PHYSHDR: WHEN LIGHTING MEETS MATERIALS AND SCENE GEOMETRY IN HDR RECONSTRUCTION

Hrishav Bakul Barua[‡†⋆], *Kalin Stefanov*[†], *Ganesh Krishnasamy*[‡], *KokSheik Wong*[‡§], *Abhinav Dhall*[†]

[‡]School of Information Technology, Monash University, Malaysia
[†]Faculty of Information Technology, Monash University, Australia
[⋆]Robotics and Autonomous Systems Lab, TCS Research, India
{hrishav.barua, kalin.stefanov, ganesh.krishnasamy, wong.koksheik, abhinav.dhall}@monash.edu

ABSTRACT

Low Dynamic Range (LDR) to High Dynamic Range (HDR) image translation is a fundamental task in many computational vision problems. Numerous data-driven methods have been proposed to address this problem; however, they lack explicit modeling of illumination, lighting, and scene geometry in images. This limits the quality of the reconstructed HDR images. Since lighting and shadows interact differently with different materials, (*e.g.*, specular surfaces such as glass and metal, and lambertian or diffuse surfaces such as wood and stone), modeling material-specific properties (*e.g.*, specular and diffuse reflectance) has the potential to improve the quality of HDR image reconstruction. This paper presents PhysHDR, a simple yet powerful latent diffusion-based generative model for HDR image reconstruction. The denoising process is conditioned on lighting and depth information and guided by a novel loss to incorporate material properties of surfaces in the scene. The experimental results establish the efficacy of PhysHDR in comparison to a number of recent state-of-the-art methods.

***Index Terms*—** latent diffusion, material modeling, high dynamic range, generative models, CLIP, depth maps

1. INTRODUCTION

Reconstructing High Dynamic Range (HDR) images from Low Dynamic Range (LDR) counterparts has gained significant attention in the vision community [1]. Applications involving medical and computational imaging [2], robotic vision and self-driving cars [3], augmented/virtual reality [4], media & entertainment [5], require high-fidelity images of real-world scenes, which LDR images generally lack. A large body of data-driven methods attempts to solve major issues in HDR imaging, pertaining to artifacts, ghosting, and blurring effects. These methods primary approximate the reverse of the image formation pipeline in standard cameras, where a camera captures HDR scenes with high intensity values and clips them to a low dynamic range [6].

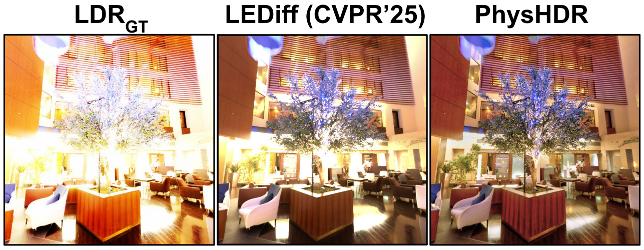

Fig. 1. PhysHDR (right) can recover the light/shadow details and light-object interactions in the rough and metallic surfaces better than the state-of-the-art [7] (middle) given an extremely over-exposed LDR image (left) as input.

The current state-of-the-art focuses on the retrieval of information in low light areas [8] or extreme lighting conditions [9]. Most methods utilize Convolutional Neural Networks (CNN) [10], Transformers [11] and Generative Adversarial Networks (GAN) [12, 13]. Some methods use single-exposed LDR [14, 6, 15], while others use multi-exposed LDR images as input [16, 17]. More recent methods are based on diffusion models [7, 18, 19], with some based on stable or latent diffusion and others on conditional diffusion [20]. These methods typically include a variational autoencoder (VAE) [21] to encode and decode the input and output, a U-Net for the denoising process, and a condition encoder such as Contrastive Language-Image Pre-Training (CLIP) [22] to embed the condition features. The denoising process takes place in the latent space instead of the pixel space. The main advantage of using diffusion over other generative-based or CNN-based methods is their ability to reconstruct high-resolution HDR images while mitigating artifacts and ghosting effects.

