LLM-HDR: Bridging LLM-based Perception and Self-Supervision for Unpaired LDR-to-HDR Image Reconstruction

Hrishav Bakul Barua
Monash University & TCS Research
hrishav.barua@monash.edu

Kalin Stefanov
Monash University
kalin.stefanov@monash.edu

Lemuel Lai En Che
Monash University
lemuel.lai03@gmail.com

Abhinav Dhall
Flinders University
abhinav.dhall@flinders.edu

KokSheik Wong
Monash University
wong.koksheik@monash.edu

Ganesh Krishnasamy
Monash University
ganesh.krishnasamy@monash.edu

Abstract

*The translation of Low Dynamic Range (LDR) to High Dynamic Range (HDR) images is an important computer vision task. There is a significant amount of research utilizing both conventional non-learning methods and modern data-driven approaches, focusing on using both single-exposed and multi-exposed LDR for HDR image reconstruction. However, most current state-of-the-art methods require high-quality paired {LDR,HDR} datasets for model training. In addition, there is limited literature on using unpaired datasets for this task, that is, the model learns a mapping between domains, i.e., LDR ↔ HDR. This paper proposes **LLM-HDR**, a method that integrates the perception of Large Language Models (LLM) into a modified semantic- and cycle-consistent adversarial architecture that utilizes unpaired {LDR,HDR} datasets for training. The method introduces novel artifact- and exposure-aware generators to address visual artifact removal and an encoder and loss to address semantic consistency, another underexplored topic. **LLM-HDR** is the first to use an LLM for the LDR ↔ HDR translation task in a self-supervised setup. The method achieves state-of-the-art performance across several benchmark datasets and reconstructs high-quality HDR images. The official website of this work is available at: https://github.com/HrishavBakulBarua/LLM-HDR*

1. Introduction

High Dynamic Range (HDR) [3, 70] images capture a wider range of intensity values compared to their Low Dynamic Range (LDR) or Standard Dynamic Range (SDR) counterparts, which have a pixel bit depth of only 2^8 intensity levels. From human vision perspective, LDR images are often not visually pleasing, while from the computer/robot vision applications perspective, they hold limited information.

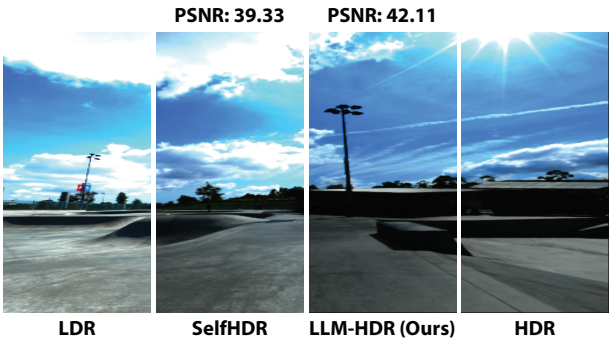


Figure 1. Qualitative comparison of the proposed LLM-HDR method and state-of-the-art method SelfHDR [116] (ICLR'24). We see, our method handles the overexposed portions in the sky more realistically.

Most commercial devices produce and display LDR content, and some specialized hardware can support HDR, *e.g.*, HDR cameras and sensors [45], which can capture images with more than 256 intensity levels, and HDR displays [14], which can display HDR intensity levels. Given that specialized hardware that captures HDR is expensive, researchers have been investigating approaches for accurate HDR reconstruction from LDR (*i.e.*, inverse tone-mapping [98]) using conventional non-learning methods and modern data-driven approaches. For display purposes, however, one often needs to tone-map [33] the HDR content to LDR to fit the intensity range supported by standard hardware.

Early HDR reconstruction methods address tasks such as image enhancement, *e.g.*, recovering missing information due to extreme lighting conditions [76] or low lighting conditions [57, 109] and compression [13]. Current applications of HDR reconstruction span multiple domains, including, robotics/machine vision [103], media and enter-

tainment [35], gaming, mixed reality, and novel view synthesis [37, 66, 88, 91], as well as medical imaging [37].

Early data-driven methods utilize Convolutional Neural Networks (CNN) [90], Transformers [65, 89] and Generative Adversarial Networks (GAN) [77, 82]. Recent approaches for HDR image and/or 3D HDR scene reconstruction are based on Diffusion Models [8, 19, 28, 95], Neural Radiance Fields (NeRF) [37], and Gaussian Splatting [91]. Some use multi-exposed [5, 22, 29, 30, 55, 60, 64, 85, 87, 118], while others use single-exposed LDR [6, 11, 116] paired with HDR images for supervised training.

Large Language Models (LLM) [75] and Vision-Language Models (VLM) [105, 114] have been widely used in many vision tasks, *e.g.*, image generation [92], deepfake detection [41], semantic segmentation [97], image relighting [15], and adverse weather condition removal [108]. However, the human-defined knowledge of such models has not been utilized for LDR-to-HDR image reconstruction.

Although current methods achieve excellent results in reconstructing HDR images, most of them require proper {LDR,HDR} paired datasets [7, 23, 64] for training. Consequently, the quality of the state-of-the-art data-driven HDR reconstruction methods depends on the quality of the available paired {LDR,HDR} datasets. There is a research gap in the field; on the one hand, the literature on unpaired HDR reconstruction is extremely limited [61, 94], and on the other, only a few approaches utilize semantic and contextual information to guide the reconstruction [27, 28, 65, 96].

To address this gap, we propose **LLM-HDR**, a method that leverages LLM-driven cycle consistency [17, 117] objective using unpaired data, where the model learns a mapping between domains, *i.e.*, LDR $\leftrightarrow$ HDR. In addition, the method ensures semantic consistency between the LDR and reconstructed HDR by utilizing Contrastive Language-Image Pretraining (CLIP) [81] embeddings and loss based on semantic segmentation. Moreover, LLM-based artifact- and exposure-aware information further supports the training process to deliver more realistic and natural HDR image generation. The proposed method reconstructs artifact-free and visually impressive HDR from single-exposed LDR images. It also outperforms the most recent state-of-the-art which predominantly uses paired datasets for training. Our work makes the following contributions:

- We introduce the first semantic- and cycle consistency-guided self-supervised learning method for unpaired {LDR,HDR} data which addresses both the inverse tone-mapping (*i.e.*, LDR $\rightarrow$ HDR) and tone-mapping (*i.e.*, HDR $\rightarrow$ LDR) tasks (Sec. 3).
- We leverage LLM-based loss and artifact- and exposure-aware saliency maps using human-defined prompting to further refine those areas in the reconstructed HDR images (Secs. 3.1 and 3.2).
- We propose a novel generator based on a modified U-Net architecture [86] that incorporates ConvLSTM-based artifact-aware feedback mechanism [36, 63] and exposure-aware skip connections to mitigate visual artifacts in the HDR reconstruction (Sec. 3.1).
- We propose a CLIP embedding encoder for contrastive learning to minimize the semantic difference between LDR and reconstructed HDR image pairs, while maximizing the difference between LDR and other {LDR,HDR} image pairs (Sec. 3.1).
- We propose a novel loss function based on the Mean Intersection over Union (mIoU) metric to further ensure semantic consistency between the LDR and reconstructed HDR images (Sec. 3.2).
- We perform thorough experimental validation for the contribution of all proposed components, both qualitatively and quantitatively (Sec. 5).

2. Related Work

Learning-based approaches use either single-exposed LDR as input [22, 29, 30, 55, 60, 64, 87, 118] or alternatively, multi-exposed LDR [5, 11, 77, 85, 116]. Some of the multi-exposed methods use novel feature fusion or aggregation methods [104, 107, 111]. Barua *et al*. [6] harnessed the concept of histogram equalization of input LDR, in addition to the original LDR, to overcome the contrast and hue miss-representation in over/underexposed areas of the LDR. Cao *et al*. [12] proposed a channel-decoupled kernel-based approach which combines the output HDR with the output of another architecture in a pixel-wise fashion. Luzardo *et al*. [67] proposed a method to enhance the artistic intent of the reconstructed HDR by enhancing the peak brightness in the reconstruction process. In addition, recently, Neural Radiance Fields [37, 66, 72, 73], Diffusion Models [19, 28, 95], and Gaussian Splatting [91] have also been used to improve the reconstruction of HDR content.

Some methods use weakly-supervised [55], self-supervised [116] and unsupervised [76, 94] approaches for HDR reconstruction. Le *et al*. [55] proposed an indirect approach for HDR reconstruction from multi-exposed LDR using weak supervision. The method first outputs a stack of multi-exposed LDR, which are then merged using the state-of-the-art tool Photomatix [34] to obtain an HDR. Zhang *et al*. [116] proposed a self-supervised approach for HDR reconstruction. The method requires multi-exposed LDR to learn the reconstruction network during training without the need for HDR counterparts. Nguyen *et al*. [76] proposed an unsupervised approach to recover image information from overexposed areas of LDR. This approach does not require any HDR for supervision and instead uses pseudo-ground truth images. Lee *et al*. [56] proposed a method with limited supervision for multi-exposed LDR generation consisting of a pair of networks that generate LDR with higher and lower exposure levels. Then the two exposure levels are

combined to generate the HDR. Wang *et al*. [94] proposed an unsupervised approach for HDR reconstruction. They train the model in such a way that when the HDR is re-projected to LDR using a camera response function model it becomes indistinguishable from the original LDR (Sec. S1 for related works on non-learning methods).

