Exposure-Limited Image Enhancement with Generative Diffusion Prior

Baiang Li, Sizhuo Ma, Yanhong Zeng, Xiaogang Xu, Youqing Fang, Zhao Zhang, *Senior Member, IEEE*, Jian Wang* and Kai Chen*

Abstract—Many consumer cameras are equipped with 8-bit image sensors, which often struggle to capture scenes with a High Dynamic Range (HDR). This limitation can result in overexposed or underexposed regions, a loss of fine details due to low bit-depth compression, skewed color distributions, and noticeable noise in dark areas. Traditional Standard Dynamic Range (SDR) image enhancement methods typically focus on color mapping by expanding the color range and adjusting brightness. However, they often fail to restore details in *dynamic range extremes*, *i.e.* regions where pixel values approach the minimum or maximum limits. We define "exposure-limited image enhancement" as the process of enhancing images with large missing areas due to exposure issues within the SDR space, which differs from existing "mis-exposed image enhancement" methods primarily aimed at correcting color distributions. To enhance these exposure-limited images and overcome the limitations of current models, we propose a novel two-stage approach. In the first stage, we remap color and brightness to a suitable range while preserving existing details. In the second stage, we use a diffusion prior to generate content in severely overexposed or underexposed regions, which are otherwise lost during capture. Notably, this generative refinement module can also serve as a plug-and-play component alongside existing enhancement methods. Extensive experiments demonstrate that our method significantly improves image quality and detail, outperforming state-of-the-art techniques in dynamic range extremes. The project page is at https://Sagiri0208.github.io.

Index Terms—Computational Photography, Image Enhancement, Generative Methods

1 Introduction

Real-world scenes often feature broad dynamic ranges, yet many smartphone cameras are equipped with 8-bit image sensors with limited dynamic ranges. As a result, these sensors cannot simultaneously capture details in both the bright sun and shaded leaves. One popular strategy for increasing the dynamic range is exposure bracketing [6], which merges multiple low-exposure and high-exposure shots into a single high dynamic range (HDR) image. However, this approach must be enabled at capture time, extends the shooting duration, and requires computationally intensive motion compensation.

Deep neural networks have been developed to restore and enhance overexposed and underexposed regions from a *single* image. This is usually done by either reconstructing an HDR image [4] followed by a tone mapping step, or by directly predicting an enhanced image [1]. However, these methods still struggle to consistently deliver a truly satisfying visual experience [7], [8]. The challenge is particularly pronounced in *dynamic range extremes*—areas where the pixel values are close to minimum or the maximum possible values, as illustrated in Figure 1. Although Convolutional Neural Networks (CNNs) and transformers excel at tone mapping (e.g., low-light enhancement [9] and denoising [10]) in dark regions, they are not capable of reconstructing large areas of content where data is essentially lost at capture. Consequently, existing methods often produce blurry, unnatural content in such regions (Figure 1(c)). In this work, we aim to extend the capabilities of *exposure-limited image enhancement*, which we define as the integration of tone mapping, noise reduction for low light conditions, detail compensation due to low bit-depth, and *generating* image details obscured or completely lost due to the camera's restricted dynamic range.

Large-scale generative models trained on extensive text-image pairs [11], [12] such as Stable Diffusion [13] excel at synthesizing realistic images and can potentially serve as a powerful tool for generating the details in dynamic range extremes. Motivated by the recent success in leveraging diffusion prior in low-level vision tasks, we propose a two-stage framework tailored to *exposure-limited SDR image enhancement*. This framework takes an SDR image as input and directly predicts an output SDR image with enhanced color distribution and details. In the first stage, we perform a global adjustment using **L**atent-**S**winIR$_c$ (LS), a transformer-based model [14] designed to harmonize color distributions and re-map extremely bright or dark areas into a visually pleasing range[1]. This is accomplished through a carefully defined color-mapping loss computed over color histograms.

In the second stage, the initially color-adjusted image is

B. Li, Y. Zeng, Y. Fang and K. Chen are with Shanghai AI Laboratory, Shanghai, China. Also, B. Li is with Hefei University of Technology, Hefei 230601, China.(e-mails: ztmotalee@gmail.com, zengyh1900@gmail.com, fangyouqing@pjlab.org.cn, chenkai@pjlab.org.cn).
X. Xu is with The Chinese University of Hong Kong, Hong Kong 100871, China; affiliated with Zhejiang University, Hangzhou 310058, China. (email: xiaogangxu00@gmail.com)
S. Ma and J. Wang are with Snap Research, Snap Inc., New York, NY 10036, USA. (e-mails: sizhuoma@gmail.com, jwang4@snapchat.com)
Z. Zhang is with Hefei University of Technology, Hefei 230601, China. (email: cszzhang@gmail.com)
** J. Wang and K. Chen are the co-corresponding authors. Besides, J. Wang initialized the project.*

