zhiyong Li[2,3], Xinhua Chen[1], Yibo Fan[1,†]
[1]Fudan University, [2]DAMO Academy, Alibaba Group, [3]Hupan Lab
rxzhu22@m.fudan.edu.cn



Linearly Scaled HDR Image | Vinker *et al.* (GAN-Based) | Proposed

Figure 1. Qualitative comparison of previous state-of-the-art [33] with our method.

Abstract

Tone mapping techniques, aiming to convert high dynamic range (HDR) images to high-quality low dynamic range (LDR) images for display, play a more crucial role in real-world vision systems with the increasing application of HDR images. However, obtaining paired HDR and high-quality LDR images is difficult, posing a challenge to deep learning based tone mapping methods. To overcome this challenge, we propose a novel zero-shot tone mapping framework that utilizes shared structure knowledge, allowing us to transfer a pre-trained mapping model from the LDR domain to HDR fields without paired training data. Our approach involves decomposing both the LDR and HDR images into two components: structural information and tonal information. To preserve the original image's structure, we modify the reverse sampling process of a diffusion model and explicitly incorporate the structure information into the intermediate results. Additionally, for improved image details, we introduce a dual-control network architecture that enables different types of conditional inputs to control different scales of the output. Experimental results demonstrate the effectiveness of our approach, surpassing previous state-of-the-art methods both qualitatively and quantitatively. Moreover, our model exhibits versatility and can be applied to other low-level vision tasks without retraining. The code is available at https://github.com/ZSDM-HDR/Zero-Shot-Diffusion-HDR.

1. Introduction

Tone mapping, a long-standing computer vision task, focuses on imaging high dynamic range (HDR) scenes by converting HDR images to high-quality low dynamic range (LDR) images. Unlike LDR, HDR images typically possess a wider bitwidth, resulting in greater diversity in the distribution of pixel values, which captures more information especially in extreme cases such as backlit scenes. While human eyes effortlessly perceive both HDR and LDR scenes, general displayers struggle to accurately present the abundant information contained in HDR images. Therefore, the crucial task of mapping this high range visual information into LDR spaces with both natural appearance and accurate structure becomes imperative for real-world vision systems.

In recent years, there has been a rise in the use of deep learning-based methods for tone mapping. However, obtaining the ideal paired training data is particularly challenging due to the inherent difficulty in capturing both HDR and high-quality LDR images for natural

This work was supported in part by the National Key R&D Program of China (2023YFB4502802), in part by the National Natural Science Foundation of China (62031009), in part by the "Ling Yan" Program for Tackling Key Problems in Zhejiang Province (No.2022C01098), in part by Alibaba Innovative Research (AIR) Program, in part by Alibaba Research Fellow (ARF) Program, in part by the Fudan-ZTE Joint Lab, in part by CCF-Alibaba Innovative Research Fund For Young Scholars.
*contribute equally. †Yibo Fan is the corresponding author.

scenes [3, 33]. To address this challenge, many researchers employ the adversarial learning framework in training their networks [25, 33, 41]. Despite the utilization of a unified loss function combining a natural style term and a structure-preserving term, the concurrent optimization of these objectives during training can diminish their individual effectiveness. This occurs due to their potential divergence in optimization directions. Consequently, the resulting output may fall short of neither a natural-look tone nor accurate structures, leading to an unsatisfying outcome. In addition, these methods require a large number of HDR images as training samples, which is also challenging as compared to readily available LDR image datasets [11].

Differently, we approach the issue of unpaired data by adopting a zero-shot method which aims to transfer a mapping model trained on the LDR domain to HDR domain with less readily training data. To achieve this, transferable knowledge is crucial to bridge the gap between the LDR and HDR domains, as shown in Fig.2. However, in the field of tone mapping, it remains an ongoing challenge to discover a shared knowledge that has an equivalent distribution in both the HDR and LDR domains.

From this perspective, we aim to identify suitable shared knowledge for addressing the zero-shot tone mapping issue. To achieve this, we propose a novel methodology that involves decomposing the original image into two distinct components: tonal information and structural information. Our analysis, illustrated in Fig.4, reveals that the structural information exhibits a high similarity in distribution over the HDR and LDR domains. This observation highlights the potential of the structural information as a unified shared knowledge that can bridge the gap between the two domains. Hence, we train a conditional diffusion model to generate images that have an immersive tone as high-quality LDR images and the same structure as the input original images under the guidance of the original structural information of images from LDR datasets. During inference, the structural information of the HDR image is extracted and fed to the model. The structural information has the same distribution over the HDR domain and the LDR domain, bridging the gap between training and inference. Under this framework, we further propose a dual-control network architecture to enhance the image quality. We also propose a structure refinement operation, seamlessly integrated into the reverse sampling iterations, that explicitly combines the structure of the original image and the tone of the predicted image. Additionally, except for HDR image tone mapping, some other low-level vision tasks can also be understood as changing the tone while preserving the structure, thus the proposed method can be used as a general solution to these tasks without re-training.

To summarize, our contributions are listed as follows:

- Under the zero-shot paradigm, we introduce a structure-preserving tone mapping framework that extracts the structure information in both HDR and LDR images as shared knowledge that guides the mapping model to yield images with both accurate structure and immersive tone.
- We propose a dual-control network architecture for better luminance retention and detail information retention.
- We propose a structure refinement operation that modifies the structure of intermediate results of the reverse sampling process.
- Experiments on the benchmark datasets illustrate the superiority of our method. We also apply our model to other tasks without re-training and achieve promising results.