GAN-Based Approaches. Generative Adversarial Networks [2, 26, 94] are well-explored in this domain. Niu *et al*. [77] proposed a GAN-based method that can handle images with foreground motions by fusing multi-exposed LDR and extracting important information from the over/underexposed areas of the LDR. Raipurkar *et al*. [82] proposed a conditional GAN architecture that adds details to the saturated regions of the input LDR using a pre-trained segmentation model to extract exposure masks. Guo *et al*. [29] proposed a two-stage pipeline that extracts the over/underexposed features with high accuracy. GAN with attention mechanism first generates the missing information in those extreme exposure areas, and in the second stage, a CNN with multiple branches fuses the multi-exposed LDR from the previous stage to reconstruct the HDR. Nam *et al*. [74] proposed a GAN method that uses exposure values for conditional generation of multi-exposure stack that adapts well to varying color and brightness levels.

Li *et al*. [61] proposed a GAN-based unsupervised method for unpaired multi-exposure LDR-to-HDR translation. The method introduces modified GAN loss and a novel discriminator to tackle ghosting artifacts caused from the misalignment in the LDR stack and HDR. Wu *et al*. [102] proposed CycleGAN-like architecture for low-light image enhancement tasks using unpaired data for training. The method fuses the Retinex theory [54] with CycleGAN [117] concept to enhance the lighting conditions globally, recover color and reduce noise in the output HDR.

Semantic and Knowledge-Based Approaches. Some methods attempt to use semantic/context information in the image or image formation process for HDR reconstruction. Wang *et al*. [96] proposed a method that approximates the inverse of the camera pipeline. Their knowledge-inspired block uses the knowledge of image formation to address three tasks during HDR reconstruction: missing details recovery, adjustment of image parameters, and reduction of image noise. Goswami *et al*. [28] proposed a method to recover the clipped intensity values of an LDR/SDR image due to the tone-mapping process in the camera. The proposed method works in two stages: first it uses a semantic graph-based guidance to help the diffusion process with the in-painting of saturated image parts, and second, the problem is formulated as HDR in-painting from SDR in-painted regions. Liu *et al*. [65] proposed a vision transformer approach with context-awareness to remove ghosting effects in the output HDR. The model is a dual-branch architecture to capture both local and global context in the input image that enables the generation process to remove unwanted information in the output HDR and avoid artifacts.

Table 1. Summary of recent state-of-the-art methods. **I/O:** LDR used as input, single-exposed (SE) and multi-exposed (ME), **O/P:** Reconstructs directly HDR (D) or multi-exposed LDR stack (I), **UP:** Can be trained with unpaired data, **LLM:** Uses LLM-based perception, **Context:** Uses local/global image information and relationship among entities in the image, **Semantics:** Uses color/texture information and identity of the items in the image, **Artifacts:** Handles visual artifacts in heavily over/underexposed areas, **TM**: Also performs tone-mapping *i.e*. HDR → LDR.

Method	I/O	O/P	UP	LLM	Context	Semantics	Artifacts	TM
PSENet [76]	SE	D	✗	✗	✗	✗	✗	✗
SingleHDR(W) [55]	SE	I	✗	✗	✗	✗	✗	✗
UPHDR-GAN [61]	ME	D	✓	✗	✗	✗	✓	✗
SelfHDR [116]	ME	I	✗	✗	✗	✗	✓	✗
KUNet [96]	SE	D	✗	✗	✗	✓	✗	✗
Ghost-free HDR [65]	ME	D	✗	✗	✓	✗	✓	✗
GlowGAN-ITM [94]	SE	D	✓	✗	✗	✗	✓	✗
DITMO [28]	SE	I	✗	✗	✗	✓	✓	✗
LLM-HDR (Ours)	**SE**	**D**	✓	✓	✓	✓	✓	✓

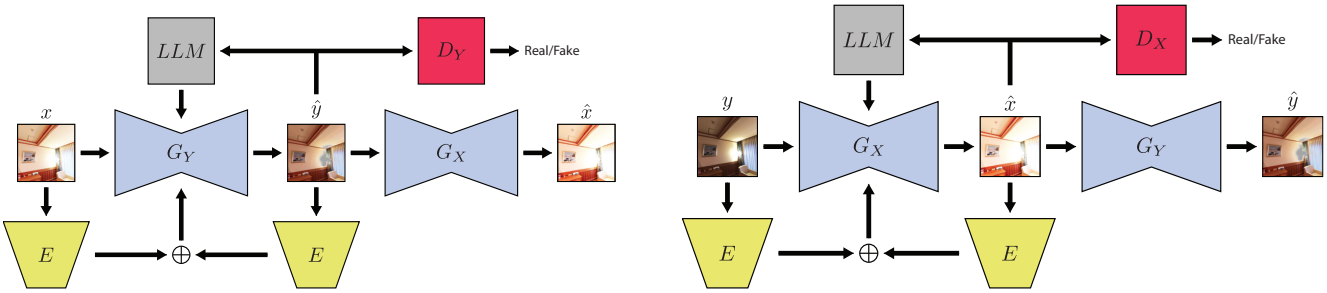
Limitations. Our analysis reveals that the methods based on single-exposed LDR fail to preserve image characteristics such as the levels of color hue and saturation, contrast, and brightness intensity in HDR. On the other hand, methods based on multi-exposed LDR produce unwanted visual artifacts such as halo in edges and boundaries, outlier pixel intensities, unwanted repeated pattens, missing details in shadow and highly bright regions, irregular color and texture transitions, and ghosting effects from dynamic scenes. While most methods produce excellent visual quality, the main limitations of these approaches are the requirement of paired {LDR,HDR} datasets for training, and the lack of semantic and contextual knowledge guidance in the reconstruction process that could significantly improve the quality of HDR output. Finally, LLM- and VLM-based [105, 114] perception capabilities have not been utilized in this problem area, which might improve performance. Tab. 1 summarizes a comparison between the state-of-the-art and our method in terms of various parameters.

3. Method

This section describes the proposed method - the first LLM-driven semantic and cycle consistency guided self-supervised learning approach for unpaired {LDR,HDR} data which addresses both the inverse tone-mapping (*i.e*., LDR → HDR) and tone-mapping (*i.e*., HDR → LDR) tasks.

3.1. Architecture Modules

We adopt an cycle-consistent adversarial [117] architecture as the basis of our method depicted in Fig. 2. The network includes two generators and two discriminators. The network also takes advantage of CLIP encoders and LLM-



(a) Forward cycle consistency: $x \rightarrow G_Y(x) \rightarrow G_X(G_Y(x)) \approx x$.

(b) Backward cycle consistency: $y \rightarrow G_X(y) \rightarrow G_Y(G_X(y)) \approx y$.

Figure 2. Overview of the proposed method architecture (Sec. 3.1, Sec. S2.2) where x and y represent LDR and HDR images, respectively. The method is trained with six objectives: adversarial, cycle consistency, identity, LLM-based, contrastive, and semantic segmentation (Sec. 3.2, Sec. S2.3).

based perception to further bridge the gap between semantic and perceptual differences between input and generated images (Sec. S2.1 for a concise background on the key concepts employed in our method). Let us denote the two domains as X for LDR images and Y for HDR images. Furthermore, let us denote the generator used in the forward cycle that maps images from LDR to HDR as G_Y and the one used in the backward cycle that maps images from HDR to LDR as G_X. Finally, let us denote the discriminator that discriminates between the reconstructed LDR and real LDR images as D_X and the one that discriminates between the reconstructed HDR and real HDR images as D_Y. The images from the two domains can be represented as $\{x_i\}_{i=1}^{N}$ where $x_i \in X$ and $\{y_j\}_{j=1}^{M}$ where $y_j \in Y$. The data distribution of the two domains can be represented as $x \sim p_{\text{data}}(x)$ and $y \sim p_{\text{data}}(y)$.

Generators. The generators G_Y and G_X are based on U-Net architecture [86] that includes an encoder and a decoder block with skip connections from each level of the encoder to the decoder. In our generators, we propose a feedback mechanism [36, 113] between the encoder and decoder block. The rationale behind the feedback is to refine the features extracted from the encoder (during the first iteration of the feedback) to guide the decoder for better output image reconstruction over the rest of the iterations. Hence, the feedback block not only iterates over its own output but also re-runs the decoder situated ahead of it in each iteration while keeping the encoder frozen until the feedback iteration completes. The feedback is implemented with a ConvLSTM [36, 63] network and the number of iterations is fixed to 4 based on an ablation experiment testing 2, 3, and 4 iterations setups (Sec. S2.2 for visual depiction of the proposed generators and implementation details).

Discriminators. The discriminators D_X and D_Y are based on [39, 61] (Sec. S2.2 for implementation details).