1. This process is analogous to tone mapping, except that our input is also an SDR image.

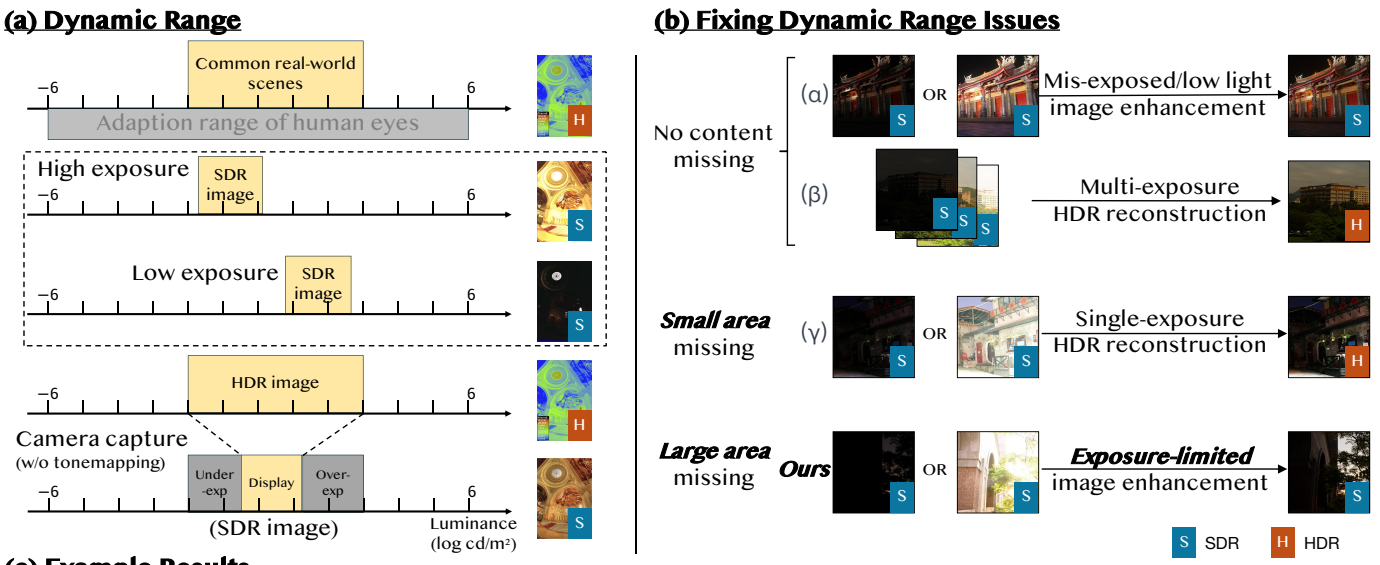


Fig. 1: **(a)** Real-world scenarios have broad dynamic ranges. However when captured by normal low bit mobile camera, the chosen exposure may often face over-saturated bright regions or heavily quantized dark areas with strong noise. **(b)** Existing works on mis-exposed/low light image enhancement [1], [2], multi-exposure HDR reconstruction [3] and single-exposure HDR reconstruction [4], [5] can enhance SDR images with no or small area of missing content, they are not designed to deal with ***large*** *dynamic range extreme areas*. **(c)** As a result, such methods often create blurry and unnatural imagery at dynamic range extremes. In this work, we aim to enhance these exposure-limited situations by analyzing and decomposing the complex task into the following challenges: (a) color and brightness mapping, (b) denoising of dark areas, (c) detail enhancement in low bit-depth regions, and (d) content generation for saturated or near-black regions.

refined by our diffusion-based model (named **Sagiri**). Even though the images are mapped to normal color distribution, content lost due to exposure limitations remains missing. Sagiri utilizes the powerful generative capabilities of pre-trained diffusion models with ControlNet [15] to further enhance the partially restored content and synthesizes realistic image details in the dynamic range extremes. Trained using a carefully crafted synthetic degradation strategy, Sagiri can also function as a plug-and-play module to improve results from existing SDR image enhancement and HDR reconstruction methods. Additionally, we introduce an adaptive regional processing technique during sampling, allowing users to guide content generation through custom prompts such as text or pixel masks. It is important to note that we intentionally limit our scope to enhancing SDR images to leverage the powerful diffusion models pretrained on SDR images. Extending this technique to the HDR domain could enable the use of flexible tone-mapping operators, which we leave for future work

Our key contributions are summarized as follows:

- We present LS-Sagiri, a novel two-stage approach for exposure-limited image enhancement. Stage 1 adjusts the overall color and brightness, while Stage 2 refines and generates missing details.
- We integrate a powerful generative diffusion based model to generate realistic content in saturated or black regions, while also enhancing fine details in areas affected by low bit-depth. Additionally we adopt a two-step training strategy to make it a plug-and-play module, effectively enhancing existing methods.
- Comprehensive experiments demonstrate that our approach delivers superior visual and quantitative performance and can flexibly enhance results from a wide range of existing techniques.

2 Related Work

In this section, we review related work on similar tasks and highlight their differences from the exposure-limited image enhancement problem we address. Additionally, we discuss image inpainting methods in supplementary material, which, although relevant, cannot directly substitute for the proposed Sagiri network.

