HDR Image Generation by consistent LDR denoising

Rupika Nilakant, EE367, Stanford University

Abstract—This poster proposes a novel method for generating high-dynamic range (HDR) images using pre-trained, black-box low-dynamic range (LDR) denoising diffusion models. The motivation is to overcome limitations in HDR image synthesis by leveraging multiple LDR brackets and ensuring consistency between them during the diffusion process. This allows the generation of HDR images without requiring large-scale HDR datasets or expensive model retraining. The key idea is to operate multiple denoising processes to generate multiple LDR brackets that together form a valid HDR result. To this end, the method introduces an exposure consistency term into the diffusion process to couple the brackets such that they agree across the exposure range they share. The brackets are then fused to produce a single HDR image using Debevec's method followed by tonemapping.

Index Terms—Diffusion Models, Computational Imaging

1 Introduction

Modern denoising diffusion models used for image generation typically generate them as Low Dynamic Range (LDR) images whereas High Dynamic Range (HDR) images are required in several applications such as advanced displays or scene reconstruction involving profound shadows and specular highlights. Common diffusion models are not HDR as there is no sufficiently large HDR image dataset available to re-train them, and, second, even if it was, re-training such models is impossible for most compute budgets.

The goal is to produce a set of individual meaningful exposure ”brackets”, i.e., LDR images, which can be merged into an HDR image. This method does not need any fine-tuning or training and considers the denoiser a black box. A ”bracket” should have all details, without noise, in the range of values it represents. To work as a combination, a value in one bracket must match its value re-exposed to another bracket and ultimately when they are merged. This is done by modifying the diffusion process based on Diffusion Posterior Sampling (DPS) [1] that operates between multiple brackets jointly by adding a consistency term in the reverse diffusion process.

2 Related Work

HDR images directly register scene radiance, typically up to a scale factor, so that image details in the darkest and brightest scene regions are visible. As sensors with HDR capabilities are relatively rare and expensive, a stack of differently exposed LDR photographs is typically merged into an HDR image [2]. An alternative solution to multi-exposure techniques is to restore HDR information from a single LDR image using deep learning techniques. Single-image HDR reconstruction can be performed directly [3], or, alternatively, by first producing a stack of different exposures that are then merged into an HDR image [4][5][6]. Even though some methods employ adversarial training [7], the key problem remains limited performance in reconstructing clamped regions. Those methods mostly require LDR and HDR image pairs for training, which is problematic due to limited datasets. Recently, GlowGAN [8] addressed the latter two problems by fully unsupervised learning a generative model of HDR images exclusively from in-the-wild LDR images. As this approach is based on StyleGAN-XL [9], it re- quires GAN training on narrow domains (e.g., lightning, fireworks) to capture the respective HDR image distribution.

2.1 Diffusion Models in HDR imaging

Denoising diffusion probabilistic models (DDPMs) [11] demonstrate huge capacity in modeling complex distributions and typically outperform other generative models in terms of image realism, diversity, and detail reproduction. DDPMs also proved useful for solving linear and non-linear [1] inverse imaging problems that are common in image restoration and enhancement tasks guided by the degraded input image.

In HDR imaging tasks, the degradation model is more complex, and existing solutions based on DDPMs are more sparse. Wang et al. [12] propose low-light image enhancement using exposure diffusion that is initialized with the noisy low-light image instead of Gaussian noise, which g simplifies denoising and consequently reduces the network complexity and the required number of inference steps. The method can be trained using pairs of low-light and normally-exposed photographs, as well as synthetic data using different noise models. Fei et al. [13] employ a pre-trained DDPM and propose the Generative Diffusion Prior (GDP) for unsupervised modeling of the natural image posterior distribution. They demonstrate the utility of this framework for low-light image enhancement and HDR image reconstruction by merging low, medium, and high exposures. A similar task, but with explicit emphasis on large motion between the three exposures and severe clamping at the same time, is addressed in Yan et al.[14]