LLM. We adopt a pre-trained LLM [38] for zero-shot Q&A sessions (using human-defined knowledge) in each iteration of the training process. We prompt the LLM to provide the pixels identified as artifacts, and over/underexposed regions for both the reconstructed HDR and LDR. Given that these three factors constitute the nature of the LDR $\leftrightarrow$ HDR translation, the pixels proposed by the LLM are used in a LLM-based loss function. In addition, the LLM generates a separate saliency map for each of those areas, *i.e.*, artifact, overexposed, and underexposed maps. The artifact map is multiplied with the input features to the feedback mechanism of the generators G_Y and G_X. The underexposed map is fed into each of the skip connections (levels 1 to 3) in the generator G_Y only and the overexposed map is multiplied with the input features to the bottleneck layer of this generator. We use element-wise multiplication as a gating mechanism to emphasize the selected areas in the maps. The rationale behind the separated saliency map fusion is based on the proposed U-Net architecture where distorted pixels are mostly refined by the feedback mechanism and later layers, high-intensity lighting details are generally recovered by the middle or later layers, and low-intensity shadows are reconstructed realistically at the starting and ending layers (Sec. S2.2 for visual depiction of the proposed LLM module and implementation details).

Encoders. We also introduce a CLIP embeddings encoder E. Specifically, we use a pre-trained CLIP encoder [81] to extract image embeddings with both local and global semantic context. For the forward cycle consistency we use $E(x)$ and $E(\hat{y})$. Then we add the embeddings and feed them back to the bottleneck layer of the G_Y decoder. Similarly, for the backward cycle consistency we add the embeddings from $E(y)$ and $E(\hat{x})$ and feed them back to the bottleneck layer of the G_X decoder (Sec. S2.2 for implementation details).

3.2. Loss Functions

We introduce six loss functions to train the proposed method. We introduce three novel loss functions for HDR reconstruction - LLM-based, contrastive, and semantic segmentation loss. Two are standard loss functions, *i.e.*, adver-

sarial loss [25] and cycle consistency loss [117]. Inspired from [117] we also use an identity loss [93].

The calculations in all loss functions are based on tone-mapped versions of the reconstructed and real HDR images. This tone-mapping is performed on the basis of the μ-law [43] and is done to avoid the high-intensity pixels of HDR images that can distort the loss calculation. The tone-mapping operator T can be represented as follows:

$$T(y_j) = \frac{log(1+\mu y_j)}{log(1+\mu)}, \tag{1}$$

where the amount of compression μ is 5000 following [47].

LLM Loss. Given a reconstructed HDR/LDR image, this loss is based on pixels identified by the LLM as artifacts and over/underexposed regions during Q&A sessions (using human-defined knowledge). The loss is defined as:

$$\mathcal{L}_{\text{llm}} = 3 \times \frac{\hat{y_{\text{af}}}}{\hat{y_{\text{total}}}} + 2 \times \frac{\hat{y_{\text{ox}}}}{\hat{y_{\text{total}}}} + 1.5 \times \frac{\hat{y_{\text{ux}}}}{\hat{y_{\text{total}}}}, \tag{2}$$

where $\hat{y_{\text{total}}}$ represents the total number of pixels in the reconstructed HDR $\hat{y}$, while $\hat{y_{\text{af}}}$, $\hat{y_{\text{ox}}}$, and $\hat{y_{\text{ux}}}$ denote the total number of pixels in the areas perceived by the LLM as artifacts, overexposed, and underexposed. The weights on each of the three components are selected after thorough ablations as well as theoretical considerations. Our analysis suggests that artifact mitigation is more important than exposure issues for the perceptual quality of the reconstructed images because artifacts make the images unnatural and unpleasing to the human eye. Humans have greater tolerance towards dark areas than faded bright regions hence we give lower weight to the third component. (Sec. S2.3 for visual depiction of the LLM loss). This loss is calculated in the backward cycle with $\hat{x}$ but only with the artifact component. The rationale behind this decision lies in the nature of the output, *i.e.*, LDR can be overexposed or underexposed.

Contrastive Loss. This loss is based on embeddings extracted using a CLIP encoder and ensures semantic information preservation across domains. Here, we do not directly extract the embeddings from LDR images but instead, use a histogram-equalized version processed using the OpenCV function `equalizeHist` [10]. Histogram equalization improves the pixel visibility in extreme lighting areas or areas with shadows/darkness of an image by re-adjusting the contrast and saturation levels. This is done by spreading out the frequent pixel intensity values across 256 bins. Equalization often leads to revealing hidden semantic information or non-perceivable objects in an image. For the image embedding $\bar{x}$ from $E(x)$ and $\bar{y}$ from $E(\hat{y})$ we first define the cosine similarity between them as follows:

$$\text{sim}(\bar{x}, \bar{y}) = \frac{\bar{x} \cdot \bar{y}}{\|\bar{x}\| \|\bar{y}\|}, \tag{3}$$

where $\cdot$ represents the dot product between the two embeddings and $\|\cdot\|$ represents the norm of the embeddings. We formulate the contrastive loss for an input batch as positive pairs of images (*e.g.*, LDR and the corresponding reconstructed HDR) and negative pairs of images (*e.g.*, each LDR with the rest of the LDR as well as the rest of the reconstructed HDR in the batch) as follows:

$$\mathcal{L}_{\text{con}} = -\frac{1}{N} \times \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\bar{x}_i, \bar{y}_i)/\tau)}{\sum_{\substack{j=1 \\ j \neq i}}^{N} (\exp(\text{sim}(\bar{x}_i, \bar{x}_j)/\tau) + \exp(\text{sim}(\bar{x}_i, \bar{y}_j)/\tau))}, \tag{4}$$

where N is the batch size, exp represents the exponential function, and τ represents the temperature parameter which controls the amount of emphasis given in distinguishing between positive and negative pairs. This loss is calculated in both the forward and backward cycles with the input and output being swapped. This loss replicates the contrastive learning paradigm of CLIP for {image,image} instead of {image,text} pairs (Secs. S2.3 and S3.3 for visual depiction of the contrastive loss and an ablation experiment for τ).

Semantic Segmentation Loss. This loss is based on segmentation masks. The Mean Intersection over Union (mIoU) metric measures the amount of overlap between ground truth and predicted segmentation masks. Similar to the previous loss function, we use histogram-equalized versions of the LDR images processed using the OpenCV function `equalizeHist` [10]. We choose equalized images instead of original LDR because segmentation of low-light or extremely bright images does not yield good results. We use Segment Anything (SAM) [50] to generate segmentation classes for the histogram-equalized LDR and reconstructed tone-mapped HDR images (Sec. S2.3 for visual depiction of the semantic segmentation loss). This loss component helps in mitigating differences in boundary and edge pixels between the LDR and HDR images. We can define the IoU metric as:

$$\text{IoU}_c = \frac{x_c \cap \hat{y_c}}{x_c \cup \hat{y_c}}, \tag{5}$$

where x_c and $\hat{y_c}$ represent the segmentation for class c in image x and $\hat{y}$, respectively. $\cap$ represents the overlapping area of predicted and ground truth pixels while $\cup$ represents the total area covered by predicted and ground truth pixels for class c. The mean IoU over all segmentation classes can be formulated as:

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \text{IoU}_c. \tag{6}$$

We define the semantic segmentation loss as:

$$\mathcal{L}_{\text{sem}} = 1 - \text{mIoU}. \tag{7}$$

This loss is calculated in both forward and backward cycles with the input and output being swapped (Sec. S2.3 for visual depiction of the semantic segmentation loss).

Adversarial Loss. We apply adversarial loss to both mappings. The mapping from LDR to HDR domain, *i.e.*, $G_Y : X \rightarrow Y$ with the discriminator D_Y, can be expressed as:

$$\begin{aligned} &\mathcal{L}_{\text{GAN}}(G_Y, D_Y, X, Y) = \\ &\mathbb{E}_{y \sim p_{\text{data}}(y)} \left[\log D_Y(y)\right] + \\ &\mathbb{E}_{x \sim p_{\text{data}}(x)} \left[\log(1 - D_Y(G_Y(x)))\right], \end{aligned} \tag{8}$$

where G_Y generates images that look similar to images from domain Y while D_Y distinguishes between generated samples $\hat{y}$ and real samples y. G_Y aims to minimize this objective against D_Y that aims to maximize it, *i.e.*, $\min_{G_Y} \max_{D_Y} \mathcal{L}_{\text{GAN}}(G_Y, D_Y, X, Y)$.

The mapping from HDR to LDR domain, *i.e.*, $G_X : Y \rightarrow X$ with the discriminator D_X, can be expressed as:

$$\begin{aligned} &\mathcal{L}_{\text{GAN}}(G_X, D_X, Y, X) = \\ &\mathbb{E}_{x \sim p_{\text{data}}(x)} \left[\log D_X(x)\right] + \\ &\mathbb{E}_{y \sim p_{\text{data}}(y)} \left[\log(1 - D_X(G_X(y)))\right], \end{aligned} \tag{9}$$

where G_X generates images that look similar to images from domain X while D_X distinguishes between generated samples $\hat{x}$ and real samples x. Similar to above, G_X aims to minimize this objective against D_X that aims to maximize it, *i.e.*, $\min_{G_X} \max_{D_X} \mathcal{L}_{\text{GAN}}(G_X, D_X, Y, X)$.