2.1 Mis-Exposed and Low Light Image Enhancement

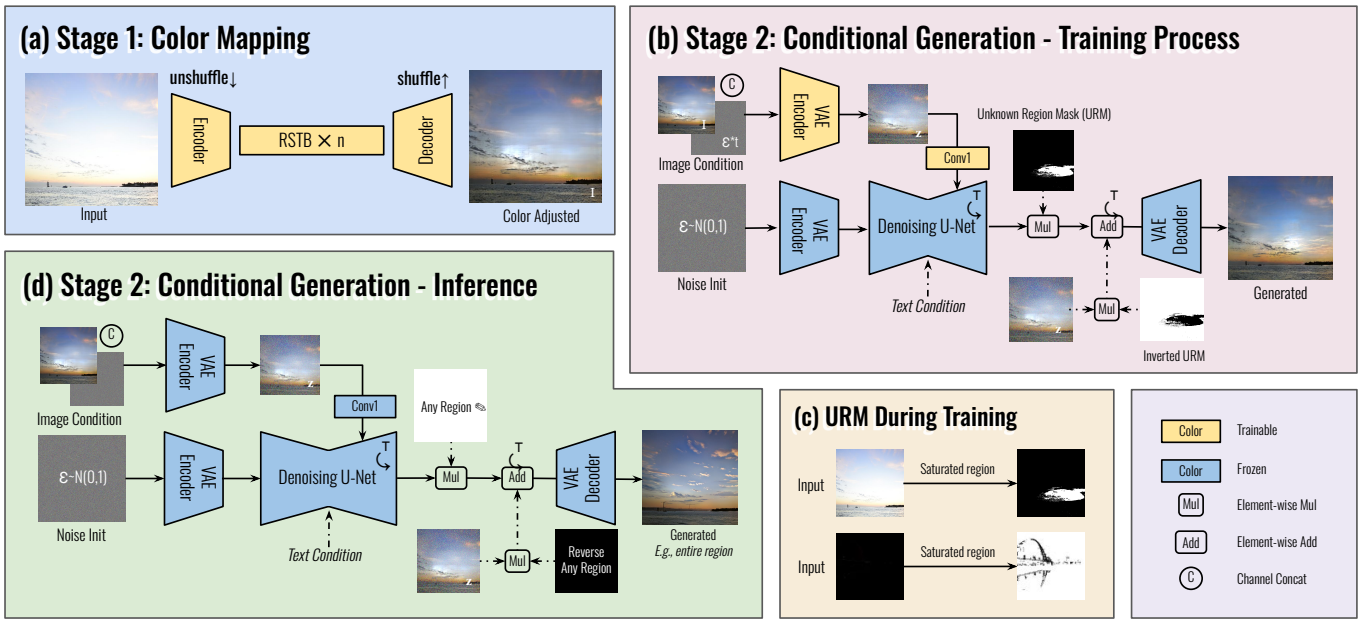


Fig. 2: Method overview. (a) Given a degraded input, the Stage 1 model performs color mapping to adjust the entire image to a more balanced color distribution. (b) In the Stage 2, the color-adjusted image is concatenated with a random noise map and sent to the parallel VAE encoder. After a shape-adjusting convolutional layer, the encoded latent feature map is then sent to the decoder in the denoising U-Net. (c) A default unknown region mask (the region where the pixel has the maximum or minimum value is selected) is used during training to indicate the dynamic range extremes. (d) During inference, the users can define their own mask. Note that the input and output of each stage is both SDR image. **Please zoom in on the images to observe the differences.**

Mis-Exposed Image Enhancement. Rather than attempting to recover or generate HDR images, some research focuses on correcting the exposure of single SDR images to produce outputs directly within the SDR domain. Wang et al. [1] observe that an image's local color distributions often contain both overexposed and underexposed areas and propose a method to *directly enhance these regions within the LDR domain*. Li et al. [16] further notice that these error exposed regions display opposite color tone distribution shifts and further propose a model to handle the shift. These methods are functionally similar to our first-stage model Latent-SwinIR$_c$ and do not have strong generative capabilities to fill in the missing details in *dynamic range extremes*.

Low Light Image Enhancement. Low light image enhancement is another related task that focuses on color mapping and noise reduction for low-light images within the SDR domain. Early methods like [17], [18] face limitations when addressing complex real-world scenarios, which later proposed deep learning methods have improved upon. LLNet [2] utilizes a deep autoencoder to transform underexposed images into enhanced versions, simultaneously reducing noise and increasing brightness. Retinex-Net [19] builds on the classical Retinex theory by decomposing images into reflectance and illumination components, allowing for targeted brightness adjustment and effective denoising. More advanced CNN and Transformer-based methods, such as [20], [21], [22], [23] and generative methods include [24], [25] have been proposed recently. However, these methods lack the capability to generate content in large blank regions and are not designed to address overexposure, thus they are not included in our comparison.

2.2 Single Image HDR Reconstruction

Non-Generative Methods. As the reconstruction of HDR images from a single input primarily involves color mapping and pixel value prediction without large exposure-limited region, many methods approach this task as a restoration problem without employing generative networks. Among these, SingleHDR [4] inverts the SDR image formation pipeline to recover HDR information. However, pipeline-based methods can lead to error accumulation at each intermediate step. HDRUNet [5] learns an end-to-end mapping for single-image HDR reconstruction, featuring denoising and dequantization. RawHDR [7] targets raw images by learning exposure masks to address challenging high dynamic range regions. Le et. al explores another direction by reconstructing images through multi-exposure generation [26], predicting images at various exposure levels from the input image and subsequently fusing them using standard HDR merging strategies. Other representative methods include [27], [28], [29], [30]. Although these methods have advanced HDR image reconstruction, their primary focus is not on addressing large exposed regions, resulting in weaker performance when applied to our specific task.

Generative Methods. Recent generative models offer new possibilities for HDR reconstruction. Fei et al. [31] introduce a diffusion-based framework for unsupervised restoration and enhancement, utilizing hierarchical guidance and patch-level operations to produce high-quality results. However, it requires multiple LDR inputs and involves a lengthy inference process. GlowGAN [8] uses a generative adversarial network to learn HDR content from in-the-wild LDR images without explicit supervision but

struggles with *large* overexposed regions. In contrast, our second-stage Sagiri model leverages a diffusion prior to effectively address this challenge.

3 METHOD

Our method aims to restore and enhance exposure-limited images using a two-stage framework. In the first stage, we employ a restoration model to adjust brightness and color distributions, yielding an image whose color statistics more closely resemble ground truth (GT) data. However, this restoration model alone has limited generative capacity and struggles to reconstruct details lost in *dynamic range extremes*. To address this, we introduce a second-stage model that leverages a diffusion prior to generate and refine missing or severely degraded regions.

Specialized Loss Functions. To effectively guide the learning process at each stage, we design distinct loss functions suited to each model's strengths. In Stage 1, our *color reconstruction loss* prioritizes brightness adjustment and color alignment by matching the predicted image's color histogram to that of the target image. In Stage 2, we define a *content enhancement loss* focused on generating high-fidelity textures and shifting the content distribution closer to that of detail-rich references. This encourages the model to synthesize missing details effectively.