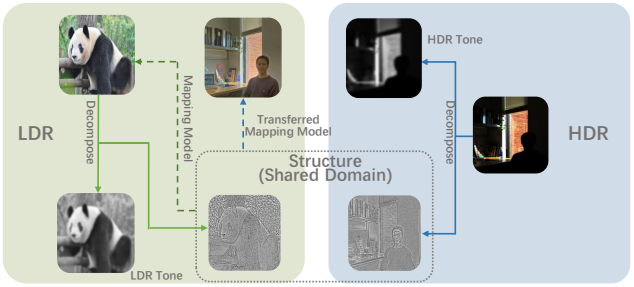


Figure 2. Schematic of our tone-structure decomposition and zero-shot framework. Structural information that distributes samely over HDR and LDR domains is utilized as the input of the network.

2. Related Work

Traditional Tone Mapping Methods. Tone mapping serves as a considerable technique for mapping the pixel values of HDR images to LDR counterparts. Due to the poor performance of simple linear scaling [36], early researchers use logarithmic tone curves [13], but these approaches exhibit limited performance due to contrast loss. Subsequently, some researchers have explored more complicated tone mapping methods [14, 24, 26] aligned with the responses of the human visual system, introducing other challenges such as the halo effect. Although numerous techniques have been put forward to mitigate the shortcomings of traditional tone mapping algorithms, such as bilateral filters [6], robust averaging [2], decomposition of image details using l0 and l1 sparsity [17], and operations at the fine scale of image gradients [29], the artifacts and deviations from reality still persist.

Deep Learning Based Tone Mapping Methods. Recently, deep learning (DL) has stepped into HDR tone mapping. Hou *et al.* [11] train a network to adjust HDR images based on their log-transformed luminance map. Panetta *et al.* view the tone mapping as an image enhancement task and train a network on a low-light image enhancement dataset. Recently, more works have emerged, focusing on

either delicate network architectures [4, 37] or more effective training strategies [3, 33].

Despite the outstanding performance that DL-based methods achieve on various low-level vision tasks, they suffer from the lack of paired training data when struggling for tone mapping. To tackle this, researchers have developed two main approaches. One approach [4] directly uses a no-reference image quality metric (such as normalized Laplacian pyramid distance) as the loss function, but risks overfitting this specific metric. More works [25, 32, 33, 40] belong to the other approach that adopts an adversarial training strategy, with an adversarial loss term to ensure the natural style and a reconstruction loss term to ensure structure consistency. These works can be further classified into the pseudo-paired scheme and the unpaired scheme. Among the former, Rana *et al.* [25] try several traditional tone mapping operators on each HDR image before training, rank them by tone mapped quality index (TMQI) [38], and pick the best-performing one as the reference to calculate the reconstruction loss. Whereas, the performance of the network is restricted by the traditional operators they use. Among the latter, Vinker *et al.* [33] proposed to calculate the reconstruction loss based on Pearson correlation that doesn't require an LDR reference. However, for both of the schemes, the commingling of the two distinct objectives, along with potential differences in their optimization directions, may result in their mutual dilution. This can lead to suboptimal outcomes where both the tone and structure of the output fall short of expectations.

Diffusion-based Generative Models. Denoising diffusion probabilistic models (DDPM) is a type of generative model that learns to yield desired data samples from a Gaussian noise through a multistep denoising process [10]. These models have demonstrated several advantages, including more stable training, resistance to overfitting, and a more interpretable latent space compared to GANs [5] and other earlier generative models. Diffusion denoising implicit models (DDIM) [31] further accelerate the sampling process for pre-trained diffusion models. Diffusion models have been applied across various tasks, such as super-resolution [34], denoising [12], image translation [28], inpainting [20] and low-light image enhancement [35]. More recently, some works [23, 39] focus on exploring versatile and effective approaches to adding control to foundation models such as Stable Diffusion [27], so as to acquire images with the desired contents more accurately.

3. Proposed Method

In this section, we present our zero-shot tone mapping framework with tone-structure decomposition. We begin by discussing the decomposition method in Section 3.1, which involves separating the HDR/LDR images into tonal and structural information. Next, in Section 3.2, we introduce the concept of utilizing the structure information as shared knowledge within a zero-shot framework. In the final Section 3.3, we present our structure-preserving framework for mapping HDR images to LDR images. The overall framework architecture is illustrated in Figure 3.

3.1. Tone-Structure Decomposition

Tone-structure decomposition, dubbed $TSD()$, is a technique introduced to disentangle the tonal information and structural information from the original image.

In the context of image processing, the term "structure" refers to the spatial relationships of visual elements within an image, which plays a crucial role in conveying information. Preserving the image structure is particularly important in tasks like tone mapping, as it ensures the retention of the original content features and details of the image.

Then, the rest "tone" refers to the distribution of light and dark values across different areas within the image. By adjusting the "tone" term, one can change the overall brightness, contrast, and distribution of tonal values. This helps achieve desired visual effects and enhances the image's appearance.