Cycle Consistency Loss. The adversarial loss does not guarantee learning without contradiction *i.e.*, the forward $G_Y : X \rightarrow Y$ and backward $G_X : Y \rightarrow X$ mappings might not be consistent with each other. Hence, we also incorporate a cycle consistency loss to prevent mutual contradiction of the learned mappings G_Y and G_X. For each image x_i from the LDR domain, the cycle of reconstruction *i.e.*, from LDR to HDR and then back to LDR must result back in the original image x_i. Hence, we can define the forward cycle consistency as: $x_i \rightarrow G_Y(x_i) \rightarrow G_X(G_Y(x_i)) \approx x_i$. Similarly, the backward cycle consistency can be represented as: $y_j \rightarrow G_X(y_j) \rightarrow G_Y(G_X(y_j)) \approx y_j$. The loss can be formulated as:

$$\begin{aligned} &\mathcal{L}_{\text{cyc}}(G_Y, G_X) = \\ &\mathbb{E}_{x \sim p_{\text{data}}(x)} \left[\|G_X(G_Y(x)) - x\|_1\right] + \\ &\mathbb{E}_{y \sim p_{\text{data}}(y)} \left[\|G_Y(G_X(y)) - y\|_1\right], \end{aligned} \tag{10}$$

where $\|\cdot\|_1$ is the L1 norm (Sec. S2.3 for visual depiction of the cycle consistency loss).

Identity Loss. For LDR $\leftrightarrow$ HDR translation tasks, we also found that adversarial and cycle consistency formulations alone cannot preserve the color and hue information. This is due to incorrect mapping of color shades from LDR to HDR domains by the generators stemming from the underlying difference in dynamic ranges. Therefore, we force the generators to replicate an identity mapping by providing target domain images. This loss can be expressed as:

$$\begin{aligned} &\mathcal{L}_{\text{id}}(G_Y, G_X) = \\ &\mathbb{E}_{y \sim p_{\text{data}}(y)} \left[\|G_Y(y) - y\|_1\right] + \mathbb{E}_{x \sim p_{\text{data}}(x)} \left[\|G_X(x) - x\|_1\right]. \end{aligned} \tag{11}$$

Final Loss. The full objective is expressed in Eq. (12), where λ scales the relative importance of the cycle consistency and identity loss. λ is set to 10 in our experiments and the weight for identity loss is 0.5 inspired from the setup in the original cycle consistency work [117]. α and β are weights for the contrastive and semantic segmentation loss, respectively. Both values are set to 2 (Sec. S3.3 for ablation experiments for α and β).

$$\begin{aligned} &\mathcal{L}_{\text{full}} = \\ &\mathcal{L}_{\text{GAN}}(G_Y, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(G_X, D_X, Y, X) + \\ &\lambda(\mathcal{L}_{\text{cyc}}(G_Y, G_X) + 0.5 \times \mathcal{L}_{\text{id}}(G_Y, G_X)) + \\ &\alpha\mathcal{L}_{\text{con}} + \beta\mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{llm}} \end{aligned} \tag{12}$$

4. Experiments

We discuss the experimental setup details in Sec. S2.4.

Datasets. For the primary comparison of our method with the state-of-the-art, we consider the HDRTV [16], NTIRE [79], and HDR-Synth & HDR-Real [64] paired datasets. HDRTV has 1235 samples for training and 117 for testing. The NTIRE dataset consists of approximately 1500 training, 60 validation, and 201 testing samples. HDR-Synth & HDR-Real dataset consists of 20537 samples. Other paired datasets used in our evaluations are DrTMO [23] (1043 samples), Kalantari [44] (89 samples), HDR-Eye [51] (46 samples), and LDR-HDR Pair [40] (176 samples). We choose these datasets because they consist of a balance between real and synthetic images as well as image and scene diversity. We used the pre-defined train/test sets of HDRTV and NTIRE. For HDR-Synth & HDR-Real, we performed a random $80/20$ split for training and testing in all experiments unless specified otherwise. For methods working with single-exposed LDR inputs, we use only one LDR from datasets with multi-exposed LDR. For methods working with multi-exposed LDR inputs, we generate the required exposures using the OpenCV function `convertScaleAbs` [10] for datasets with only single-exposed LDR images.

Table 2. HDR reconstruction result. Comparison with supervised (gray) and unsupervised, weakly-supervised and self-supervised (black) learning methods trained and evaluated on the paired datasets HDRTV [16], NTIRE [79] and HDR-Synth & HDR-Real [64]. **LP:** Supervised (S), unsupervised (US), weakly-supervised (WS), and self-supervised (SS).

Method	LP	PSNR↑	SSIM↑	LPIPS↓	Q-Score↑
HDRCNN (*ACM TOG'17*) [22]	S	20.12	0.724	0.422	60.12
DrTMO (*ACM TOG'17*) [23]	S	20.43	0.711	0.412	60.81
ExpandNet (*Eurographics'18*) [69]	S	21.44	0.854	0.451	58.12
FHDR (*GlobalSIP'19*) [47]	S	20.84	0.841	0.307	62.45
SingleHDR (*CVPR'20*) [64]	S	21.01	0.851	0.402	64.21
Two-stage (*CVPRW'21*) [1]	S	34.29	0.856	0.287	61.78
HDR-GAN (*IEEE TIP'21*) [77]	S	30.11	0.912	0.411	65.01
KUNet (*IJCAI'22*) [96]	S	37.21	0.974	0.051	62.14
Ghost-free (*ECCV'22*) [65]	S	**40.32**	0.966	0.043	63.02
ArtHDR-Net (*APSIPA'23*) [5]	S	37.12	0.961	0.321	63.43
HistoHDR-Net (*ICIP'24*) [6]	S	38.21	0.968	0.301	65.15
UPHDR-GAN (*IEEE TCSVT'22*) [61]	US	39.98	0.965	0.044	63.54
PSENet (*WACV'23*) [76]	US	27.35	0.856	0.274	62.89
SingleHDR(W) (*WACV'23*) [55]	WS	30.79	0.952	0.055	62.95
GlowGAN-ITM (*ICCV'23*) [94]	US	30.19	0.901	0.064	60.05
SelfHDR (*ICLR'24*) [116]	SS	39.51	0.972	0.037	64.77
LLM-HDR (Ours)	SS	40.11	**0.979**	**0.020**	**68.51**

Metrics. We use four metrics to report the results. High Dynamic Range Visual Differences Predictor (HDR-VDP-2) [68] or Mean Opinion Score Index (Q-Score) is used for evaluation replicating the human vision model. Structural Similarity Index Measure (SSIM) [99–101] is used to compare block-wise correlations on the basis of structural similarity, luminance, and contrast information. Peak Signal-to-Noise Ratio (PSNR) [31] (in dB) is used for a pixel-to-pixel comparison for noise in the signals[1]. Learned Perceptual Image Patch Similarity (LPIPS) [115] is used to compare between the high-level features in the images, such as the semantic entities present in the scene, and evaluates our semantic and contextual knowledge-based contributions. This metric aligns with the perceptual judgement of humans.

5. Results

5.1. HDR Reconstruction

For an extensive comparison we selected a combination of methods that utilize different learning paradigms including, HDRCNN [22], DrTMO [23], ExpandNet [69], FHDR [47], SingleHDR [64], Two-stage [1], KUNet [96], HistoHDR-Net [6], HDR-GAN [77], ArtHDR-Net [5], Ghost-free HDR [65], SelfHDR [116], UPHDR-GAN [61], GlowGAN-ITM [94], SingleHDR(W) [55], and PSENet [76]. These state-of-the-art approaches include methods with single-exposed and multi-exposed LDR inputs, direct and indirect (*i.e.*, either directly reconstructing HDR images or generating multi-exposed LDR stacks), GAN-based, Diffusion-based, Transformer-based, unsupervised, weakly-supervised, self-supervised, unpaired learning, and knowledge-conditioned.

[1]Unless specified otherwise, for HDR evaluation PSNR is computed on tone-mapped original and reconstructed HDR image pairs.

Table 3. HDR reconstruction results. Comparison with unsupervised learning methods trained on mixed datasets.

Method	PSNR↑	SSIM↑	LPIPS↓	HDR-VDP-2↑
UPHDR-GAN [61]	42.51	0.980	0.043	66.13
GlowGAN-ITM [94]	33.45	0.910	0.061	62.13
LLM-HDR (Ours)	**43.96**	**0.989**	**0.020**	**69.32**

Paired Datasets. Tab. 2 summarizes the results of evaluation on HDRTV [16], NTIRE [79], HDR-Synth & HDR-Real [64] paired datasets, the most widely used datasets by the state-of-the-art (Sec. S3.1 for an extensive qualitative evaluation). In the top part of the table, we compare our method with 11 supervised learning approaches. The proposed method outperforms those approaches in terms of SSIM, LPIPS, and HDR-VDP-2. In terms of PSNR, it is second best after Ghost-free HDR. In the bottom part of the table, we compare our method with 5 unsupervised, weakly-supervised, and self-supervised learning methods. The proposed method outperforms those approaches on all evaluation metrics. It is worth noting that we used paired datasets in order to compare our method with the state-of-the-art, which predominantly requires paired samples for training. However, in all experiments, our method is trained in a self-supervised fashion with unpaired samples (Sec. 3).