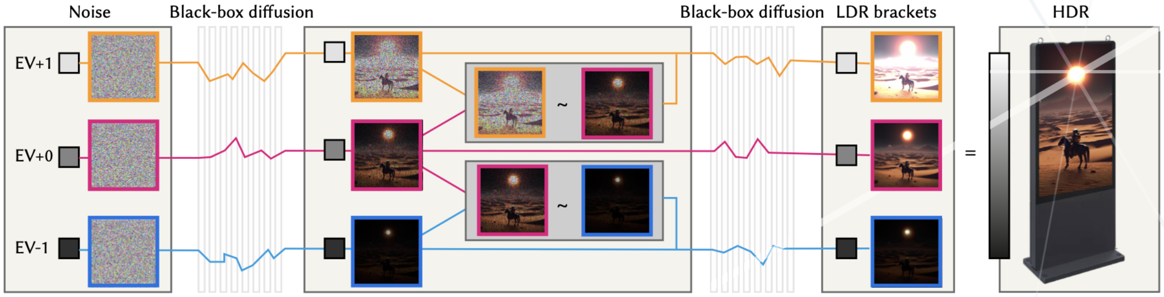

Fig. 1. Diffusion occurs from left to right and across multiple exposure levels (brackets), shown vertically.The process starts with three independent noises. At each diffusion step (one is shown), denoising is guided by an exposure consistency term (middle block). In this term, brackets are made consistent when re-exposed. When diffusion has finished, the brackets form an HDR image under a common HDR fusion.

where train a DDPM to capture the distribution of natural HDR environment maps, but are limited to rather narrow classes (e.g., urban streets) due to scarcity of available HDR training data.

3 Proposed Method

This project follows [1] and relies on off-the-shelf pre-trained diffusion models that feature better domain generalizability due to intensive training on large datasets than explicit training on small datasets of LDR–HDR image pairs. The proposed method does not require any HDR images at the training stage. Instead, we implicitly leverage the exposure statistics of real-world photographs used for DDPM training, which allows the model to reason on the underlying radiance distributions. In single-image reconstruction, we require as the input just one LDR exposure and then generate a stack of different spatially consistent LDR exposures.

3.1 Guided Diffusion

Data generation with a pre-trained DDPM [11] amounts to gradual denoising of a sample $\mathbf{x} \in \mathbb{R}^u$ using

$$\mathbf{x}_{t-1} := \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - (1-\alpha_t)\nabla_{\mathbf{x}_t}\log p_t\left(\mathbf{x}_t\right)\right) + \mathbf{z}_t \text{ (1)}$$

This update rule involves a noise schedule $\alpha_\mathbf{t} \in \mathbb{R}_+$, random vectors $\mathbf{z_t} \in \mathbb{R}^u$, and, a score function, $\nabla_{\mathbf{x}_t}\log p_t\left(\mathbf{x}_t\right)$. In the framework of diffusion posterior sampling (DPS) [1], an additional guiding signal $\mathbf{y} \in \mathbb{R}^w$ such as a partial observation of $\mathbf{x}$, is incorporated into the denoising process to arrive at the posterior score.

$$\nabla_{\mathbf{x}_t}\log p_t\left(\mathbf{x}_t \mid \mathbf{c}, \mathbf{y}\right) \approx \mathbf{s}_\theta\left(\mathbf{x}_t, \mathbf{c}, t\right) - \lambda\nabla_{\mathbf{x}_t}C\left(\hat{\mathbf{x}}_t, \mathbf{y}\right) \text{ (2)}$$

Here, $C \in \left(\mathbb{R}^u \times \mathbb{R}^w\right) \to \mathbb{R}$ is a problem-specific consistency measurement term that drives the denoising process towards solutions that incorporate the guiding signal $\mathbf{y}$, and $\lambda \in \mathbb{R}_+$ is a balancing term. For increased stability, [1] propose to feed the current estimate of the clean sample,

$$\hat{\mathbf{x}}_t = \frac{1}{\sqrt{\overline{\alpha}_t}}\left(\mathbf{x}_t + (1-\overline{\alpha_t})\, s_\theta\left(\mathbf{x}_t, \mathbf{c}, t\right)\right) \text{ (3)}$$

to C, where $\overline{\alpha_t}$ is derived from α_t.