Therefore, to extract the "structure" information from the "tone" information, we adopt a value-independent representation, so-called mean subtracted contrast normalized coefficients (MSCN) [21]. We choose the mean value and standard deviation to represent the distributions of the tone. Moreover, we calculate them for each local patch to retain spatial information that is crucial in low-level vision tasks. Mathematically, in this paper, the tonal information of an image is represented by the local mean value map and the local standard deviation map, given as:

$$\boldsymbol{\mu}(i,j) = \sum_{x=-K}^{K} \sum_{y=-K}^{K} \boldsymbol{\omega}(x,y)\boldsymbol{I}(i+x, j+y), \tag{1}$$

$$\boldsymbol{\sigma}(i,j) = \sqrt{\sum_{x=-K}^{K} \sum_{y=-K}^{K} \boldsymbol{\omega}(x,y)(\boldsymbol{I}(i+x, j+y) - \boldsymbol{\mu}(i,j))^2}, \tag{2}$$

where w denotes a Gaussian filter kernel. To extract the structural information and reduce its correlation with the tonal information, we normalize the pixel values using the local mean values and local standard deviations, given as:

$$\hat{\boldsymbol{I}}(i,j) = \frac{\boldsymbol{I}(i,j) - \boldsymbol{\mu}(i,j)}{\boldsymbol{\sigma}(i,j) + \varepsilon} \tag{3}$$

where ε is a small value to avoid dividing by zero. This formula is the same as the definition of MSCN coefficients introduced in [21].

3.2. Zero-Shot Strategy for Tone Mapping

In this subsection, from the perspective of zero-shot approach, we explain how our proposed tone-structure decom-

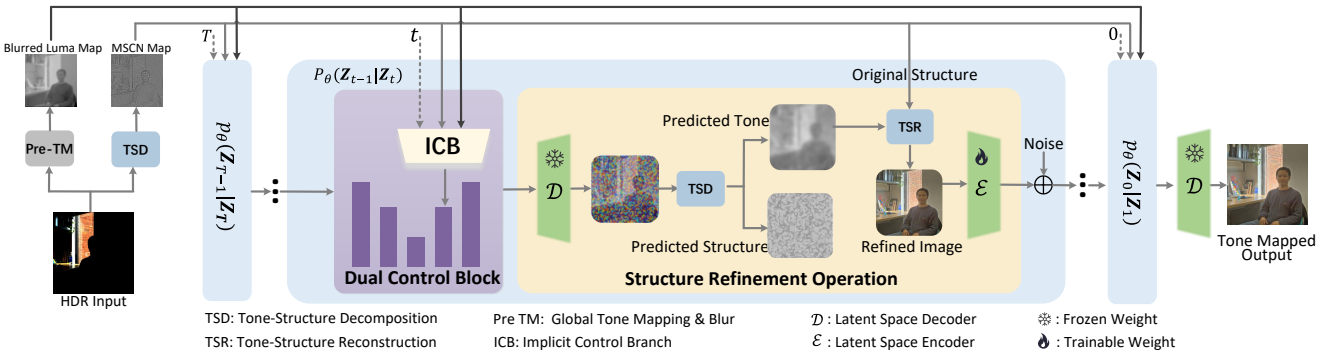


Figure 3. Inference pipeline of the proposed structure-preserving diffusion model. In each iteration, an intermediate result is first generated by a dual-control block, then modified by a structure refinement operation to acquire more accurate structural information.

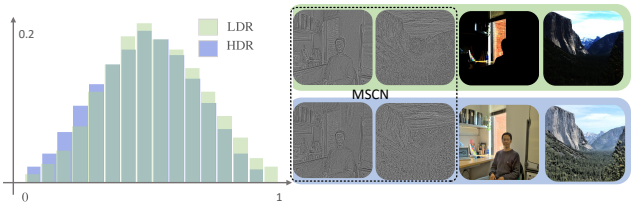


Figure 4. Left: distributions of MSCN coefficients (linearly normalized) over HDR and LDR datasets. Right: HDR image and its LDR counterpart produced by [33], together with their MSCN maps (enhanced for visualization).

position enables the model trained with natural LDR images to process HDR images.

To explain the rationality of the proposed tone-structure decomposition, an example is given in Fig.4, including HDR images and their LDR counterparts produced by the tone mapping method of [33], together with their MSCN maps. Despite the huge difference between the HDR and the tone mapped image, their MSCN maps are nearly the same, with only minor differences caused by imperfect tone mapping. Moreover, we calculated the local mean values, local standard deviations and MSCN coefficients of 105 HDR images from HDRPS dataset [7] and 100 high-quality natural LDR images randomly selected from Flick2K dataset [18] respectively. Bar graphs of MSCN coefficients are illustrated in Fig.4, showing that the distributions of MSCN coefficients are approximately the same over the two domains. To indicate the similarity of distributions quantitatively, we calculate the Jensen-Shannon (JS) divergence of local mean values, local standard deviations and MSCN coefficients, which are 0.365, 0.184 and 0.002 respectively. Considering that the JS divergence of the two distributions of MSCN is far below 0.1, which is a commonly used threshold to judge whether two distributions are similar, the MSCN coefficients can be approximately considered to be samely distributed over the HDR and LDR domains.

Therefore, we treat the extracted "structure" information from the HDR and LDR images as a form of shared knowledge, serving as a crucial bridge between the two domains, illustrated in Fig.2. To begin, we independently train a mapping model within the LDR domain, utilizing paired LDR images and the corresponding decomposed "structure" information as the ground truth and the input respectively. Subsequently, by sharing a shared distribution, we replace the LDR "structure" features with the HDR "structure" features. This substitution enables the transfer of the pre-trained model from the LDR domain to the HDR domain. Consequently, we obtain a mapping model that effectively transforms imperceptible HDR inputs into high-quality LDR outputs by leveraging the shared information contained in the "structure" features. By employing this approach, we exploit the extracted "structure" information to bridge the gap between HDR and LDR domains, facilitating the generation of superior LDR images from HDR inputs.