Mixed Datasets. To further demonstrate the strength of our approach we extended the train set used in the previous experiment (paired data) with: 1) The HDR images from DrTMO [23], Kalantari [44], HDR-Eye [51], LDR-HDR Pair [40], and GTA-HDR [7] (1500 random samples), *i.e.*, additional 2854 HDR images; and 2) LDR images from GTA-HDR [7] (2000 random samples) and 50% of the LDR images in DrTMO, Kalantari, HDR-Eye, and LDR-HDR Pair, *i.e.*, additional 2677 LDR images. There is no overlap between LDR and HDR in this additional data, *i.e.*, it is an unpaired dataset. The final dataset used for training in this experiment has an approximately 87% overlap between LDR and HDR images. We kept the test set identical to the previous experiment, reported in Tab. 2, in order to compare with the state-of-the-art methods based paired data. Given that the only other methods that support this type of unpaired train data are UPHDR-GAN [61] and GlowGAN-ITM [94], we perform a direct comparison with these approaches in Tab. 3. The results demonstrate that our method outperforms all unpaired and paired data methods, setting new state-of-the-art on all metrics (Sec. S3.1 for more mixed data experiments).

Unpaired Datasets. We performed cross-dataset experiment where LDR and HDR for training are sourced from different datasets. We kept the test set identical to

Table 4. HDR reconstruction results. Comparison with unsupervised learning methods trained on unpaired datasets.

Method	PSNR↑	SSIM↑	LPIPS↓	HDR-VDP-2↑
UPHDR-GAN [61]	37.51	0.941	0.101	61.63
GlowGAN-ITM [94]	32.15	0.923	0.137	60.11
LLM-HDR (Ours)	**39.32**	**0.952**	**0.082**	**64.32**

Table 5. LDR reconstruction results. Comparison with the state-of-the-art tone-mapping operators and GlowGAN-prior [94].

Method	PSNR↑	SSIM↑	LPIPS↓
μ-law operator [43]	28.68	0.891	0.311
Reinhard's operator [84]	30.12	0.903	0.342
Photomatix [34]	**32.11**	0.922	0.136
GlowGAN-prior [94]	32.03	0.936	0.108
LLM-HDR (Ours)	31.81	**0.942**	**0.091**

the previous experiments. LDR images are sourced from HDRTV [16] and NTIRE [79], while the HDR are taken from HDR-Synth & HDR-Real [64]. The direct comparison with UPHDR-GAN [61] and GlowGAN-ITM [94] which have unpaired training capabilities is reported in Tab. 4. The results demonstrate that our method outperforms both unpaired data methods on all metrics.

5.2. LDR Reconstruction

We report results for LDR reconstruction in Tab. 5 and compare with the state-of-the-art tone-mapping operators μ-law [43], Reinhard's [84], and Photomatix [34]. We also consider the HDR-to-LDR prior from GlowGAN [94] which is used to re-project reconstructed HDR to LDR space for inverse learning. The test data is identical to the one used in the HDR reconstruction experiments. The results demonstrate that our method outperforms the other approaches in terms of SSIM and LPIPS. In terms of PSNR, it is second best after Photomatix (Sec. S3.2 for more LDR reconstruction experiments).

5.3. Ablation Results

Generators. We propose a novel generator that modifies the U-Net architecture by introducing an artifact-aware feedback mechanism (Sec. 3.1). The feedback helps in avoiding artifacts in image-to-image translation tasks where huge amounts of training data might make the process of hallucinating details difficult. To test this hypothesis, we first replaced the original U-Net used in the SingleHDR(W) [55] method with our U-Net generator. We see that this leads to improvements in all but one metrics as shown in Tab. 6. The table also demonstrates the improvement in our model when we use the feedback U-Net instead of the original U-Net of SingleHDR(W). Fig. 3 illustrates the improvement in SingleHDR(W) [55] when we use the proposed feedback U-Net instead of the original U-Net of

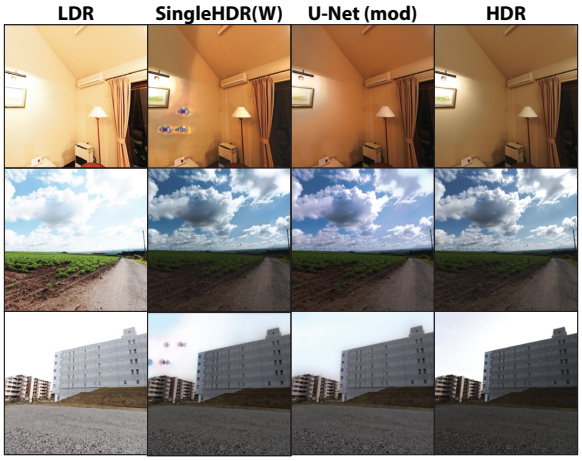


Figure 3. Comparison of the SingleHDR(W) [55] U-Net with and without our feedback mechanism on images from the DrTMO [23].

SingleHDR(W). The original U-Net produces many artifacts in the output HDR images whereas our modified version with feedback reconstructs artifact-free HDR images. Sec. S3.3 provides extensive ablation experiments for the proposed architecture and loss functions.

6. Conclusion

This paper proposes **LLM-HDR**, the first LLM-driven semantic and cycle consistency guided self-supervised learning approach for unpaired {LDR,HDR} data which addresses both the inverse tone-mapping (*i.e.*, LDR $\rightarrow$ HDR) and tone-mapping (*i.e.*, HDR $\rightarrow$ LDR) tasks. The proposed method utilizes novel generators based on modified U-Net architecture incorporating ConvLSTM-based artifact-aware feedback mechanism and exposure-aware skip connections to mitigate visual artifacts, CLIP embedding encoder for contrastive learning to minimize the semantic difference between LDR and reconstructed HDR pairs, and a novel loss function based on the Mean Intersection over Union metric to further ensure semantic consistency between the LDR and reconstructed HDR. It also utilizes LLM-based loss and artifact- and exposure-aware saliency maps to bring more realism and naturalness in the reconstructed HDR images. The thorough experimental validation demonstrates the contributions of the proposed method that achieves state-of-the-art results across several benchmark datasets and reconstructs high-quality HDR and LDR images (Sec. S4 for future work).

References

[1] SM A Sharif, Rizwan Ali Naqvi, Mithun Biswas, and Sungjun Kim. A two-stage deep network for high dynamic range image reconstruction. In *Proceedings of the*

Table 6. Quantitative evaluation of our feedback U-Net-based generators for HDR reconstruction. For evaluation we use the dataset employed by SingleHDR(W), *i.e.*, DrTMO [23]. The models marked with "(mod)" use feedback U-Net-based generators and the rest use the same U-Net generator as SingleHDR(W).

Method	PSNR↑	SSIM↑	LPIPS↓	HDR-VDP-2↑
SingleHDR(W) [55]	30.79	0.952	0.055	62.95
SingleHDR(W) (mod)	30.03	0.961	0.036	63.91
LLM-HDR	35.11	0.969	0.029	67.88
LLM-HDR (mod)	**43.01**	**0.988**	**0.023**	**68.21**

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 550–559, 2021. 7

[2] Mrinal Anand, Nidhin Harilal, Chandan Kumar, and Shanmuganathan Raman. HDRVideo-GAN: Deep Generative HDR Video Reconstruction. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2021. 3

[3] Alessandro Artusi, Rafał K Mantiuk, Thomas Richter, Philippe Hanhart, Pavel Korshunov, Massimiliano Agostinelli, Arkady Ten, and Touradj Ebrahimi. Overview and evaluation of the jpeg xt hdr image compression standard. *Journal of Real-Time Image Processing*, 16: 413–428, 2019. 1

[4] Francesco Banterle, Alessandro Artusi, Alejandro Moreo, and Fabio Carrara. Nor-Vdpnet: A No-Reference High Dynamic Range Quality Metric Trained On Hdr-Vdp 2. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 126–130. IEEE, 2020. 7

[5] 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)*, pages 806–812. IEEE, 2023. 2, 7

[6] 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)*, pages 2730–2736. IEEE, 2024. 2, 7

[7] Hrishav Bakul Barua, Kalin Stefanov, KokSheik Wong, Abhinav Dhall, and Ganesh Krishnasamy. Gta-hdr: A large-scale synthetic dataset for hdr image reconstruction. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 7865–7875, 2025. 2, 7, 5

[8] Mojtaba Bemana, Thomas Leimkühler, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. Exposure diffusion: Hdr image generation by consistent ldr denoising. *arXiv preprint arXiv:2405.14304*, 2024. 2