3.1 Color Mapping

Our first-stage model is based on SwinIR [14] but with modified pre- and post-processing. Specifically, we apply a pixel unshuffle operation to downsample the original low-quality input by a factor of 8. Subsequently, a 3×3 convolutional layer extracts shallow features in color space. These features pass through Residual Swin Transformer Blocks (RSTB) for processing. Nearest-neighbor upsampling and another 3×3 convolutional layer are then repeated three times to return the features to the original resolution. We refer to this modified model as *Latent-SwinIR$_c$* (seen in top left of the Figure 2), which focuses on global color and brightness adjustment while offering preliminary content recovery. The core processing is performed in a downsampled latent space to reduce the computational cost, as the network must take the entire image as input rather than tiles to capture the color distribution accurately. To improve color mapping and brightness adjustment, we introduce the following loss function:

$$L_{\text{color}} = \lambda_1 L_{\text{mse}} + \lambda_2 L_{\text{cd}} + \lambda_3 L_{\text{fdp}}, \tag{1}$$

where L_{mse}, L_{cd}, and L_{fdp} are the MSE, Color Distribution, and Frequency Domain Preservation losses, respectively. The coefficients λ_1, λ_2, and λ_3 balance each component. *Further details are provided in the supplementary material.*

3.2 Conditional Generation

After the first stage, the restored image may still contain regions of poor visual quality, particularly in dynamic range extremes. Traditional methods often struggle to synthesize high-fidelity details in these areas [1], [7]. We harness the generative power of a pretrained diffusion model to overcome this limitation (Figure 2).

First, we encode the Stage 1 output via a Variational Autoencoder (VAE) [32] to obtain a latent representation. This latent is then combined with noise and fed into a parallel encoder module, which mirrors the encoder architecture in the U-Net denoiser. The outputs from different encoder blocks serve as latent controls, concatenated with the U-Net's decoder features. Newly introduced parameters are initialized to zero, while the pretrained denoising U-Net remains frozen except for 1×1 convolutional layers added before each concatenation.

During training, we mark the pixels that exceed the dynamic range (which will has value of 0 or reach the maximum in a SDR image) as "unknown" regions (See Figure 2 (c)). To ensure that the detail generation is focused to the unknown regions while being semantically and aesthetically harmonious with the known regions, at the t-th step, the known regions are preserved by directly diffusing the initial latent feature, while the unknown regions are inferred from the model's denoised output [33]:

$$x_{t-1}^{\text{known}} \sim \mathcal{N}\big(\sqrt{\overline{\alpha}_t}\, x_0,\, (1-\overline{\alpha}_t)I\big), \tag{2}$$

$$x_{t-1}^{\text{unknown}} \sim \mathcal{N}\big(\mu_\theta(x_t, t),\, \Sigma_\theta(x_t, t)\big), \tag{3}$$

$$x_{t-1} = m_{\text{latent}} \odot x_{t-1}^{\text{known}} \;+\; (1 - m_{\text{latent}}) \odot x_{t-1}^{\text{unknown}}, \tag{4}$$

where the known pixels x_{t-1}^{known} are directly diffused from x_0, and the unknown pixels x_{t-1}^{unknown} are sampled from the diffusion model, blended by a mask m_{latent}. After t steps, the final denoised latent is decoded by the LDM decoder.

To guide Sagiri toward realistic content generation, we introduce the following loss:

$$L_{\text{content}} = \lambda_4 L_{\text{mse}} + \lambda_5 L_{\text{ssim}} + \lambda_6 L_{\text{fdp}}, \tag{5}$$

where L_{mse}, L_{ssim}, and L_{fdp} ensure structural fidelity, realistic textures, and frequency consistency. Coefficients $\lambda_4, \lambda_5, \lambda_6$ balance these terms. *Details are provided in the supplementary material.*

3.3 Training Strategy

Overall Pipeline. We train Latent-SwinIR$_c$ on the HDR-Real dataset [4] to learn appropriate color and brightness mappings. For Sagiri, we first pretrain it on the large-scale Places365 dataset [34] to expand its capability of generating diverse scenes, and then finetune it on HDR-Real.

Degradation Generation in Pre-training. To bridge the domain gap, we simulate Latent-SwinIR$_c$ outputs during Sagiri's pre-training on Places365 by crafting various realistic degradation patterns. Specifically, we generate a "degradation mask" using random lines of varied thickness, followed by dilation and Gaussian blurring. This mask is used to blend the original image with a heavily blurred version, producing blur-like artifacts mimicking over- or underexposed areas. This approach improves Sagiri's ability to serve as a *plug-and-play* module not only for Latent-SwinIR$_c$ but also for other methods we compare with (Figure 4).

Unknown Region Mask. During Sagiri's pre-training, we do not apply the unknown region mask; the model is encouraged to autonomously identify and handle low-quality areas. During finetuning, however, we apply the binary mask to emphasize the unique challenge of inpainting over-/under-exposed regions.