Although these methods reconstruct excellent HDR images, they lack explicit modeling of scene geometry (*i.e.*,

[⋆]This research is supported by the Global Excellence and Mobility Scholarship (GEMS), Monash University (Australia & Malaysia).

[§]This research is supported, in part, by the E-Science fund under the project: *Innovative High Dynamic Range Imaging - From Information Hiding to Its Applications* (Grant No. 01-02-10-SF0327).

depth information), illumination conditions, and scene material properties. HDR reconstruction is an ill-posed problem because multiple real-world lighting conditions can result in the same LDR pixel values (especially in under- and over-exposed regions). Therefore, it is technically challenging to disambiguate the pixel intensity values corresponding to the lighting and shadow areas in the reconstructed HDR. To this end, we propose PhysHDR, a latent diffusion approach which harnesses the power of both depth and illumination information to reconstruct HDR images in a more physics-informed manner. Inspired by the fact that different surfaces respond differently to light depending on their material properties (*e.g.*, lambertian and specular), PhysHDR introduces a novel loss for material properties (*i.e.*, albedo, roughness, and metallic) to guide the model in disambiguating the pixel intensities in the HDR image. Lambertian (diffuse) materials such as wood and stone reflect light in all directions uniformly, in contrast, specular materials such as glass and metal reflect light only in one direction. Depth combined with illumination information provides explicit 3D scene geometry, which helps to disentangle shading, lighting, and reflectance components. Fig. 1 illustrates the quality of HDR images reconstructed with PhysHDR compared to the recent state-of-the-art [7]. Our **key-contributions** are as follows: **(a)** proposing a novel latent diffusion method conditioned on depth and illumination information to model the shading, lighting, and reflectance properties of materials in an unambiguous manner; **(b)** proposing a new loss function based on material properties to further strengthen the efficacy in reconstructing light and shadow interactions with lambertian and specular surfaces, and; **(c)** presenting an extensive analysis of the method to highlight the contribution of each of the components.

2. METHOD

The goal is to reconstruct an HDR image $\hat{h} \in \mathbb{R}^{H\times W\times 3}$ with $\gg 2^8$ radiance values (having luminance and color information for each pixel) given a single LDR image $l \in \mathbb{R}^{H\times W\times 3}$ with 2^8 intensity values. The proposed latent diffusion model PhysHDR, similar to [23, 24], is conditioned on the input LDR image and its properties, *i.e.*, illumination and scene geometry (depth maps containing surface normal information [25]) beneficial for learning light-object interactions. The reconstructed HDR $\hat{h}$ and ground truth HDR h are decomposed into three material maps (albedo, roughness, and metallic) using [23], and employed in a novel loss function to further guide the diffusion process in reconstructing physics-informed light-object interactions. To the best of our knowledge, PhysHDR (see Fig. 2) is the first method to use material-based properties along with scene geometry (depth) and illumination information for HDR reconstruction.

Architecture: PhysHDR uses stable diffusion U-Net [20] as its base architecture. Diffusion models have shown excellent performance as effective priors trained using huge amounts of real data [26, 20].They transform Gaussian noise from the training distribution to data samples (*i.e.*, $\hat{h} \sim q(h \mid l)$) using an iterative denoising process. Similar to [20] we adopt the latent diffusion process with trainable encoder $\mathcal{E}(\cdot)$ (encodes the HDR image h in latent representation), and decoder $\mathcal{D}(\cdot)$ (decodes latent representation to HDR image $\hat{h}$). We use the LDR image l as a condition in three different ways. First, a pre-trained encoder $\bar{\mathcal{E}}(\cdot)$ (same architecture as $\mathcal{E}(\cdot)$) extracts features for l, which are concatenated with noise-induced features from h, making the input channel size six. Second, illumination features l_{ill} are extracted through the pre-trained ViT encoder from [27] and depth information l_{dep} is extracted with the pre-trained model Depth Anything (DA) [28]. Finally, the outputs of 1×1 convolutional layers are concatenated (*i.e.*, $l_{ill} \oplus l_{dep}$) and used in a pre-trained CLIP [22] encoder to extract image embedding l_{emb}. This embedding is used as a condition in the diffusion model's cross-attention U-Net during denoising. The illumination and depth-aware features provide the diffusion process with scene geometry and disentangled lighting and shadow effects with various objects and surfaces.