[9] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 6

[10] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 5, 6, 4

[11] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a Deep Single Image Contrast Enhancer from Multi-Exposure Images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018. 2

[12] Gaofeng Cao, Fei Zhou, Kanglin Liu, Anjie Wang, and Leidong Fan. A Decoupled Kernel Prediction Network Guided by Soft Mask for Single Image HDR Reconstruction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s):1–23, 2023. 2

[13] Peibei Cao, Haoyu Chen, Jingzhe Ma, Yu-Chieh Yuan, Zhiyong Xie, Xin Xie, Haiqing Bai, and Kede Ma. Learned hdr image compression for perceptually optimal storage and display. *arXiv preprint arXiv:2407.13179*, 2024. 1

[14] 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*, pages 22433–22443, 2024. 1

[15] Junuk Cha, Mengwei Ren, Krishna Kumar Singh, He Zhang, Yannick Hold-Geoffroy, Seunghyun Yoon, HyunJoon Jung, Jae Shin Yoon, and Seungryul Baek. Text2relight: Creative portrait relighting with text guidance. *arXiv preprint arXiv:2412.13734*, 2024. 2

[16] Xiangyu Chen, Zhengwen Zhang, Jimmy S Ren, Lynhoo Tian, Yu Qiao, and Chao Dong. A New Journey From SDRTV to HDRTV. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4500–4509, 2021. 6, 7, 8, 5

[17] Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. Mocycle-gan: Unpaired video-to-video translation. In *Proceedings of the 27th ACM international conference on multimedia*, pages 647–655, 2019. 2, 1

[18] Dinu Coltuc, Philippe Bolon, and J-M Chassery. Exact histogram specification. *IEEE Transactions on Image processing*, 15(5):1143–1152, 2006. 7

[19] Dwip Dalal, Gautam Vashishtha, Prajwal Singh, and Shanmuganathan Raman. Single Image LDR to HDR Conversion Using Conditional Diffusion. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3533–3537. IEEE, 2023. 2

[20] Khursheed Ahmad Dar and Sumit Mittal. An enhanced adaptive histogram equalization based local contrast preserving technique for hdr images. In *IOP Conference Series: Materials Science and Engineering*, page 012119. IOP Publishing, 2021. 1

[21] Frédéric Drago, Karol Myszkowski, Thomas Annen, and Norishige Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. In *Computer graphics forum*, pages 419–426. Wiley Online Library, 2003. 1

[22] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017. 2, 7

[23] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6):177–1, 2017. 2, 6, 7, 8, 9, 4, 5

[24] Andrew S Glassner. *An introduction to ray tracing*. Morgan Kaufmann, 1989. 7

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville,

and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 5

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[27] Abhishek Goswami, Erwan Bernard, Wolf Hauser, Frederic Dufaux, and Rafal Mantiuk. G-semtmo: Tone mapping with a trainable semantic graph. *arXiv preprint arXiv:2208.14113*, 2022. 2

[28] Abhishek Goswami, Aru Ranjan Singh, Francesco Banterle, Kurt Debattista, and Thomas Bashford-Rogers. Semantic aware diffusion inverse tone mapping. *arXiv preprint arXiv:2405.15468*, 2024. 2, 3

[29] B-C Guo and C-H Lin. Single-image hdr reconstruction based on two-stage gan structure. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 91–95. IEEE, 2023. 2, 3

[30] Cheng Guo and Xiuhua Jiang. Lhdr: Hdr reconstruction for legacy content using a lightweight dnn. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 3155–3171, 2022. 2

[31] Prateek Gupta, Priyanka Srivastava, Satyam Bhardwaj, and Vikrant Bhateja. A modified psnr metric based on hvs for quality assessment of color images. In *2011 International Conference on Communication and Industrial Application*, pages 1–4. IEEE, 2011. 7

[32] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 6

[33] Xueyu Han, Ishtiaq Rasool Khan, and Susanto Rahardja. High Dynamic Range Image Tone Mapping: Literature review and performance benchmark. *Digital Signal Processing*, page 104015, 2023. 1

[34] HDRsoft. Photomatix. https://www.hdrsoft.com/. [Online; accessed 3-Nov-2023]. 2, 8, 5

[35] 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*, pages 2890–2898, 2022. 2

[36] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2, 4

[37] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. HDR-NeRF: High Dynamic Range Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18398–18408, 2022. 2

[38] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 4

[39] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 4

[40] Hanbyol Jang, Kihun Bang, Jinseong Jang, and Dosik Hwang. Dynamic Range Expansion Using Cumulative Histogram Learning for High Dynamic Range Image Generation. *IEEE Access*, 8:38554–38567, 2020. 6, 7, 1, 5, 8

[41] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4324–4333, 2024. 2

[42] Takao Jinno and Masahiro Okuda. Multiple exposure fusion for high dynamic range image acquisition. *IEEE Transactions on image processing*, 21(1):358–365, 2011. 1

[43] 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*, 94(3): 972–978, 2011. 5, 8

[44] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep High Dynamic Range Imaging of Dynamic Scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017. 6, 7, 5

[45] Agastya Kalra, Guy Stoppi, Dmitrii Marin, Vage Taamazyan, Aarrushi Shandilya, Rishav Agarwal, Anton Boykov, Tze Hao Chong, and Michael Stark. Towards co-evaluation of cameras hdr and algorithms for industrial-grade 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22691–22701, 2024. 1

[46] Suthum Keerativittayanun, Toshiaki Kondo, Kazunori Kotani, and Teera Phatrapornnant. An innovative of pyramid-based fusion for generating the hdr images in common display devices. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pages 53–56. IEEE, 2015. 1

[47] Zeeshan Khan, Mukul Khanna, and Shanmuganathan Raman. FHDR: HDR Image Reconstruction from a Single LDR Image using Feedback Network. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE, 2019. 5, 7

[48] Kyungman Kim, Jonghyun Bae, and Jaeseok Kim. Natural hdr image tone mapping based on retinex. *IEEE Transactions on Consumer Electronics*, 57(4):1807–1814, 2011. 1

[49] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[50] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5, 1, 2, 4

[51] Pavel Korshunov, Hiromi Nemoto, Athanassios Skodras, and Touradj Ebrahimi. Crowdsourcing-based evaluation of

privacy in hdr images. In *Optics, photonics, and digital technologies for multimedia applications III*, page 913802. SPIE, 2014. 6, 7, 5

[52] Jung Gap Kuk, Nam Ik Cho, and Sang Uk Lee. High dynamic range (hdr) imaging by gradient domain fusion. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1461–1464. IEEE, 2011. 1

[53] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (ToG)*, 39(6):1–14, 2020. 7

[54] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977. 3

[55] Phuoc-Hieu Le, Quynh Le, Rang Nguyen, and Binh-Son Hua. Single-Image HDR Reconstruction by Multi-Exposure Generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4063–4072, 2023. 2, 3, 7, 8, 9, 5

[56] Siyeong Lee, So Yeon Jo, Gwon Hwan An, and Suk-Ju Kang. Learning to generate multi-exposure stacks with cycle consistency for high dynamic range imaging. *IEEE Transactions on Multimedia*, 23:2561–2574, 2020. 2

[57] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[58] Fangya Li, Ruipeng Gang, Chenghua Li, Jinjing Li, Sai Ma, Chenming Liu, and Yizhen Cao. Gamma-enhanced spatial attention network for efficient high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1032–1040, 2022. 6

[59] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 1

[60] Jinghui Li and Peiyu Fang. HDRNET: Single-Image-based HDR Reconstruction Using Channel Attention CNN. In *Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing*, pages 119–124, 2019. 2

[61] Ru Li, Chuan Wang, Jue Wang, Guanghui Liu, Heng-Yu Zhang, Bing Zeng, and Shuaicheng Liu. Uphdr-gan: Generative adversarial network for high dynamic range imaging with unpaired data. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7532–7546, 2022. 2, 3, 4, 7, 8, 5

[62] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. 1

[63] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback Network for Image Super-Resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3867–3876, 2019. 2, 4

[64] 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*, pages 1651–1660, 2020. 2, 6, 7, 8, 5, 9

[65] 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*, pages 344–360. Springer, 2022. 2, 3, 7

[66] Zhan Lu, Qian Zheng, Boxin Shi, and Xudong Jiang. Panonerf: Synthesizing high dynamic range novel views with geometry from sparse low dynamic range panoramic images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3927–3935, 2024. 2

[67] Gonzalo Luzardo, Jan Aelterman, Hiep Luong, Sven Rousseaux, Daniel Ochoa, and Wilfried Philips. Fully-automatic inverse tone mapping algorithm based on dynamic mid-level tone mapping. *APSIPA Transactions on Signal and Information Processing*, 9:e7, 2020. 2

[68] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011. 7

[69] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, pages 37–49. Wiley Online Library, 2018. 7

[70] John J McCann and Alessandro Rizzi. *The art and science of HDR imaging*. John Wiley & Sons, 2011. 1

[71] Laurence Meylan and Sabine Susstrunk. High dynamic range image rendering with a retinex-based adaptive filter. *IEEE Transactions on image processing*, 15(9):2820–2830, 2006. 1