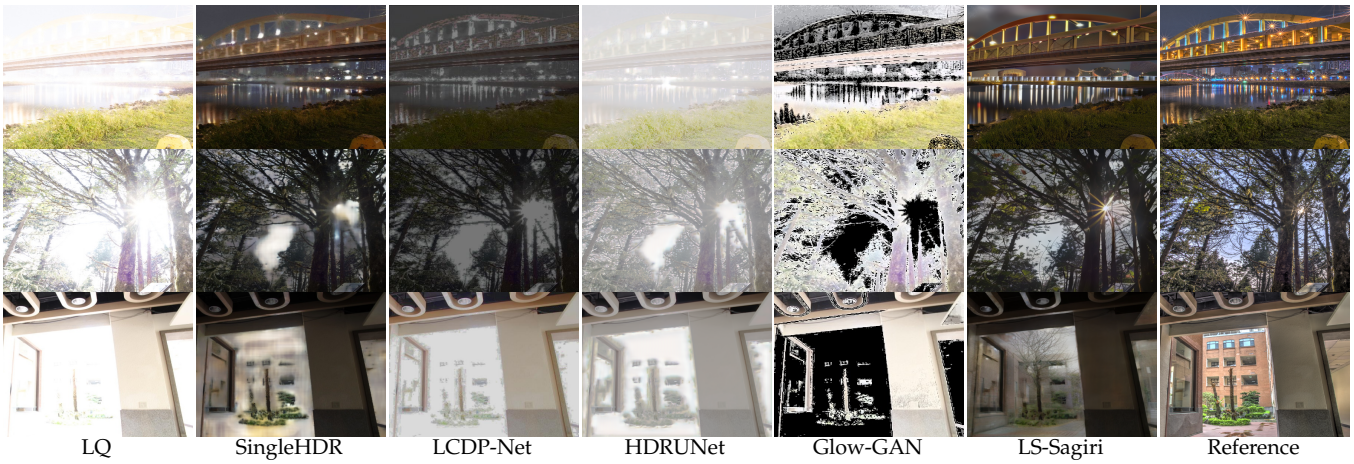


Fig. 3: **Performance of LS-Sagiri.** Previous restoration-based methods can only restore exposure-limited areas to blurry content. Although Glow-GAN [8] is a generative method, it fails to handle *large* overexposed regions, often rendering them black. In contrast, our method can generate realistic content based on existing information and specified mask areas.

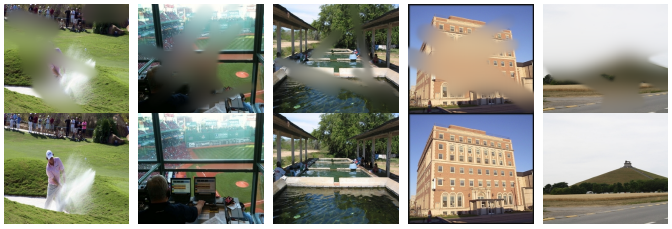


Fig. 4: The first row is the result obtained using our degradation strategy, while the second row is the reference images. We aim to simulate the degradation caused by other models in dynamic range extremes during SDR enhancement and train Sagiri to handle these situations effectively.

4 Experiments

4.1 Training and Inference Settings

Training. We train Latent-SwinIR$_c$ on the HDR-Real training set [4] for 150,000 iterations using a batch size of 16. Our Sagiri model is initialized from pretrained Stable Diffusion v2.1. We first pretrain Sagiri on 250,000 randomly selected images from the Places365 dataset [34] for 70,000 steps, then fine-tune it on the HDR-Real training set for an additional 20,000 steps. All training stages use the Adam optimizer [35] with a learning rate of 1×10^{-4}, conducted on four NVIDIA A100 GPUs. The loss coefficients λ_1, λ_2, λ_3, λ_4, λ_5, and λ_6 are set to 10, 1, 0.1, 1, 1, and 0.01, respectively. The total training process takes 3 days using 4 NVIDIA A100 GPUs.

Inference. During inference, our model processes an SDR input image, applying an unknown region mask that detects pixels with values of 0 or 255. We use just 30 steps of DDPM sampling [36]. Note that all the visualizations are from the test set or real-world examples.

Datasets. We evaluate performance on HDR-Real [4], NTIRE [37], HDR-Eye [38], Eye-over, and Eye-under. The latter two are constructed by uniformly adjusting the exposure values of HDR-Eye to create overexposed or underexposed images. Since existing datasets lack a large number of test images with severe content loss, these newly created sets better assess our method's effectiveness in extreme conditions.

Prompt Usage in Training and Inference. We employ CogVLM [39] to summarize input images into prompts. During Sagiri's fine-tuning on HDR-Real, these prompts are generated using the ground-truth images to align the model with prompt inputs. For inference on the HDR-Real testing set [4], prompts are produced from low-quality inputs. For the HDR-Eye [38], Eye-over, Eye-under, and NTIRE [37] datasets, we do not provide prompts. This design aims to evaluate Sagiri's adaptability under different scenarios.

4.2 Results

Baseline Methods. We compare our approach with SingleHDR [4], LCDPNet [1], HDRUNet [5], and GlowGAN [8]. For fairness, we uniformly apply the Reinhard tone mapping function [40] across the training and testing sets to generated pairs of SDR images for training and evaluating SDR methods (LCDPNet and ours), and as a common post-processing step for evaluating HDR methods (SingleHDR, HDRUNet and GlowGAN). Additional comparisons with GDP [31] are provided in the supplementary material.

Metrics. To assess both our complete LS-Sagiri pipeline and the generalizability of Sagiri as a refinement module for other models, we rely on no-reference metrics such as BRISQUE [41], NIQE [42], MANIQA [43], and CLIP-IQA [44], which focus on visual quality. We do not use PSNR, SSIM [45], or LPIPS [46] because prior studies [37], [47], [48] have shown that these metrics can be unreliable for generative models and this challenging generative task.

Performance of Latent-SwinIR$_c$. Figures 7(a–f) demonstrate the ability of Latent-SwinIR$_c$ to correct image color distributions. Rows 1–3 show LQ images captured in extremely dark conditions, a challenging scenario for image enhancement. SingleHDR reveals more details yet suffers from low contrast. LCDP-Net and HDRUNet fail to adequately adjust the brightness, leaving the result underexposed. GlowGAN struggles to restore quantized details, leading to images lacking essential textures. Thanks to our novel loss formulation, Latent-SwinIR$_c$ achieves superior brightness and color mapping without compromising detail, delivering high-quality, natural-looking images.

TABLE 1: Quantitative results on HDR-Real [4], NTIRE [37], HDR-Eye [38], Eye-over and Eye-under datasets. The latter two datasets are made by uniformly adjusting the exposure value of HDR-Eye dataset to synthesize datasets with large areas at dynamic range extremes. In addition to comparing the performance of our pipeline with existing methods, we plugged Sagiri into each model to see performance improvements. The results show that (1) Sagiri enhances the performance of each method, and (2) LS-Sagiri achieves the best overall results.