3.3. Structure Preserving Diffusion Framework

In this section, we detail our structure-preserving diffusion framework, including a generation process with a dual-control block and a structure refinement process in each sampling iteration, demonstrated in Fig.3.

3.3.1 Dual-Control Network Architecture

As shown in Fig.5, we utilize two different types of conditions to control the output. The main network can be bifurcated into a generation branch and an implicit control branch. The former is implemented as a fixed-weight Stable Diffusion v1.5, which samples a natural image from Gaussian noise in the latent space. The latter receives two different conditions at different scales to control the spatial information of different frequencies by different hints. The adopted pre-trained foundation model holds a wealth of prior tone information, enabling our model to yield more natural-look images.

Fine-Scale Control: Basically, the MSCN map is utilized as the main conditional input, which is fed to the first

layer of the implicit control branch. The intermediate results of each scale are added to the feature maps of the corresponding decoder layers of the generation branch, controlling the contents of the generated image in an implicit manner.

Coarse-Scale Control: However, MSCN maps are calculated by a local normalization operation that significantly eliminates the luminance information. Thus, if using the MSCN map as the only condition, though the structure of the output images is consistent with the original image, the luminance may suffer from distortion. For example, the relative intensity of luminance between different parts of the image may be incorrect, leading to an unnatural appearance. Thus, another type of conditional input containing luminance information is required.

An intuitive and straightforward approach is to directly use the luma map (Y channel in the YUV color space) of the original image as conditional input. However, this results in domain shift between training and inference, because an accurate luma map is not available during inference. Thus, we design a novel approach discussed below.

In the training phase, we use the Gaussian blurred luma map as luminance guidance. In the inference phase, similar to [33], we first adjust the Y channel of the HDR image using a global tone mapping curve given by

$$\boldsymbol{Y}_c(i,j) = \log(\lambda \frac{\boldsymbol{Y}(i,j)}{max(\boldsymbol{Y})} + \varepsilon)/\log(\lambda + \varepsilon), \quad (4)$$

where λ is decided by minimizing the cross entropy of the histogram of the HDR image and the histogram of natural LDR images:

$$\min_{\lambda} - \sum_{l} H_l(\boldsymbol{Y}_c) \log H_l(LDR). \quad (5)$$

Then the adjusted Y channel is blurred by Gaussian filter and input to the network. This strategy can solve the domain shift problem in that the luminance guidance is a low-pass luma map during both training and inference. The network is expected to generate fine-scale textures using the information contained in MSCN maps and to generate coarser-scale components by leveraging the low-pass luma maps.

Besides, to avoid the blurred luma map affecting the generation of image details, we input it at a coarser-grained scale compared with MSCN maps. More specifically, the luma map is first convoluted and downsampled, then concatenated with the feature maps of the first scale and fed to the second scale. This architecture ensures that the contents added to the finest scale of the generation branch are only decided by the MSCN map that contains abundant details.

3.3.2 Structure-Refinement Operation

Though MSCN maps are taken as the conditional input, it is an implicit constraint and cannot ensure that the result is completely consistent with the original structure. Therefore, we further propose a structure refinement operation (SRO), seamlessly integrated with the diffusion sampling process, by explicitly injecting the original structural information into both the intermediate results of each diffusion sampling iteration and the final output. This structure-preserving diffusion process is summarized in Alg.1, whose technical details are discussed below.

Consider a reverse diffusion sampling process that starts from a pure Gaussian noise. This process is random, thus the results may suffer from structure distortion even if conditional hints are input to the network. Besides, the structural difference between the generated intermediate results and the desired output may accumulate during the iterative process, which may lead to unacceptable distortion in the final output. Therefore, we propose to refine the structure of the intermediate results in the reverse sampling process to avoid the accumulation of structural distortion.

In each sampling iteration $p_\theta(z_{t-1}|z_t)$, we first subtract the predicted noise from the result of the previous iteration

Algorithm 1 Structure-Preserving Sampling Process

$\boldsymbol{\mu}_{ori}, \boldsymbol{\sigma}_{ori}, \hat{\boldsymbol{I}}_{ori} = TSD(\boldsymbol{I}_{ori})$
for $t = T, T-1, ..., 1$ **do**
  $z'_{t-1} = \frac{1}{\sqrt{\alpha_t}}(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(z_t, t, c))$
  if $t > t_0$ **then**
    $\boldsymbol{\mu}_{t-1}, \boldsymbol{\sigma}_{t-1}, \hat{\boldsymbol{I}}_{t-1} = TSD(\mathcal{D}(z'_{t-1}))$
    $z_{t-1} = \mathcal{E}(\boldsymbol{\mu}_{t-1} + \gamma \boldsymbol{\sigma}_{t-1} \hat{\boldsymbol{I}}_{ori}) + \varepsilon_t$
  else
    $z_{t-1} = z'_{t-1} + \epsilon_t$
  end if
end for
$\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, \hat{\boldsymbol{I}}_0 = TSD(\mathcal{D}(z'_0))$
$\boldsymbol{I}_{pred} = \boldsymbol{\mu}_0 + \gamma \boldsymbol{\sigma}_0 \hat{\boldsymbol{I}}_{ori}$
return $\boldsymbol{I}_{pred}$