[72] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[73] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022. 2, 7

[74] 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. 3

[75] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023. 2, 1

[76] Hue Nguyen, Diep Tran, Khoi Nguyen, and Rang Nguyen. Psenet: Progressive self-enhancement network for unsuper-

vised extreme-light image enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1756–1765, 2023. 1, 2, 3, 7

[77] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. HDR-GAN: HDR Image Reconstruction from Multi-Exposed LDR Images with Large Motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021. 2, 3, 7

[78] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 2

[79] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, and Radu Timofte. Ntire 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 691–700, 2021. 6, 7, 8, 5

[80] Motong Qiao and Michael K Ng. Tone mapping for high-dynamic-range images using localized gamma correction. *Journal of electronic imaging*, 24(1):013010–013010, 2015. 1

[81] 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*, pages 8748–8763. PMLR, 2021. 2, 4, 1

[82] Prarabdh Raipurkar, Rohil Pal, and Shanmuganathan Raman. HDR-cGAN: Single LDR to HDR Image Translation using Conditional GAN. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2021. 2, 3

[83] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[84] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. *Photographic Tone Reproduction for Digital Images*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. 8, 2, 5

[85] Haoyu Ren, Yi Fan, and Stephen Huang. Robust real-world image enhancement based on multi-exposure ldr images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1715–1723, 2023. 2

[86] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2, 4

[87] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single Image HDR Reconstruction Using a CNN with Masked Features and Perceptual Loss. *ACM Transactions on Graphics (TOG)*, 39(4):80–1, 2020. 2

[88] Pinar Satilmis and Thomas Bashford-Rogers. Deep Dynamic Cloud Lighting. *arXiv preprint arXiv:2304.09317*, 2023. 2

[89] Shuaikang Shang, Xuejing Kang, and Anlong Ming. Hdr-transdc: High dynamic range image reconstruction with transformer deformation convolution. *arXiv preprint arXiv:2403.06831*, 2024. 2

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

[91] Shreyas Singh, Aryan Garg, and Kaushik Mitra. Hdr-splat: Gaussian splatting for high dynamic range 3d scene reconstruction from raw images. *arXiv preprint arXiv:2407.16503*, 2024. 2

[92] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 2

[93] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 5

[94] 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*, pages 10509–10519, 2023. 2, 3, 7, 8, 5

[95] Chao Wang, Zhihao Xia, Thomas Leimkühler, Karol Myszkowski, and Xuaner Zhang. Lediff: Latent exposure diffusion for hdr generation. *arXiv preprint arXiv:2412.14456*, 2024. 2

[96] Hu Wang, Mao Ye, Xiatian Zhu, Shuai Li, Ce Zhu, and Xue Li. KUNet: Imaging Knowledge-Inspired Single HDR Image Reconstruction. In *The 31st International Joint Conference On Artificial Intelligence (IJCAI/ECAI 22)*, 2022. 2, 3, 7

[97] Junchi Wang and Lei Ke. Llm-seg: Bridging image segmentation and large language model reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1765–1774, 2024. 2

[98] 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*, 44 (12):8874–8895, 2021. 1

[99] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009. 7

[100] 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*, 13(4):600–612, 2004.

[101] Zhou Wang, Ligang Lu, and Alan C Bovik. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication*, 19(2):121–132, 2004. 7

[102] Kangle Wu, Jun Huang, Yong Ma, Fan Fan, and Jiayi Ma. Cycle-retinex: Unpaired low-light image enhancement via

retinex-inline cyclegan. *IEEE Transactions on Multimedia*, 26:1213–1228, 2023. 3

[103] 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*, 5(4):5113–5119, 2020. 1

[104] Jun Xiao, Qian Ye, Tianshan Liu, Cong Zhang, and Kin-Man Lam. Deep progressive feature aggregation network for multi-frame high dynamic range imaging. *Neurocomputing*, 594:127804, 2024. 2

[105] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021. 2, 3, 1

[106] Qingsen Yan, Bo Wang, Lei Zhang, Jingyu Zhang, Zheng You, Qinfeng Shi, and Yanning Zhang. Towards accurate hdr imaging with learning generator constraints. *Neurocomputing*, 428:79–91, 2021. 1

[107] Qingsen Yan, Weiye Chen, Song Zhang, Yu Zhu, Jinqiu Sun, and Yanning Zhang. A unified hdr imaging method with pixel and patch level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22211–22220, 2023. 2

[108] Hao Yang, Liyuan Pan, Yan Yang, and Wei Liang. Language-driven all-in-one adverse weather removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24902–24912, 2024. 2

[109] Shaoliang Yang, Dongming Zhou, Jinde Cao, and Yanbu Guo. Lightingnet: An integrated learning method for low-light image enhancement. *IEEE Transactions on Computational Imaging*, 9:29–42, 2023. 1

[110] W Yao, ZG Li, and S Rahardja. Intensity mapping function based weighted frame averaging for high dynamic range imaging. In *2011 6th IEEE Conference on Industrial Electronics and Applications*, pages 1574–1577. IEEE, 2011. 1

[111] Qian Ye, Jun Xiao, Kin-man Lam, and Takayuki Okatani. Progressive and selective fusion network for high dynamic range imaging. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5290–5297, 2021. 2

[112] Phyo Thet Yee, Sudeepta Mishra, and Abhinav Dhall. Clipswap: Towards high fidelity face swapping via attributes and clip-informed loss. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024. 1

[113] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 4

[114] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024. 2, 3, 1

[115] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[116] 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. 1, 2, 3, 7

[117] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 3, 5, 6, 1

[118] 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)*, pages 12334–12344, 2023. 2

LLM-HDR: Bridging LLM-based Perception and Self-Supervision for Unpaired LDR-to-HDR Image Reconstruction

Supplementary Material

S1. Related Work

This section complements Sec. 2 in the main paper and includes an overview of non-learning approaches for HDR reconstruction. The most popular approach in this category consists of multi-exposed LDR fusion to achieve high dynamic range in the output image [42]. This approach involves image feature alignment, calculating weights for feature mapping on the basis of image characteristics, and fusing the images on the basis of the weights to get the most appropriate exposures for each part of the image. Another approach is using histogram equalization [20] that allows for contrast and brightness levels re-adjustment in the over/underexposed regions of the image. Gradient domain manipulation [52] techniques enhance the granular details of an LDR image and expand its dynamic range. Pyramid-based image fusion [46] is a technique where image features are extracted into Laplacian pyramids [46] and then fused for the HDR. Some methods use Retinex-based [48, 71] tuning to effectively produce and display HDR images on consumer screens [48, 71]. Intensity mapping function-based approaches [21, 80, 110] use transformations on image pixels either by using Logarithmic mapping or Gamma correction to stretch the intensity range or enhance lower intensity pixels. The methods in this category, although time-efficient, face some issues including ghosting effects, halo artifacts, and blurring in the reconstruction, and they do not generalize well for variety of input LDR images.

S2. Method

This section complements Sec. 3 in the main paper. It provides a concise background on the key concepts employed in our method, *i.e.*, cycle consistency [117], Large Language Models [75, 105], contrastive language-image pretraining [62, 81, 112], and image semantic segmentation [50]. The section also offers implementation details for the architecture modules.

S2.1. Background

Cycle Consistency. The concept of cycle consistency in image-to-image [117] and video-to-video [17] translation tasks is a powerful and elegant design for unpaired data. Previously, cycle consistency has been applied to paired LDR $\leftrightarrow$ HDR translation [106], where the authors used 3 LDR images with different exposures as input to reconstruct an HDR. The reconstructed HDR is then used by 3 different generators to reconstruct the original 3 multi-exposed LDR images. However, the focus of our work is on the more general task of unpaired LDR $\leftrightarrow$ HDR translation. Given the images $\{x_i\}_{i=1}^{N}$ where $x_i \in X$ and $\{y_j\}_{j=1}^{M}$ where $y_j \in Y$, the goal is to design two generators, *i.e.*, G_Y and G_X such that, $G_Y(x) \rightarrow y$ and $G_X(y) \rightarrow x$. Fig. S1 depicts this concept for the LDR-to-HDR translation task.

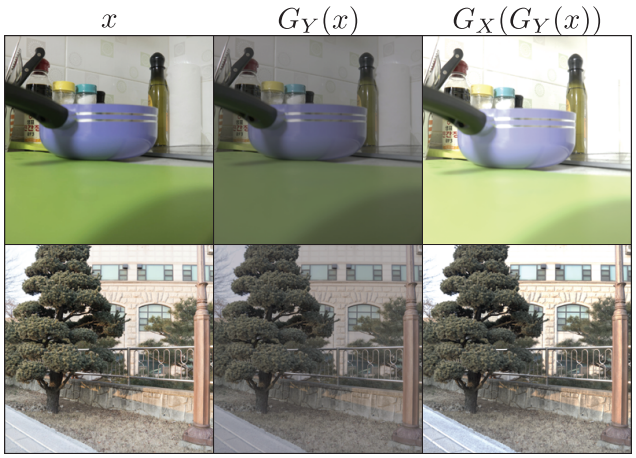


Figure S1. G_Y translates LDR to HDR images and G_X HDR to LDR. The LDR images are from the LDR-HDR pair [40] dataset.