Datasets	HDR-Real				NTIRE			
Metrics	BRISQUE↓	NIQE↓	MANIQA↑	CLIP-IQA↑	BRISQUE↓	NIQE↓	MANIQA↑	CLIP-IQA↑
SingleHDR [4]	23.597	20.839	0.367	0.387	22.730	**21.399**	0.250	0.411
SingleHDR+Sagiri	**19.855**	**20.326**	**0.556**	**0.649**	**10.211**	21.622	**0.385**	**0.676**
LCDPNet [1]	30.704	20.660	0.344	0.383	19.237	**20.978**	0.267	0.415
LCDPNet+Sagiri	**24.464**	**20.318**	**0.542**	**0.641**	**9.951**	21.622	**0.385**	**0.674**
HDRUNet [5]	41.521	21.388	0.341	0.361	52.898	22.752	0.229	0.377
HDRUNet+Sagiri	**24.935**	**20.704**	**0.503**	**0.609**	**21.353**	**21.749**	**0.397**	**0.650**
GlowGAN [8]	36.727	21.774	**0.470**	0.503	21.769	**24.053**	**0.403**	0.478
GlowGAN+Sagiri	**22.840**	**21.602**	0.443	**0.554**	**15.549**	24.078	0.354	**0.511**
Latent-SwinIR$_c$	35.407	21.457	0.291	0.303	31.298	22.000	0.224	0.392
LS-Sagiri	**19.725**	**20.309**	**0.569**	**0.670**	**9.724**	**21.652**	**0.395**	**0.671**

Datasets	HDR-Eye			Eye-over			Eye-under		
Metrics	BRISQUE↓	MANIQA↑	CLIP-IQA↑	BRISQUE↓	MANIQA↑	CLIP-IQA↑	BRISQUE↓	MANIQA↑	CLIP-IQA↑
SingleHDR [4]	18.338	0.452	0.466	20.573	0.447	0.428	33.675	0.244	0.244
SingleHDR+Sagiri	**15.092**	**0.570**	**0.697**	**14.969**	**0.557**	**0.676**	**13.477**	**0.339**	**0.523**
LCDPNet [1]	20.672	0.453	0.475	26.374	0.398	0.365	54.493	0.311	0.335
LCDPNet+Sagiri	**14.137**	**0.543**	**0.665**	**14.973**	**0.478**	**0.638**	**37.825**	**0.382**	**0.552**
HDRUNet [5]	27.672	0.418	0.390	24.545	0.454	0.410	72.920	0.364	0.403
HDRUNet+Sagiri	**14.846**	**0.555**	**0.662**	**15.905**	**0.560**	**0.668**	**40.954**	**0.460**	**0.610**
GlowGAN [8]	**16.042**	**0.506**	**0.536**	**16.930**	**0.503**	**0.561**	46.667	**0.356**	**0.483**
GlowGAN+Sagiri	19.775	0.430	0.473	20.040	0.401	0.466	**37.745**	0.286	0.432
Latent-SwinIR$_c$	25.870	0.329	0.286	25.345	0.321	0.286	45.168	0.256	0.252
LS-Sagiri	**14.777**	**0.538**	**0.675**	**14.667**	**0.535**	**0.669**	**12.066**	**0.462**	**0.660**

In Rows 4–8, five images with decreasing exposure times are tested. As the exposure decreases, existing methods exhibit fading contrast (SingleHDR), dull brightness (LCDP-Net and HDRUNet), or further detail loss (GlowGAN). In contrast, Latent-SwinIR$_c$ consistently preserves robust color and brightness distributions, demonstrating remarkable robustness to varying exposure levels.

Performance of LS-Sagiri. Figures 7(g) and 3 present the performance of the complete LS-Sagiri pipeline. Despite the overall robust performance of Latent-SwinIR$_c$, it is notable that the results still contain certain degradations, such as blurry areas in exposure-limited regions, as illustrated in Figure 7. In Figure 7(g), Sagiri clearly refines Latent-SwinIR$_c$'s outputs by synthesizing realistic details, notably improving the perceived image quality. Figure 3 features inputs with large overexposed regions, where SingleHDR introduces blurry areas, LCDP-Net and HDRUNet fail to regulate brightness, and GlowGAN cannot recover heavily quantized details. Only LS-Sagiri succeeds in filling oversaturated regions with coherent, realistic details. Table 1 shows that LS-Sagiri achieves the top performance across nearly all metrics, confirming its efficacy in enhancing LDR images and its broad applicability to various datasets.

We note that Latent-SwinIR$_c$ alone does not always yield top metric scores. This may stem from the limitations of no-reference image quality metrics, which do not thoroughly account for global brightness distribution. Future research is needed to explore more comprehensive metrics to better evaluate such enhancements.

Sagiri as a Plug-and-Play Module. Beyond refining Latent-SwinIR$_c$ outputs, Sagiri can seamlessly integrate with other models. Figure 8 shows how Sagiri corrects dynamic range extremes even when the initial enhanced images vary significantly in quality, ultimately improving the overall perceptual quality. Table 1 supports this versatility: Sagiri significantly boosts nearly every baseline's output. Minor exceptions (e.g., GlowGAN) are detailed in the supplementary material.

Further evaluations. A user study on the subjective quality of the methods and a comparison with inpainting methods can be found in the supplementary material.

4.3 Ablation Studies

Importance of the Two-Stage Model. To verify whether our two-stage pipeline is essential, we tested Sagiri alone for both color restoration and detail enhancement. As shown in Figure 5 (Top), Sagiri alone struggles with color restoration and brightness adjustment, demonstrating the necessity of Latent-SwinIR$_c$ as the first stage.