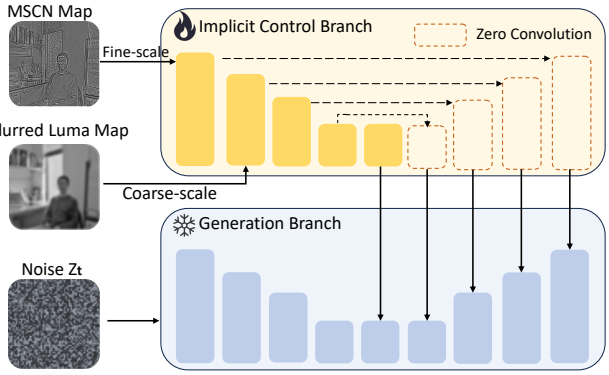


Figure 5. Network architecture of the dual-control block, which receives two types of conditions to control the generated image implicitly.

as the original DDPM does:

$$z'_{t-1} = \frac{1}{\sqrt{\alpha_t}}(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(z_t, t, c)). \quad (6)$$

The initial estimate of the intermediate embedding z'_{t-1} is then converted from the latent space to the image domain using the pre-trained decoder $\mathcal{D}$, and further decomposed into tonal information and structural information using our proposed tone-structure decomposition $TSD()$ introduced in 3.1, given as:

$$\boldsymbol{\mu}_{t-1}, \boldsymbol{\sigma}_{t-1}, \hat{\boldsymbol{I}}_{t-1} = TSD(\mathcal{D}(z'_{t-1})), \quad (7)$$

where $\mathcal{D}$ denotes the decoder part of the variational autoencoder used by Stable Diffusion. Among the three components of the predicted intermediate result, $\boldsymbol{\mu}_{t-1}$ and $\boldsymbol{\sigma}_{t-1}$ remain unchanged while the MSCN $\hat{\boldsymbol{I}}_{t-1}$ is replaced by the MSCN of the original HDR image $\hat{\boldsymbol{I}}_{ori}$. This step preserves the tonal style of the predicted image and refines its structure. The recombined components are then reconstructed to form a modified image. We introduced a tunable parameter γ that is multiplied by $\boldsymbol{\sigma}_{t-1}$ to control the strength of edge enhancement. Afterward, the modified intermediate result is converted to the latent space and mixed with random noise for the subsequent iterations:

$$z_{t-1} = \mathcal{E}(\boldsymbol{\mu}_{t-1} + \gamma\boldsymbol{\sigma}_{t-1}\hat{\boldsymbol{I}}_{ori}) + \varepsilon_t, \quad (8)$$

where $\mathcal{E}$ is an auxiliary encoder trained along with the implicit control branch to match $\mathcal{D}$, and ε_t is timestep-related random noise.

Besides, in practice, we found that if the SRO is conducted in each iteration, though the hallucinating artifacts can be reduced in the final result, the contrast and edge intensities of the whole image would also be reduced. This may be ascribed to the excessively strong constraint imposed by the SRO, diminishing the diversity of the generated contents. Hence, to trade-off between structural accuracy and subjective visual effect, we empirically set a critical point t_0, and the SRO would not be conducted after the timestep reaches t_0, allowing the diffusion model to generate more visually pleasing textures in the last few iterations.

Note that the previous discussion and the illustration of Fig. 3 is based on DDPM sampling process for simplicity. In case of DDIM in real practice, the SRO is slightly modified, which is detailed in our supplementary material.

4. Experiments

In this section, we first illustrate the implementation details of the proposed method. Then we compare our results with previous state-of-the-art tone mapping methods both qualitatively and quantitatively. Ablation studies are then conducted to validate the contributions of the proposed designs. Besides, we apply our model to a different task without retraining, showing the generalization ability of our method.

Table 1. Quantitative comparisons on the HDRPS dataset. The TMQI scores are taken from [37]

Method	TMQI↑	NIQE↓	HDR-free
Liang *et al.* [16]	0.8650	-	-
Shibata *et al.* [30]	0.877	-	-
ETMO [32]	0.8652	-	✗
DeepTMO [25]	0.88	2.519	✗
Vinker *et al.* [33]	0.8861	3.174	✗
LA-Net [37]	0.8759	2.524	✗
Proposed	**0.8915**	**2.340**	✓

4.1. Implementation Details

We train our model on the Flickr2K dataset [18] consisting of 2650 high-quality natural LDR images. The MSCN maps and low-pass luma maps of the images in this dataset are used as conditional inputs in the training phase, and no HDR images are required during training. In the initial 200 epochs, we only train the implicit control branch using the AdamW optimizer [19] with batch size 8 and learning rate $1e^{-5}$. The auxiliary encoder is not included in the network in this stage. Afterward, we plug the decoding-encoding process into the pipeline and train the encoder jointly with the control branch for another 200 epochs. The weights of the generation branch are always fixed. More settings of hyperparameters can be found in the supplementary material.

Pre-Processing: During inference, considering that the diffusion models are adept at processing images of fixed sizes, we first divide the image into overlapping patches with the same size as the training patches. Each of the patches is generated respectively, and the overlapped parts are weighted-summed to avoid discontinuity.