3.2 Exposure diffusion

The above equations Eq. 1 and Eq. 2 are valid for producing a single LDR result image $\mathbf{x}$. The key idea is to produce HDR by fusing multiple generated LDR results. Hence, we operate on a set of LDR images, $\{\mathrm{x}^{-m}, \ldots, \mathrm{x}^{0}, \ldots, \mathrm{x}^{n}\}$, called "brackets". Positive and negative superscripts denotes positive and negative EVs, respectively. All brackets are initialized to noise with mean zero and standard deviation one. They, further, need to be gamma-corrected sRGB LDR images, as we consider the score function a black box that cannot be retrained to work on linear HDR.

Score term The first term in Eq. 2 is the common score function that points from the current solution into the direction of a more plausible one. It may or may not be conditioned as per the second column of Tab. 1, leading to different application scenarios. It is a black box we do not need to know any details of, nor differentiate, as it already encodes a gradient. We only need to know its noise schedule α_t to also use $\hat{\mathrm{x}}$ from Eq. 3. The score function is hence simply computed on each bracket independently. The scoring function in our implementation is a pretrained diffusion model trained on a specific dataset.

Posterior term The second term in Eq. 2 is very specific to this problem, the exposure consistency cost term. The consistency of two brackets measures how much $\hat{\mathbf{x}}^i$, a free variable, is compatible with another bracket $\hat{\mathbf{x}}^r$ that is assumed fixed. For each bracket $\hat{\mathbf{x}}^i$, the reference bracket $\hat{\mathbf{x}}^r$ is exposed to another bracket, and the resulting differences are checked using a function $exco$.

$$C\left(\hat{\mathbf{x}}^i, \mathbf{y}\right) = \begin{cases} C_\downarrow\left(\hat{\mathbf{x}}^i \hat{\mathbf{x}}^{i+1}\right) & \text{, if } i < 0 \text{, see Eq. 5,} \\ C_\uparrow\left(\hat{\mathbf{x}}^i, \hat{\mathbf{x}}^{i-1}\right) & \text{, if } i > 0 \text{, see Eq. 6 and} \\ C_0\left(\hat{\mathbf{x}}^i, \mathbf{y}\right) & \text{, if } i = 0 \text{, see Eq. 7.} \end{cases}$$

Both positive and negative posterior make use of two mask functions sat and dark which are one for saturated and near-zero pixels, respectively, and zero otherwise. In practice, we use smooth versions of that for better differentiability; a very smooth function sat(x)= x and