Effect of Pre-training and Prompts. Figure 5 (Middle) shows that our pre-training approach and prompt-guided generation significantly improve visual quality, while the content reconstruction loss (ConRLoss) further enhances structural integrity. In addition, Figure 5 (Bottom) illustrates how users can control (1) the region to be generated by

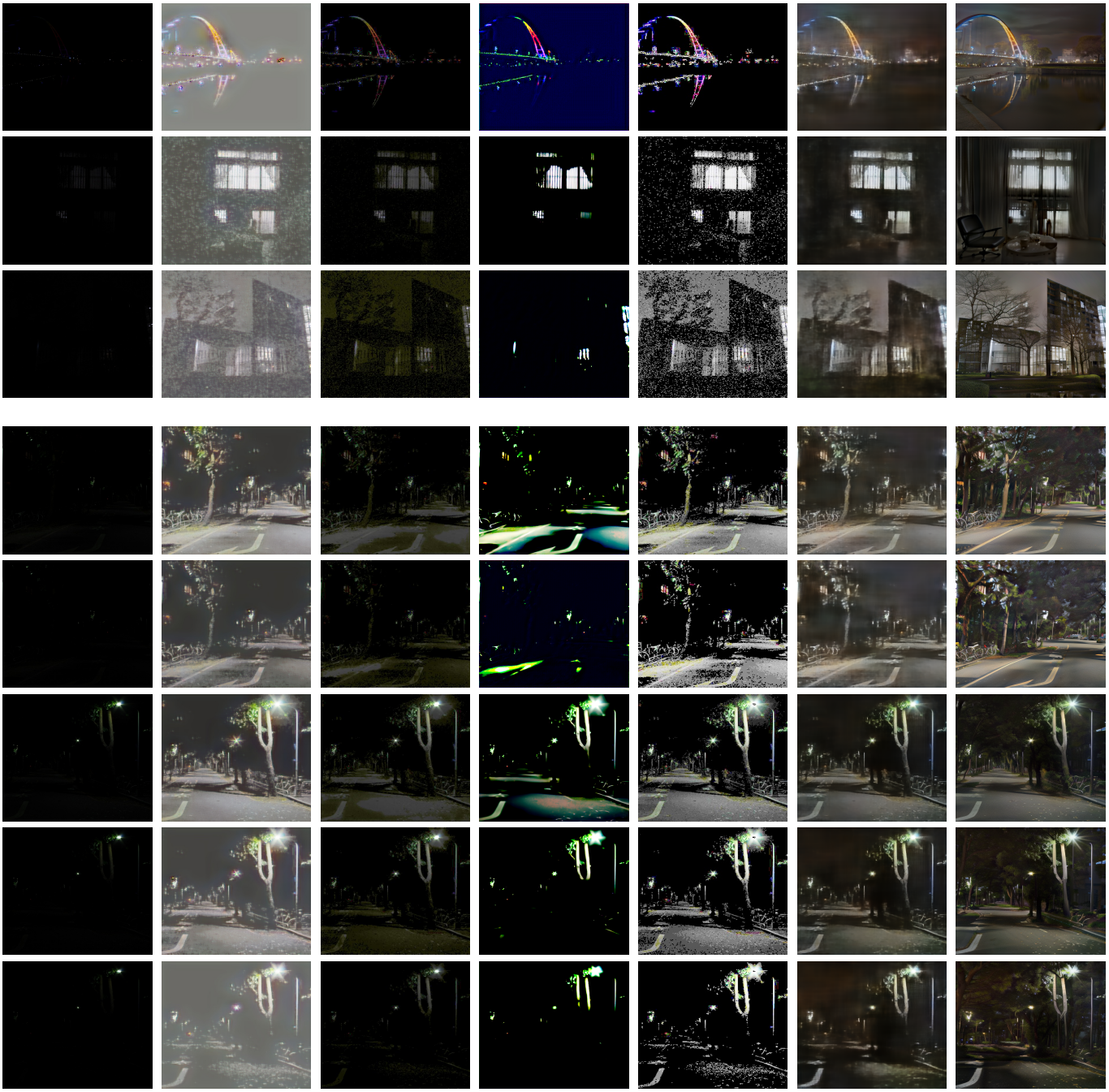
Fig. 5: **Ablation studies. (Top)** We enforce Sagiri to learn both color distribution correction and details generation, which leads to weak color mapping capabilities. **(Middle)** Ablation of pretraining, text prompt and content reconstruction loss. Prompts generated by CogVLM [39]: "A white waterfall is flowing down from the cliff, surrounded by rocks and trees." **(Bottom)** We use different user-defined unknown region mask and different prompts on Sagiri to refine SingleHDR's [4] results. Left: We manually select the red box. Right: We select the entire image. Prompt a: "The sky is filled with clouds." Prompt b: "The sun is setting, and the sky is filled with clouds." Please zoom in to see more details.

Fig. 6: Use different prompts to control the generated results. Prompt a: "A building with a **red** brick exterior, white columns, and a **black** door..." Prompt b: "A building with a **black** brick exterior, white columns, and a **red** door...". Furthermore, The model has poor responsiveness to prompts that do not fit the current context, as we found. Prompt c: "The **sun** is setting in the forest, and the trees are **black**." Prompt d: "The **moon** is setting in the forest, and the trees are **green**". Please zoom in to see more details.

customizing the unknown-region mask, and (2) the content to be generated by supplying a user-defined text prompt. More results featuring prompt-guided generation in extreme regions are shown in Figure 6.

5 Conclusion

We propose a novel pipeline for exposure-limited image enhancement, anchored by our robust and flexible Sagiri model. The pipeline comprises two stages: Stage 1 (Latent-SwinIR$_c$, LS) rectifies brightness and color distributions, while Stage 2 (Sagiri) synthesizes content in missing or severely degraded regions and refines overall details. By design, Sagiri is also compatible as a plug-and-play module, allowing it to enhance outputs from diverse restoration methods. Our experiments confirm the superior performance of the two-stage LS-Sagiri framework and demonstrate Sagiri's remarkable ability to generate realistic details. **Besides, please note that given the challenging task, which is open end, and the generative character of stable diffusion based method, our approach is mainly designed for beautify the given input, not to restrictly following the reference image.** Currently, Sagiri produces only SDR images due to limitations of the Stable Diffusion on which it is built. A promising future direction is to extend Sagiri to generate HDR outputs, thus giving users greater flexibility in applying customized tone mapping.