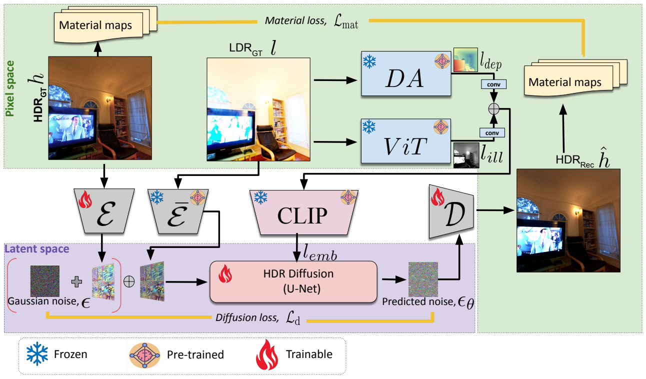
Material maps
Material loss, $\mathcal{L}_{\text{mat}}$
Pixel space
HDR_{GT} h
LDR_{GT} l
DA
l_{dep}
conv
$\oplus$
conv
ViT
l_{ill}
Material maps
HDR_{Rec} $\hat{h}$
$\mathcal{E}$
$\bar{\mathcal{E}}$
CLIP
$\mathcal{D}$
Latent space
l_{emb}
Gaussian noise, ϵ
HDR Diffusion
(U-Net)
Predicted noise, ϵ_θ
Diffusion loss, $\mathcal{L}_{\text{d}}$
Frozen
Pre-trained
Trainable

Fig. 2. Architecture of the proposed PhysHDR method.

Diffusion: During the forward process, in each timestep $t \sim [1, 1000]$ a Gaussian noise $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$ is added to $\mathcal{E}(h)$. In the reverse process, $\epsilon_\theta(\cdot)$, an U-Net [20], predicts the added noise, where θ denotes the model parameters. The denoising is conditioned on the features extracted from l and the noisy h. The training objective is:

$$\mathcal{L}_{\text{d}} = \mathbb{E}_{h,\, \epsilon\sim\mathcal{N}(0,I),\, t} \left[\|\epsilon - \epsilon_\theta(\mathcal{E}(h) + \epsilon, t, \bar{\mathcal{E}}(l), l_{emb})\|_2^2 \right] \quad (1)$$

During inference, we use the regular diffusion process [20] to sample HDR features $\hat{h}$ conditioned on the LDR image l. The advantage of using a pre-trained diffusion prior lies in the fact that it has been trained using a huge volume of real-world high resolution data [20, 23], *i.e.*, the LDR encoder $\bar{\mathcal{E}}(\cdot)$ is frozen. However, the HDR encoder $\mathcal{E}(\cdot)$ and decoder $\mathcal{D}(\cdot)$ are trained with our data. Therefore, the process finetunes only selected modules as illustrated in Fig. 2.

Table 1. Intra-dataset comparison with the state-of-the-art. The best results are in **bold** and the second best are underlined.

Method	City Scene [29]				HDR-Synth & HDR-Real [6]			
	PSNR↑	SSIM↑	LPIPS↓	VDP-3↑	PSNR↑	SSIM↑	LPIPS↓	VDP-3↑
Ghost-free [11]	**40.11**	0.955	0.143	7.47	38.12	0.943	0.157	7.44
GlowGAN [13]	34.01	0.902	0.167	7.41	33.10	0.901	0.171	7.32
HistoHDR-Net [14]	35.14	0.940	0.311	7.50	33.48	0.910	0.342	7.34
LEDiff [7]	35.91	0.922	0.121	7.52	35.11	0.905	0.138	7.48
PhysHDR (Ours)	39.01	**0.971**	**0.081**	**7.85**	**38.14**	**0.967**	**0.102**	**7.66**