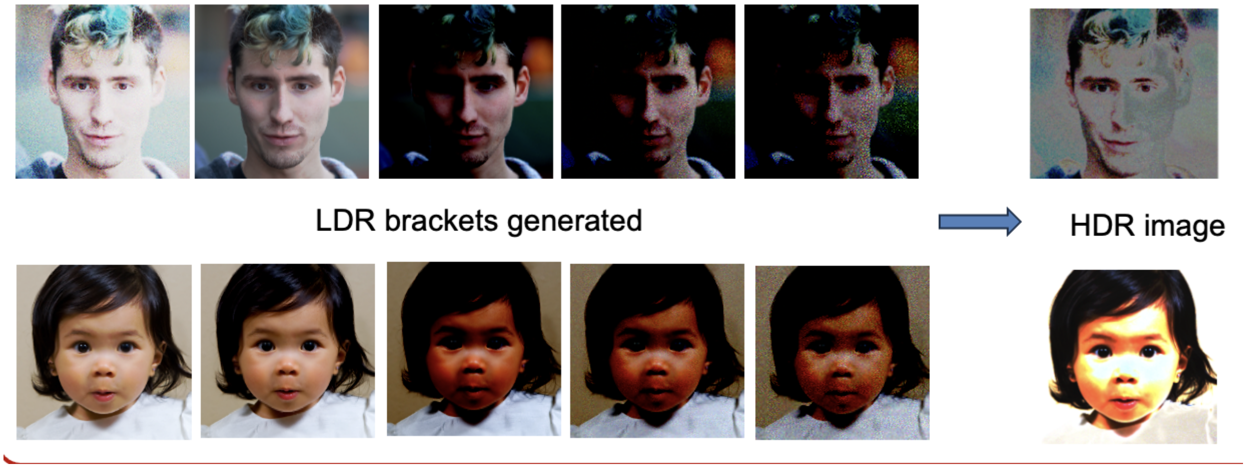

Fig. 2. The LDR brackets are generated with the guiding term given as a sample image from the Diffusion project. 5 ”brackets” are created for each image and they are fused using Debevec’s technique and Drago tonemapping

dark(x)= 1 - x provides the best results.

The posterior for decreasing exposure is

$$C_{\downarrow}\left(\hat{\mathbf{x}}^i, \hat{\mathbf{x}}^r\right) = \left\|\operatorname{sat}\left(\hat{\mathbf{x}}^r\right) \cdot \max\left(\operatorname{exco}\left(\hat{\mathbf{x}}^r \rightarrow \hat{\mathbf{x}}^i\right), 0\right)\right\|_2 + \lambda_{\mathrm{s}} \cdot \left\|\left(1 - \operatorname{sat}\left(\hat{\mathbf{x}}^r\right)\right) \cdot \left(\operatorname{exco}\left(\hat{\mathbf{x}}^r \rightarrow \hat{\mathbf{x}}^i\right)\right)\right\|_2, \quad (5)$$

while the one to increase exposure is

$$C_{\uparrow}\left(\hat{\mathbf{x}}^i, \hat{\mathbf{x}}^r\right) = \left\|\operatorname{dark}\left(\hat{\mathbf{x}}^r\right) \cdot \left(\operatorname{exco}\left(\hat{\mathbf{x}}^r \rightarrow \hat{\mathbf{x}}^i\right)\right)\right\|_2 + \lambda_{\mathrm{d}} \cdot \left\|\left(1 - \operatorname{dark}\left(\hat{\mathbf{x}}^r\right)\right) \cdot \left(\operatorname{exco}\left(\hat{\mathbf{x}}^r \rightarrow \hat{\mathbf{x}}^i\right)\right)\right\|_2, \quad (6)$$

where λ_s and λ_d are set to 1 and 2, respectively. Note that in Eq. 6, the two terms are weighted differently. This is because the darker regions dark $\hat{\mathbf{x}}^r$ are usually noisy or unreliable; thus, we impose less exposure consistency prior on these regions compared to the brighter regions.

The consistency of exposure exco of one LDR bracket $\hat{\mathbf{x}}^i$ with respect to a reference $\hat{\mathbf{x}}^r$ (which can both be higher or lower EV) is defined as

$$\operatorname{exco}\left(\hat{\mathbf{x}}^r \rightarrow \hat{\mathbf{x}}^i\right) := \left(\min\left(\left(\frac{\beta^i}{\beta^r} \odot \left(\hat{\mathbf{x}}^r\right)^{-\gamma}\right), 1\right)\right)^{\gamma} - \hat{\mathbf{x}}^i$$

where β stands for exposure time. We first undo the gamma (γ= 2.2), as the solution has to live in non-linear space for the black box score. Next, we scale by the ratio between the exposure times and then clamp and apply gamma again, as a real camera would. The result has returned to the domain an LDR score function can handle and is compared to the bracket $\hat{\mathbf{x}}^i$ in question.

Finally, we can also define an optional posterior term on the original image by applying a function f:

$$C_0\left(\hat{\mathbf{x}}^i, \mathbf{y}\right) = \lambda_c \cdot \left\|f\left(\hat{\mathbf{x}}^i\right) - \mathbf{y}\right\|_2 \ (7)$$

In this implementation, this Eq 7’s guiding term y is the initial LDR image and we use the additional DPS step of subtracting the gradient of the likelihood prior.