Post-Processing: After the whole image is generated, we recover each color channel using a formula widely adopted by previous methods [4, 8, 33]:

$$\boldsymbol{C}_{out}^{(i)} = \boldsymbol{Y}_{pred}(\boldsymbol{C}_{in}^{(i)}/\boldsymbol{Y})^s, \quad (9)$$

where $i \in \{R, G, B\}$ denotes the index of the color channel, $\boldsymbol{Y}$ and $\boldsymbol{Y}_{pred}$ are the Y channel of the original HDR image and the output of the diffusion model respectively.

4.2. Quantitative Comparisons

We test our model on the HDRPS dataset [7], a benchmark test set including 105 HDR images. We choose state-of-the-art deep learning (DL) based tone mapping methods for comparison: DeepTMO [25], Vinker *et al.* [33] and LA-Net [37]. Also, we compare to state-of-the-art DL-free methods: Liang *et al.* [16] and Shibata *et al.* [30]. We calculate the TMQI [38] and NIQE [22] as metrics of their performance, listed in Tab.1. TMQI is an objective image quality metric for tone mapped images, taking both structural fidelity

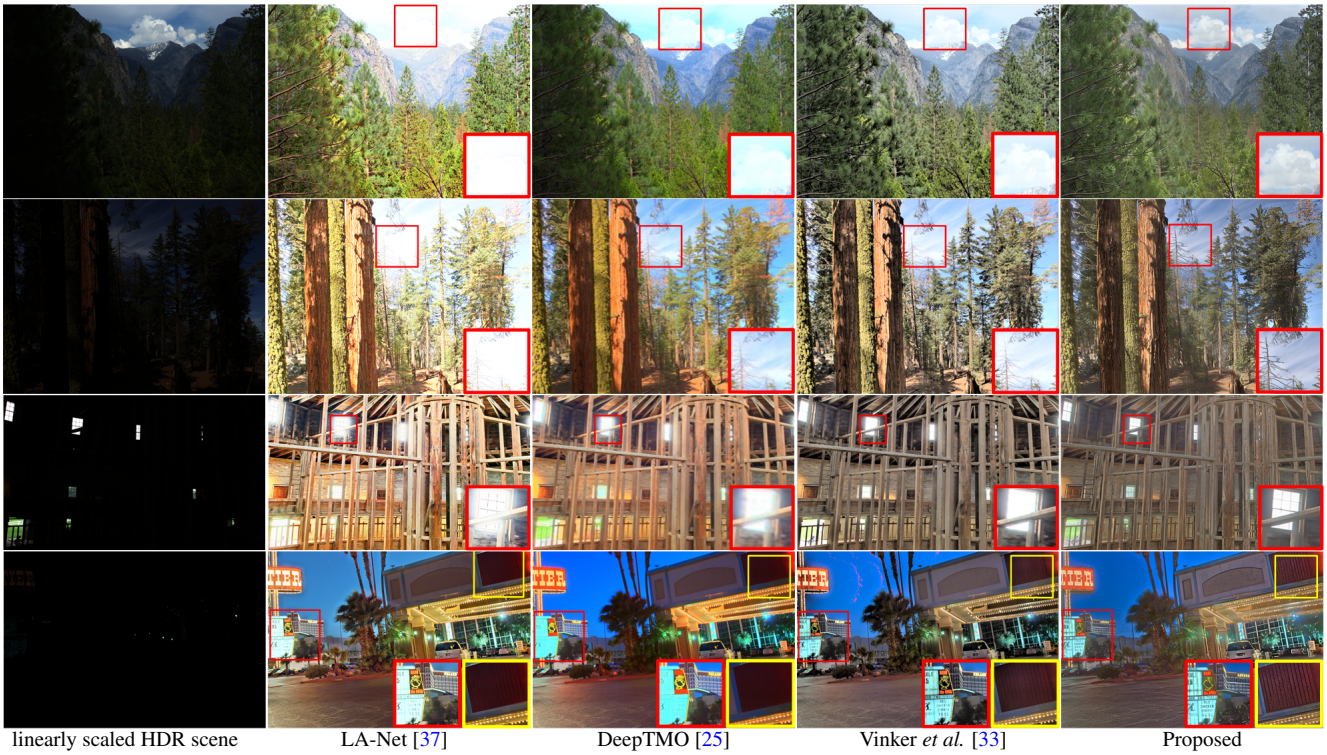


Figure 6. Visual comparisons on HDRPS dataset [7]. Our method performs best in both bright regions and dark regions.

and statistical naturalness into consideration. NIQE is a no-reference image quality metric based on natural scene statistic model, which is suitable for evaluating the tone mapping results in that no ground truth is available.

As illustrated in Tab.1, our algorithm performs best on both metrics, indicating that the proposed method is superior to previous methods in terms of both structure retention and the naturalness of appearance. Moreover, our algorithm is a zero-shot framework, requiring only LDR training data that are more readily available than HDR images.

4.3. Qualitative Comparisons

We demonstrate the visual results on the benchmark dataset in Fig.6. As depicted, LA-Net[37] and DeepTMO[33] tend to produce images with higher saturation, but the structure of bright regions is hardly visible (bounded in red boxes), resulting in considerable information loss. Vinker *et al.*'s method performs slightly better in bright regions, but the image details in dark regions (bounded in yellow boxes) are still not satisfying enough, and unacceptable artifacts sometimes occur (note the bottom row). Our method, however, can preserve or even enhance the details in both bright and dark regions, yielding more informative images compared with previous state-of-the-arts. More results are included in the supplementary material.