Material loss: Apart from the diffusion loss in Eq. 1, we also propose a novel objective function based on material properties of the objects and surfaces in the HDR images. We first extract the material maps (*i.e.*, albedo, roughness, and metallic) from the ground truth HDR h and the reconstructed HDR $\hat{h}$ image using a state-of-the-art method which exhibits high accuracy in the task [23]. We get h_{al}, h_{ro}, and h_{met} from h and $\hat{h}_{al}$, $\hat{h}_{ro}$, and $\hat{h}_{met}$ from $\hat{h}$. Albedo maps $\langle h_{al}, \hat{h}_{al} \rangle$ consists of the base color information of the objects in the scene, while roughness $\langle h_{ro}, \hat{h}_{ro} \rangle$ and metallic $\langle h_{met}, \hat{h}_{met} \rangle$ maps represent the degree of roughness and smoothness in any objects or surfaces. While the diffusion loss ensures visual realism and produces attractive images, the material loss preserves physics-based properties of light and materials in the reconstructed HDR. This loss is computed on the tone-mapped versions of the material maps. This tone-mapping is performed using the μ-law [30] and is done to avoid the high-intensity pixels of HDR images that can distort the loss calculation. We define the loss between the material maps as:

$$\mathcal{L}_{\text{mat}} = \frac{1}{N}\sum_{n=1}^{N}\Big(\big\|h^n_{\text{al}} - \hat{h}^n_{\text{al}}\big\|_1 + \big\|h^n_{\text{ro}} - \hat{h}^n_{\text{ro}}\big\|_1 + \big\|h^n_{\text{met}} - \hat{h}^n_{\text{met}}\big\|_1\Big), \tag{2}$$

where N is the number of maps in each batch. The total objective of the model is:

$$\mathcal{L}_{\text{full}} = \mathcal{L}_d + \lambda_{\text{mat}}\,\mathcal{L}_{\text{mat}}, \tag{3}$$

where λ_{mat} is empirically set to 0.2 after experimenting with values from 0.1 to 0.5.

3. EXPERIMENTS AND RESULTS

Implementation: The finetuning of the pre-trained stable diffusion model [20] (implemented in PyTorch) was done for 200 epochs with a batch size of 10, using AdamW optimizer [31] with a learning rate of $1e{-}5$.

Datasets: To evaluate the performance of PhysHDR we used two datasets, including both real and synthetic images: City Scene dataset [29] (20K LDR/HDR image pairs) and HDR-Synth & HDR-Real dataset [6] (9785 LDR/HDR real image pairs and around 500 synthetic pairs). All images were resized to a resolution of 512×512. We compared the performance of different methods in two experiments: intra- and cross-dataset evaluation. For the intra-dataset experiment, we created 80% train and 20% test splits. For the cross-dataset experiment, we considered an additional dataset, DrTMO [32] (1043 LDR/HDR pairs). For methods designed to use single-exposure LDR inputs, we provided one LDR image from the datasets that contain multiple exposures. For methods that require multi-exposure LDR inputs (such as Ghost-free [11]), we synthetically generated (using the OpenCV function `convertScaleAbs`) the additional exposures for the datasets that contain only single-exposed LDR images. In all ablation studies, we used the same test set sampled from the City Scene dataset [29].

Metrics: We used four different metrics, *i.e.*, Peak Signal-to-Noise Ratio in dB (PSNR), Structural Similarity Index Measure (SSIM) [33], Learned Perceptual Image Patch Similarity (LPIPS), and High Dynamic Range Visual Differences Predictor (HDR-VDP-3) [34]. These metrics cover a wide range of evaluation parameters such as pixel-level similarity, structural similarity, and semantic and contextual similarity, and human-level perceptual judgement. We calculated HDR-VDP-3, SSIM, and LPIPS scores using ground truth and reconstructed HDR images in the linear domain. PSNR scores are obtained on μ-law tone-mapped ground truth and reconstructed HDR images.