4 Implementation and Results

The baseline code used for this implementation is the DDPM and DPS task from the Diffusion Project. the LDR to HDR fusion is implemented using Debevec’s technique seen in the EE367 class and homework.

For this implementation, we use the pretrained diffusion model trained on the FFHQ dataset directly given in the Diffusion Project. The input images are downsampled to 256 x 256, and I have implemented 1500 denoising steps to produce the results.

In the original implementation of the paper: ”Exposure Diffusion: HDR Image Generation by Consistent LDR denoising”, they use text-based or histogram-based conditioning for initial image generation. In my implementation, due to the limited time, I’ve used two LDR images from the sample data provided in the diffusion project. Eq 1 through 7 shown above are implemented in python The hyper-parameterλ in Eq. 1 balances between the diffusion prior and our posterior term. This is the $lambda_{dps}$ that’s a tuning parameter in the code. It’s set to 3 for the current implementation.

For the two qualitative results (Fig 2) , we compute five exposure brackets: EV-4, EV-2, EV+0, EV+2, and EV+4, unless otherwise specified. These exposure brackets are merged using the standard technique [Debevec and Malik 1997] to create our HDR image. Drago tonemapping technique is used from openCV, and fused and tone mapped image for various scales and gammas are generated. the visually better one is picked for demonstrating the results here, but the code generates all the combinations (as seen in a previous homework).

Code The code to the exposure diffusion implementation as well as the LDR to HDR fusion can be found in the drive link here:

Click here to visit drive link

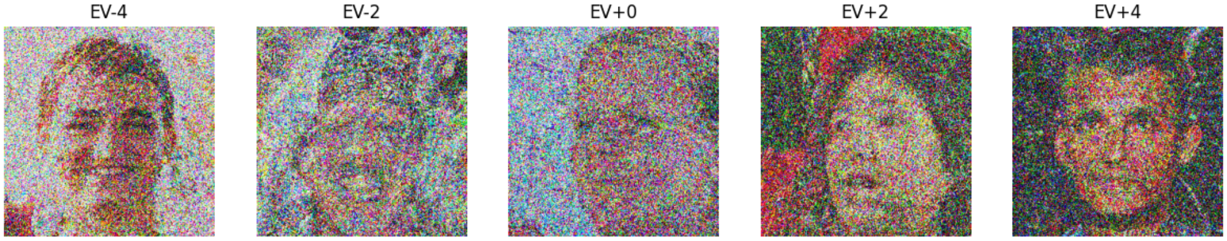

Fig. 3. Unconditional generation of exposure brackets without any guidance

5 Discussion and Conclusion

This project is able to provide a qualitative result and discussion. Since it is image generation using DPS, and since we dont have a ground truth HDR image to compare with, I'm not able to provide quantitative results.

One way to evaluate the model here is to take a good, non-saturated, no shadow image, change it's exposure ratio to make it saturated, run the saturated image through the DPS denoising steps, and fuse it to form HDR image. then compare the HDR image with the initial non-saturated, no shadow good image. This however has not been implemented in this case.

One issue that is noted is that without Eq 7's implementation, or without introducing the conditioning using the initial sample image, each exposure bracket is different from the other, which implies that the consistency term plays a significant role in ensuring that the same image is generated across the different brackets using the prior of the sample image passed to the model. An example of random generation is shown here, where all the brackets are quite noisy. Another concern is that even in the correctly generated exposure brackets, some low exposure brackets are quite noisy, and the tuning parameter for both the consistency term as well as number of sampling steps had to be varied to produce a reasonable set of brackets for HDR fusion.

Since the brackets themeselves are noisy, the fusion using Debevec's method did not yield visually good results for both the samples. This could be due to not enough tuning of the scale, saturation, bias and gamma, but could also be due to the generated brackets.