Moreover, in Fig.7, we illustrate the results on some images with extremely high luminance contrast from HDR+ dataset [9]. Previous methods encounter difficulties in such scenarios, leading to significant banding artifacts and texture loss. However, our method successfully produces visually appealing results, which can be attributed to the proposed tone-structure decomposition that reduces the correlation between structure and exposure.

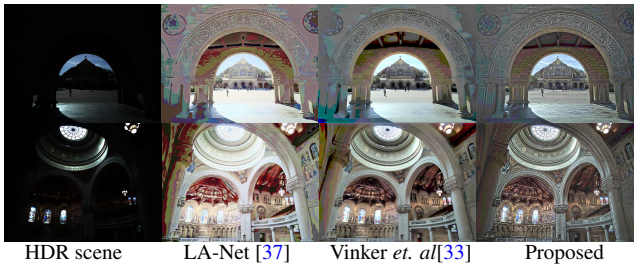


Figure 7. Visual comparisons on HDR+ dataset.

4.4. Ablation Study

To validate the contribution of the proposed network architecture (introduced in 3.3.1) and structure-refinement operation (introduced in 3.3.2) respectively, we conducted ablation studies.

Dual-Type Dual-Scale Conditional Input. To evaluate the efficacy of the proposed dual-type dual-scale con-

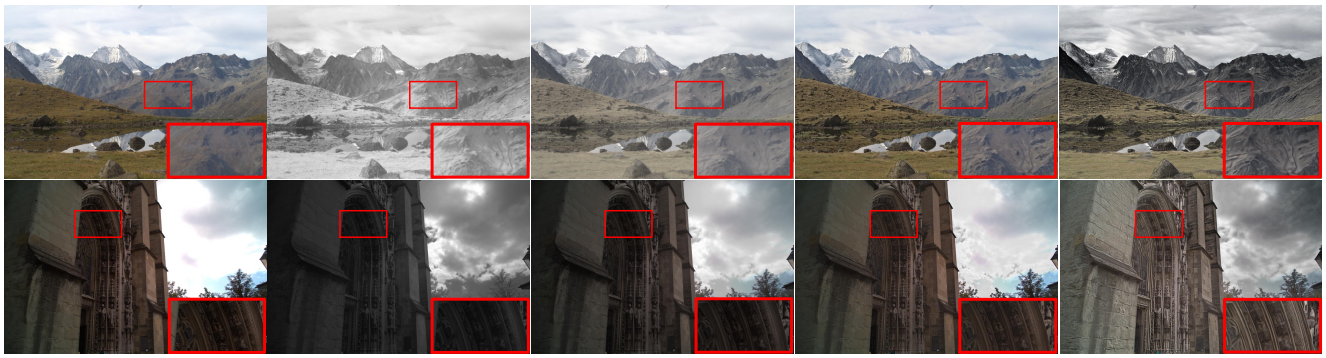


Figure 8. Comparisons of different VIS-NIR fusion methods on VIS-NIR Scene dataset [1]. Our method produces highly detailed images.

Figure 9. Ablation study of dual-type dual-scale conditional input. The proposed architecture generates images with more natural luminance and more abundant details.

ditional input architecture, we train three versions of networks: (a). only using MSCN as conditional input; (b). inputting MSCN and luma map at the same scale; (c). inputting MSCN and luma map at different scales. Visual comparisons are shown in Fig.9, indicating that the generated image would suffer from luminance distortion if only inputting MSCN maps. One failure case is shown in the left of each group, where the lit letters appear to be darker than the background, contradicting the real-world case. The generated luminance would appear more realistic if feeding luma maps additionally. The proposed architecture further enhances the image details by inputting the low-pass luminance map in a coarser scale, bounded in the yellow box.

Structure-Preserving Diffusion Process with SRO. In our default implementation, the SRO is conducted in the earlier iterations and the last iteration of DDIM. To study the potential alternatives, we try different strategies by conducting SRO at different iterations in the reverse sampling process. Quantitative results are reported in Tab.2, indicating that our default strategy leads to the best image quality.

4.5. Generalization to Other Tasks

Except for HDR image tone mapping, our model can be directly applied to some other tasks without re-training, owing to the generalization ability of the proposed algorithm. Take visible (VIS) and near-infrared (NIR) image fusion as an example, which aims to yield an image combining the structure of both modalities while remaining the same style as the VIS image. This is similar to tone mapping in that the structure and tone should be processed differently. We first simply average a pair of VIS and NIR images to generate a reference image that contains all of the structures but has a dissatisfying tone, and then use this reference image as the input of our model to produce a visually pleasing image.

Table 2. Ablation study on structure-refinement operation indicates our default strategy of SRO is the most effective.

Strategy	TMQI↑
no SRO	0.7210
at last iteration	0.8257
at earlier iterations	0.7297
at all iterations	0.8260
proposed ($t_0 = 10, T = 20$)	**0.8915**

The output images are shown in Fig.8, along with the results of previous methods. The proposed method can better combine the detail of the VIS and NIR images and enhance them, producing superior results to previous methods. The promising results indicate that our structure-preserving diffusion model has a strong generalization ability. We believe that our method can be utilized as a general approach for image enhancement.