Methods: We selected four different state-of-the-art methods: Ghost-free [11], GlowGAN [13], HistoHDR-Net [14], and LEDiff [7] in our evaluations. The selected methods include a CNN-based approach and generative approaches using GAN, Transformer, and Diffusion.

Quantitative Results: We present the intra-dataset results in Table 1. The proposed method outperforms the selected state-of-the-art methods in terms of PSNR, SSIM, LPIPS, and HDR-VDP-3 for HDR-Synth & HDR-Real [6] and SSIM, LPIPS, and HDR-VDP-3 for City Scene [29]. The Ghost-free [11] method outperforms our model in terms of PSNR on City Scene [29]. LEDiff [7] performs second best for semantic similarity and human vision-based metrics while Ghost-free [11] performs second best for the pixel-based and structural similarity metrics. The results of cross-dataset evaluation are presented in Table 2, and they show a similar trend as observed for intra-dataset evaluation, *i.e.*, our method PhysHDR outperforms the state-of-the-art in all the metrices.

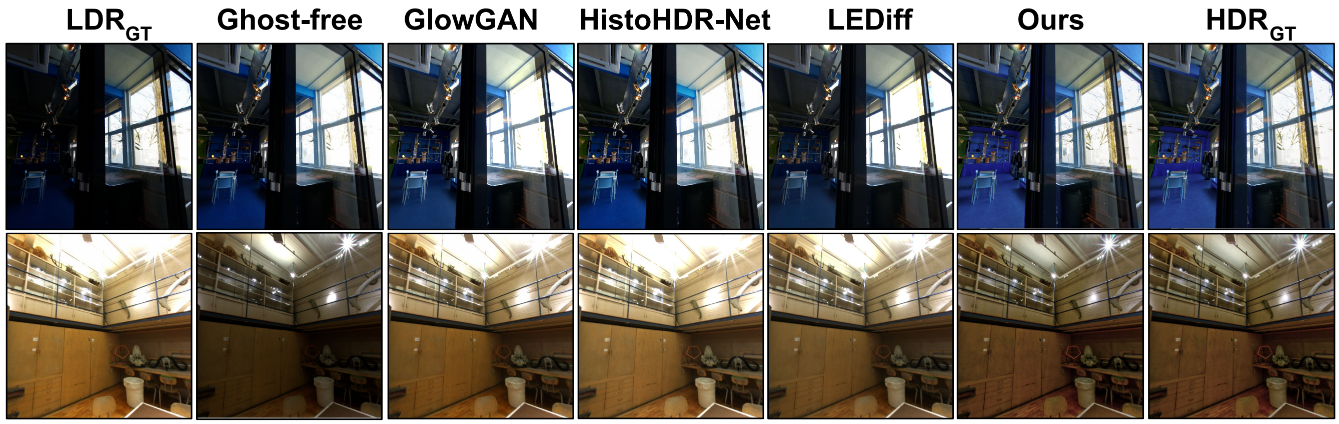

Fig. 3. HDR images reconstructed by the proposed PhysHDR and state-of-the-art methods. For our method, we specifically observe that the areas where light interacts with different surfaces are reconstructed realistically based on material properties.

Table 2. Cross-dataset evaluation of the proposed PhysHDR and state-of-the-art on unseen dataset (DrTMO [32]). The best results are in **bold** and the second best underlined.

Method	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**VDP-3↑**
Ghost-free [11]	36.77	0.935	0.155	7.39
GlowGAN [13]	33.21	0.899	0.177	7.35
HistoHDR-Net [14]	33.43	0.908	0.351	7.41
LEDiff [7]	34.12	0.909	0.142	7.52
PhysHDR (Ours)	**37.89**	**0.964**	**0.109**	**7.63**

Table 3. Architecture ablation results for different components. The best results are in **bold**. Baseline: [20] + $\mathcal{E}$ + $\bar{\mathcal{E}}$, and; CLIP: CLIP embedding from l.