(a) LQ (b) SingleHDR (c) LCDP-Net (d) HDRUNet (e) Glow-GAN (f) Latent-SwinIR$_c$ (g) LS-Sagiri

Fig. 7: **(a-f) Performance of Latent-SwinIR**c (LS). Existing methods often struggle to recover content in regions with extreme dynamic range. In contrast, Latent-SwinIRc, thanks to its uniquely designed loss function, captures a more balanced color distribution. This is evident in the Row 1–3. Moreover, in Row 4–8, where the input low-quality (LQ) images in column **(a)** exhibit a gradual decrease in exposure, columns **(b-f)** show that the performance of existing methods deteriorates with decreasing exposure. Although SingleHDR provides results closest to our method, it still produces low-contrast "hazy" outputs at low exposures. In **(f)**, LS demonstrates robust preservation of color and brightness despite the decreasing exposure levels. Additionally, **(g)** highlights the Sagiri model's excellence in generating detailed content across large regions, thereby enhancing the overall quality. **Please Zoom in the figures for details.**

Fig. 8: **Sagiri as a plug-and-play module.** Although the images generated by the baselines significantly differ from each other, Sagiri shows strong versatility and improves the visual quality of almost all of them. Additionally, the combination of LS-Sagiri surpasses the performance of other models integrated with Sagiri, confirming the robustness and adaptability of our framework.

REFERENCES

[1] H. Wang, K. Xu, and R. W. Lau, "Local color distributions prior for image enhancement," in *ECCV*, 2022.

[2] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[3] Z. Liu, Y. Wang, B. Zeng, and S. Liu, "Ghost-free high dynamic range imaging with context-aware transformer," in *ECCV*, 2022.

[4] Y.-L. Liu, W.-S. Lai, Y.-S. Chen, Y.-L. Kao, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, "Single-image hdr reconstruction by learning to reverse the camera pipeline," in *CVPR*, 2020.

[5] X. Chen, Y. Liu, Z. Zhang, Y. Qiao, and C. Dong, "Hdrunet: Single image hdr reconstruction with denoising and dequantization," in *CVPR*, 2021.

[6] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *SIGGRAPH*, 1997.

[7] Y. Zou, C. Yan, and Y. Fu, "Rawhdr: High dynamic range image reconstruction from a single raw image," in *ICCV*, 2023.

[8] C. Wang, A. Serrano, X. Pan, B. Chen, K. Myszkowski, H.-P. Seidel, C. Theobalt, and T. Leimkühler, "Glowgan: Unsupervised learning of hdr images from ldr images in the wild," in *ICCV*, 2023.

[9] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[10] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[12] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*, 2021.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[14] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *ICCV*, 2021.

[15] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023.

[16] Y. Li, K. Xu, G. P. Hancke, and R. W. Lau, "Color shift estimation-and-correction for image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 389–25 398.

[17] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *IEEE transactions on image processing*, vol. 6, no. 3, pp. 451–462, 1997.

[18] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

[19] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," *arXiv preprint arXiv:1808.04560*, 2018.

[20] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1780–1789.

[21] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5637–5646.

[22] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Snr-aware low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 714–17 724.

[23] B. Li, H. Zheng, Z. Zhang, Y. Zhao, Z. Zhao, and H. Zhang, "Dynamic grouped interaction network for low-light stereo image enhancement," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2468–2476.

[24] G. Kim, D. Kwon, and J. Kwon, "Low-lightgan: Low-light enhancement via advanced generative adversarial network with task-driven training," in *2019 IEEE International conference on image processing (ICIP)*. IEEE, 2019, pp. 2811–2815.

[25] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 302–12 311.

[26] P.-H. Le, Q. Le, R. Nguyen, and B.-S. Hua, "Single-image hdr reconstruction by multi-exposure generation," in *WACV*, 2023.

[27] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.

[28] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," *ACM Transactions on Graphics (Proc. of SIGGRAPH ASIA 2017)*, vol. 36, no. 6, Nov. 2017.

[29] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "Hdr image reconstruction from a single exposure using deep cnns," *ACM transactions on graphics (TOG)*, vol. 36, no. 6, pp. 1–15, 2017.

[30] M. S. Santos, R. Tsang, and N. Khademi Kalantari, "Single image hdr reconstruction using a cnn with masked features and perceptual loss," *ACM Transactions on Graphics*, vol. 39, no. 4, 7 2020.

[31] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang, and B. Dai, "Generative diffusion prior for unified image restoration and enhancement," in *CVPR*, 2023.

[32] D. P. Kingma, M. Welling *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, 2019.

[33] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *CVPR*, 2022.

[34] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE TPAMI*, 2017.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *ICML*, 2021.

[37] J. Gu, H. Cai, C. Dong, J. S. Ren, R. Timofte, Y. Gong, S. Lao, S. Shi, J. Wang, S. Yang *et al.*, "Ntire 2022 challenge on perceptual image quality assessment," in *CVPR*, 2022.

[38] H. Nemoto, P. Korshunov, P. Hanhart, and T. Ebrahimi, "Visual attention in ldr and hdr images," in *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2015.

[39] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song *et al.*, "Cogvlm: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.

[40] E. Reinhard and K. Devlin, "Dynamic range reduction inspired by photoreceptor physiology," *IEEE transactions on visualization and computer graphics*, vol. 11, no. 1, pp. 13–24, 2005.

[41] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE TIP*, 2012.

[42] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, 2012.

[43] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *CVPR*, 2022.

[44] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *AAAI*, 2023.

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, 2004.

[46] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[47] G. Jinjin, C. Haoming, C. Haoyu, Y. Xiaoxing, J. S. Ren, and D. Chao, "Pipal: a large-scale image quality assessment dataset for perceptual image restoration," in *ECCV*, 2020.

[48] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *CVPR*, 2018.