5.1 Limitations of proposed method

The proposed method uses a pretrained diffusion model trained only on face data (the FFHQ dataset). The consistency term does not take into account face-specific features when dealing with face data. The model is only trained on faces, so it lacks knowledge of general scenes with diverse textures, lighting, and objects.The exposure correction relies on a simple gamma-based transformation, which does not capture the diverse lighting effects seen in real-world scenes.Exposure consistency relies on L2 loss, which assumes small perturbations—but faces do not have extreme lighting variations like general HDR scenes.The posterior correction assumes that the noisy input follows a learned prior distribution, but faces have much less lighting variance than general images.You may find that $grad_{exco}$ is always None or zero, meaning the exposure correction has no effect on guiding the diffusion process.

5.2 Future Work

The exposure consistency term assumes a simplified gamma correction model, which does not accurately reflect real-world sensor noise, highlight clipping, or tone mapping behaviors.Replacing gamma-based exposure correction with a learned differentiable exposure transformation would be useful. Introducing adaptive weighting for the exposure consistency loss, so it ignores extreme outliers (e.g., saturated regions) may improve the quality of the generated brackets.

In future work, It would be interesting to extend the presented ideas to other modalities involving multiple images, like multi-spectral, stereo, light fields, and combinations thereof.

References

[1] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. 2023. Diffusion posterior sampling for general noisy inverse problems. (2023). arXiv:2209.14687 [cs.CV]

[2] Paul E. Debevec and Jitendra Malik. 1997. Recovering High Dynamic Range Radiance Maps from Photographs. In Proc. ACM SIGGRAPH. 369–378.

[3] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. 2022. Text2Light: Zero-Shot Text- Driven HDR Panorama Generation. ACM Trans. Graph. 41, 6, Article 195 (2022).

[4] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. 2017. Deep Reverse Tone Mapping. ACM Trans. Graph. (Proc. SIGGRAPH Asia) 36, 6 (2017).

[5] So Yeon Jo, Siyeong Lee, Namhyun Ahn, and Suk-Ju Kang. 2021. Deep Arbitrary HDRI: Inverse Tone Mapping with Controllable Exposure Changes. IEEE Trans. Multimedia (2021).

[6] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. 2018b. Deep recursive HDRI: Inverse tone mapping using generative adversarial networks. In ECCV. 596–611.

[7] Yang Zhang and TO Aydın. 2021. Deep HDR estimation with generative detail recon- struction. In Comp. Graph. Forum, Vol. 40. 179–190.

[8] Chao Wang, Ana Serrano, Xingang Pan, Bin Chen, Hans-Peter Seidel, Christian Theobalt, Karol Myszkowski, and Thomas Leimkuehler. 2023a. GlowGAN: Un- supervised Learning of HDR Images from LDR Images in the Wild.

(2023).

[9] Axel Sauer, Katja Schwarz, and Andreas Geiger. 2022. Stylegan-xl: Scaling StyleGAN to large diverse datasets. In Proc. SIGGRAPH. 1–10.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. NeurIPS 33 (2020), 6840–6851.

[12] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C. Kot, and Bihan Wen. 2023c. ExposureDiffusion: Learning to Expose for Low-light Image Enhance- ment. In ICCV.

[13] B. Fei, Z. Lyu, L. Pan, J. Zhang, W. Yang, T. Luo, B. Zhang, and B. Dai. 2023. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In CVPR. 9935– 9946.

[14] Qingsen Yan, Tao Hu, Yuan Sun, Hao Tang, Yu Zhu, Wei Dong, Luc Van Gool, and Yanning Zhang. 2023. Towards High-quality HDR Deghosting with Conditional Diffusion Models. arXiv:2311.00932 [cs.CV]

This project is implementing the paper:

[15]Bemana, M., Leimkühler, T., Myszkowski, K., Seidel, H.-P., Ritschel, T. (Year). Exposure diffusion: HDR image generation by consistent LDR denoising.Computer Vision and Pattern Recognition (cs.CV); 2024