Arch. components	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**HDR-VDP-3↑**
Baseline	27.62	0.857	0.403	6.71
+ CLIP	29.23	0.897	0.307	6.84
+ l_{dep}	29.11	0.891	0.288	6.91
+ l_{ill}	33.12	0.912	0.276	7.21
+ $l_{dep} \oplus l_{ill}$	34.12	0.934	0.126	7.43
+ l_{emb}	**39.01**	**0.971**	**0.081**	**7.85**

Table 4. Loss ablation results. The best results are in **bold**.

Loss	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**HDR-VDP-3↑**
$\mathcal{L}_\text{d}$	37.75	0.962	0.159	7.42
$\mathcal{L}_\text{d}$ + $\mathcal{L}_\text{mat}$	**39.01**	**0.971**	**0.081**	**7.85**

Qualitative Results: The visual quality of PhysHDR is illustrated in Fig. 3. Our outputs closely resemble the ground truth HDR in terms of lighting and shadow quality as well as physically plausible light-object interactions on rough and metallic surfaces. The base color (*i.e.*, albedo) of the objects in the scene are also preserved with high fidelity. We can also see the over-exposed and under-exposed areas with clarity. We display the ground truth and generated HDR images using Reinhard's tone-mapping algorithm [35].

Ablation Study: We performed ablation studies for architectural and loss components. Table 3 summarizes the results, where the proposed components are added one by one, and to compare the improvement over the baseline model (*i.e.*, U-Net-based denoising process with encoders $\mathcal{E}$ and $\bar{\mathcal{E}}$, and decoder $\mathcal{D}$). The second row illustrates the contribution of CLIP embeddings extracted from the LDR l, leading to a significant improvement in all metrics. The third and fourth rows illustrate the contribution of depth and illumination extracted from the LDR l, with small improvement for depth and a significant improvement for illumination. The fifth row illustrates the contribution of combined depth and illumination resulting in improvement in all metrics. The last row provides the results obtained when using all components of PhysHDR. PhysHDR is trained with two objectives, $\mathcal{L}_\text{d}$ and $\mathcal{L}_\text{mat}$. Table 4 provides an analysis of the contribution of the objectives.

4. CONCLUSION

Scene geometry (depth maps) and illumination information from the input image play a pivotal role in improving the performance of models for LDR to HDR reconstruction. CLIP-based information from the input LDR further improves the efficacy. The proposed material properties-based loss function ensures high-quality and perceptually realistic scene reconstruction, respecting the physics-based properties such as light-object and shadow-object interactions for different surfaces (*i.e.*, metallic or rough). Future work includes the study of diffusion models for HDR video reconstruction, as well as using material information directly as prior information in model training. Another important consideration will be the use of normal maps explicitly in model training.

5. REFERENCES

[1] Lin Wang and Kuk-Jin Yoon, "Deep Learning for HDR Imaging: State-of-the-Art and Future Trends," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8874–8895, 2021.

[2] Zhan Lu, Qian Zheng, Boxin Shi, and Xudong Jiang, "Pano-nerf: Synthesizing high dynamic range novel views with geometry from sparse low dynamic range panoramic images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38-4, pp. 3927–3935.

[3] Xuesong Wu, Hong Zhang, Xiaoping Hu, Moein Shakeri, Chen Fan, and Juiwen Ting, "Hdr reconstruction based on the polarization camera," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5113–5119, 2020.

[4] Shreyas Singh, Aryan Garg, and Kaushik Mitra, "Hdrsplat: Gaussian splatting for high dynamic range 3d scene reconstruction from raw images," *BMVC*, 2024.

[5] Gang He, Kepeng Xu, Li Xu, Chang Wu, Ming Sun, Xing Wen, and Yu-Wing Tai, "Sdrtv-to-hdrtv via hierarchical dynamic context feature mapping," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2890–2898.

[6] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang, "Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1651–1660.

[7] Chao Wang, Zhihao Xia, Thomas Leimkuhler, Karol Myszkowski, and Xuaner Zhang, "Lediff: Latent exposure diffusion for hdr generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 453–464.