5. Conclusion

In this paper, we propose a structure-preserving diffusion model mainly for HDR image tone mapping, tailored in a zero-shot framework that requires no HDR images during training. A tone-structure decomposition method is proposed for both structure-preserving and mapping the source and target samples to a shared domain. The network architecture and reverse sampling steps are further modified for better image quality. Results on the benchmark dataset for HDR tone mapping show the superiority of our method, and results on other tasks indicate the strong generalization ability of the method.

References

[1] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR) 2011*, pages 177–184, 2011. 8

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Fredo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *Conference on Computer Vision and Pattern Recognition (CVPR) 2011*, pages 97–104, 2011. 2

[3] Cong Cao, Huanjing Yue, Xin Liu, and Jingyu Yang. Unsupervised hdr image and video tone mapping via contrastive learning. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 2, 3

[4] Yuming Fang Chenyang Le, Jiebin Yan and Kede Ma. Perceptually optimized deep high-dynamic-range image tone mapping. *arXiv:2109.00180*, 2021. 3, 6

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 3

[6] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 257–266, 2002. 2

[7] M.D. Fairchild. The hdr photographic survey. *Color and Imaging Conference*, 15:233–238, 2007. 4, 6, 7

[8] Raanan Fattal, Dani Lischinski, and Michael Werman. Gradient domain high dynamic range compression. *ACM Trans. Graph.*, 21(3):249–256, 2002. 6

[9] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 35(6), 2016. 7

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 3

[11] Xianxu Hou, Jiang Duan, and Guoping Qiu. Deep feature consistent deep image transformations: Downscaling, decolorization and hdr tone mapping. *arXiv preprint arXiv:1707.09482*, 2017. 2

[12] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, pages 23593–23606. Curran Associates, Inc., 2022. 3

[13] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36:1–12, 2017. 2

[14] Edwin Land and John McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11, 1971. 2

[15] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. 8

[16] Zhetong Liang, Jun Xu, David Zhang, Zisheng Cao, and Lei Zhang. A hybrid l1-l0 layer decomposition model for tone mapping. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4758–4766, 2018. 6

[17] Zhetong Liang, Jun Xu, David Zhang, Zisheng Cao, and Lei Zhang. A hybrid l1-l0 layer decomposition model for tone mapping. In *Conference on Computer Vision and Pattern Recognition (CVPR) 2018*, 2018. 2

[18] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017. 4, 6

[19] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 6

[20] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 3

[21] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21 (12):4695–4708, 2012. 3

[22] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 6

[23] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[24] Sumanta Pattanaik, James Ferwerda, Mark Fairchild, and Donald Greenberg. A multiscale model of adaptation and spatial vision for realistic image display. *Computer Graphics (SIGGRAPH 98 Conference Proceedings)*, 32, 1998. 2

[25] Aakanksha Rana, Praveer Singh, Giuseppe Valenzise, Frederic Dufaux, Nikos Komodakis, and Aljosa Smolic. Deep tone mapping operator for high dynamic range images. *IEEE Transactions on Image Processing*, 29:1285–1298, 2020. 2, 3, 6, 7

[26] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 21, 2002. 2

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3

[28] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021. 3

[29] Takashi Shibata, Tanaka Masayuki, and Masatoshi Okutomi. Gradient-domain image reconstruction framework with intensity-range and base-structure constraints. In *Conference on Computer Vision and Pattern Recognition (CVPR) 2016*, pages 2745–2753, 2016. 2

[30] Takashi Shibata, Masayuki Tanaka, and Masatoshi Okutomi. Gradient-domain image reconstruction framework

with intensity-range and base-structure constraints. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2745–2753, 2016. 6

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2022. 3

[32] Chien-Chuan Su, Ren Wang, Hung-Jin Lin, Yu-Lun Liu, Chia-Ping Chen, Yu-Lin Chang, and Soo-Chang Pei. Explorable tone mapping operators. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10320–10326, 2021. 3, 6

[33] Yael Vinker, Inbar Huberman-Spiegelglas, and Raanan Fattal. Unpaired learning for high dynamic range image tone mapping. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14637–14646, 2021. 1, 2, 3, 4, 5, 6, 7

[34] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 3

[35] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C. Kot, and Bihan Wen. Exposurediffusion: Learning to expose for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12438–12448, 2023. 3

[36] Greg Ward. A contrast-based scalefactor for luminance display. In *Graphics gems*, 1994. 2

[37] Kai-Fu Yang, Cheng Cheng, Shi-Xuan Zhao, Hong-Mei Yan, Xian-Shi Zhang, and Yong-Jie Li. Learning to adapt to light. *Int. J. Comput. Vision*, 131(4):1022–1041, 2023. 3, 6, 7

[38] Hojatollah Yeganeh and Zhou Wang. Objective quality assessment of tone-mapped images. *IEEE Transactions on Image Processing*, 22(2):657–667, 2013. 3, 6

[39] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[40] Ning Zhang, Chao Wang, Yang Zhao, and Ronggang Wang. Deep tone mapping network in hsv color space. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2019. 3

[41] Ning Zhang, Chao Wang, Yang Zhao, and Ronggang Wang. Deep tone mapping network in hsv color space. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2019. 2

[42] Ruoxi Zhu, Yi Ling, Xiankui Xiong, Dong Xu, Xuanpeng Zhu, and Yibo Fan. Luminance-preserving visible and near-infrared image fusion network with edge guidance. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1155–1159, 2023. 8