[8] Shaoliang Yang, Dongming Zhou, Jinde Cao, and Yanbu Guo, "Lightingnet: An integrated learning method for low-light image enhancement," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 29–42, 2023.

[9] Hue Nguyen, Diep Tran, Khoi Nguyen, and Rang Nguyen, "Psenet: Progressive self-enhancement network for unsupervised extreme-light image enhancement," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1756–1765.

[10] Seungjun Shin, Kyeongbo Kong, and Woo-Jin Song, "Cnn-based ldr-to-hdr conversion system," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2018, pp. 1–2.

[11] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu, "Ghost-free high dynamic range imaging with context-aware transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 344–360.

[12] YoonChan Nam, JoonKyu Kim, Jae-hun Shim, and Suk-Ju Kang, "Deep conditional hdri: Inverse tone mapping via dual encoder-decoder conditioning method," *IEEE Transactions on Multimedia*, 2024.

[13] Chao Wang, Ana Serrano, Xingang Pan, Bin Chen, Karol Myszkowski, Hans-Peter Seidel, Christian Theobalt, and Thomas Leimkühler, "Glowgan: Unsupervised learning of hdr images from ldr images in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10509–10519.

[14] Hrishav Bakul Barua, Ganesh Krishnasamy, KokSheik Wong, Abhinav Dhall, and Kalin Stefanov, "Histohdr-net: Histogram equalization for single ldr to hdr image translation," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 2730–2736.

[15] Yunhao Zou, Chenggang Yan, and Ying Fu, "Rawhdr: High dynamic range image reconstruction from a single raw image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 12334–12344.

[16] Hrishav Bakul Barua, Ganesh Krishnasamy, KokSheik Wong, Kalin Stefanov, and Abhinav Dhall, "ArtHDR-Net: Perceptually Realistic and Accurate HDR Content Creation," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 806–812.

[17] Zhilu Zhang, Haoyu Wang, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo, "Self-supervised high dynamic range imaging with multi-exposure images in dynamic scenes," in *ICLR*, 2024.

[18] Qingsen Yan, Tao Hu, Yuan Sun, Hao Tang, Yu Zhu, Wei Dong, Luc Van Gool, and Yanning Zhang, "Toward high-quality hdr deghosting with conditional diffusion models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 4011–4026, 2023.

[19] Mojtaba Bemana, Thomas Leimkühler, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel, "Bracket diffusion: Hdr image generation by consistent ldr denoising," in *Computer Graphics Forum*. Wiley Online Library, 2025, p. e70086.

[20] 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*, 2022, pp. 10684–10695.

[21] Diederik P Kingma, Max Welling, et al., "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[23] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner, "Intrinsic image diffusion for indoor single-view material estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5198–5208.

[24] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 9298–9309.

[25] Hyunho Ha, Joo Ho Lee, Andreas Meuleman, and Min H Kim, "Normalfusion: Real-time acquisition of surface normals for high-resolution rgb-d scanning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15970–15979.

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[27] Ziteng Cui, Kunchang Li, Lin Gu, Shenghan Su, Peng Gao, ZhengKai Jiang, Yu Qiao, and Tatsuya Harada, "You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction," in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. 2022, BMVA Press.

[28] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.

[29] Jinsong Zhang and Jean-François Lalonde, "Learning High Dynamic Range from Outdoor Panoramas," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4519–4528.

[30] Takao Jinno, Hironori Kaida, Xinwei Xue, Nicola Adami, and Masahiro Okuda, "μ-Law Based HDR Coding and Its Error Analysis," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 94, no. 3, pp. 972–978, 2011.

[31] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017.

[32] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani, "Deep reverse tone mapping," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 177–1, 2017.

[33] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[34] Peibei Cao, Rafal K Mantiuk, and Kede Ma, "Perceptual assessment and optimization of hdr image rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22433–22443.

[35] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, jul 2002.