LLM and VLM. Large Language Models [75] and Vision-Language Models [105, 114] are trained on vast amounts of textual and/or visual data (*i.e.*, images and videos) and achieve excellent zero-shot performance. We leverage the text and visual perception capabilities of an LLM to evaluate the reconstruction of our method and refine it with better awareness of artifacts and over/underexposed regions.

Contrastive Language-Image Pretraining. CLIP [81] is a method that finds the closeness between visual and textual data in the embedding space. The method is trained on image-text pairs to associate images with their descriptions. The CLIP loss maximizes the cosine similarity between correct (positive) pairs and minimize the similarity between incorrect (negative) pairs. We extract the CLIP embeddings from both the LDR and reconstructed HDR image pairs and maximize their cosine similarity in the embedding space for preservation of semantic information across domains. We also combine the CLIP embeddings extracted from the LDR and reconstructed HDR to guide the decoder of the generators in the next step of the reconstruction.

Semantic Segmentation. Semantic segmentation [50, 59, 83] is a vision problem where each pixel of an image or video frame is segmented into meaningful sub-parts such

as objects, roads, water bodies, buildings, humans, and animals. Segment Anything (SAM) [50] is a state-of-the-art method for semantic segmentation of images. Segmented objects in a scene not only inform the number and location of semantically meaningful concepts, but also provide a holistic semantic description of the scene. Therefore, we can use that information to improve image reconstruction. We introduce a loss on the basis of the Mean Intersection over Union metric to preserve the image integrity (based on the segmented objects) between the two domain X and Y.

S2.2. Architecture Modules

Generators. The U-Net of each generator consists of 7 levels. Each level consists of 2 `Conv` layers with 3×3 kernel followed by a `ReLU` activation and a `BN` layer. The input images of size 512×512 are first converted into 32-dimensional features. The features are doubled in each level of the encoder until reaching size of 512. The decoder levels then up-sample the data by performing channel-wise concatenation with the extracted feature from previous levels of the encoder. The intuition behind using feedback mechanism is to refine the decoder layers of the generator based on the input from encoder layers and the feedback mechanism itself. The output of each iteration of the feedback mechanism is not only fed into the first level of the decoder block, but also sent back through the hidden state to guide its next input. Feedback mechanisms perform better than simple feed-forward networks with less trainable parameters [63]. The feedback network consists of 3 dilated dense blocks each with dilation rate of 3. The dilated blocks are made of 3×3 `Conv` layers followed by `ReLU` activation function. Each dilated block contains two 1×1 feature compression layers at the start and end. Dilation enhances the network's receptive field which helps to capture broader context from the input features without increasing the computational complexity. Finally, the feedback starts with a 1×1 `Conv` and ends with 3×3 `Conv` for compression and final processing, respectively. Fig. S2 depicts and overview of the proposed generators architecture.

Discriminators. The discriminators consist of 5 `Conv` layers. We input the reconstructed LDR/HDR along with a real (unpaired) LDR/HDR in the respective discriminator. The `Conv` layers have 4×4 filters and stride of 2 for the first 3 layers and 1 henceforth. The output of the first layer has features of size 64, which is doubled in each of the next layers until 512. The output of the fifth layer is of size 1. The first four layers have `LeakyReLU` activation with the final layer using a `Sigmoid` activation for outputting the probabilities of a particular image being real or fake. The training process is stabilized using `BN` following all the `Conv` layers except the first one.

LLM. The artifact-aware saliency map LLM_{af} is extracted as feature of size 512 with 1×1 `Conv` and we perform an element-wise multiplication with the output features of the bottleneck layer (*i.e.*, layer 4) of the U-Net generators G_Y and G_X which goes into the feedback mechanism. The exposure-aware saliency map LLM_{ox} (for overexposure) (which goes into the generator G_Y only) is extracted as feature of size 256 with 1×1 `Conv` and fused with the output features of layer 3 henceforth concatenated with the encoder embeddings which goes into the bottleneck layer. Finally, LLM_{ux} (for underexposure) is extracted as feature of sizes 64, 128, and 256, with 1×1 `Conv` and fused with the skip connections in each level of the U-Net generator G_Y (*i.e.*, level 1, 2, and 3). Fig. S3 depicts the overview of the proposed LLM module.

Encoders. The encoders extract embeddings of size 512 from the input and output domain images. Given the input size of the generators, we re-project these embeddings to a size of 256 using 1×1 `Conv`. Then we add the embeddings and concatenate them with the input features of the bottleneck layer (*i.e.*, layer 4) of our feedback-based U-Net generators.

S2.3. Loss Functions

Figs. S4 to S7 illustrate the formulation of the LLM-based loss, contrastive, semantic segmentation, and cycle consistency functions, respectively.

S2.4. Implementation Details

The proposed method is implemented in PyTorch [78] on Ubuntu 20.04.6 LTS workstation with Nvidia Quadro P5000 GPU with 16GB memory, Intel® Xeon® W-2145 CPU at 3.70GHz with 16 CPU cores, 64GB RAM, and 2.5TB SSD. We used Adam optimizer [49] to train the models for 170 epochs. Batch size of 1 is expected to work well in cycle consistency models [117] however, since we are using contrastive loss on positive and negative pairs, the batch size is set to 4. We used learning rate of 4×10^{-4} for the generators and 2×10^{-4} for the discriminators. The learning rate is kept constant for the first 100 epochs and subsequently set to decay linearly to 0. For all models, we resized the input images to 512×512. The HDR images used in the text are tone-mapped with Reinhard's operator [84].

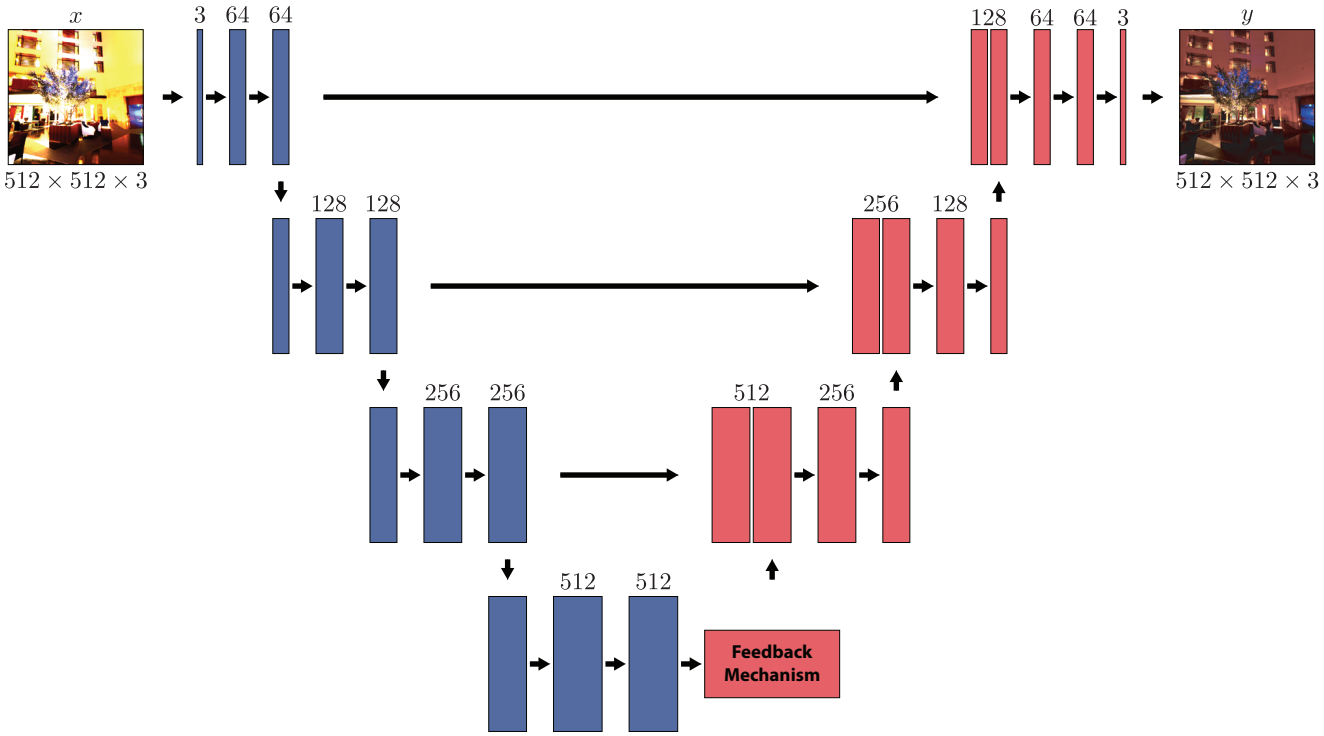


Figure S2. Overview of the proposed generators based on our novel feedback based U-Net architecture. Left part (blue) is the encoder and right part (red) is the decoder. The decoder is inside the feedback iteration loop.

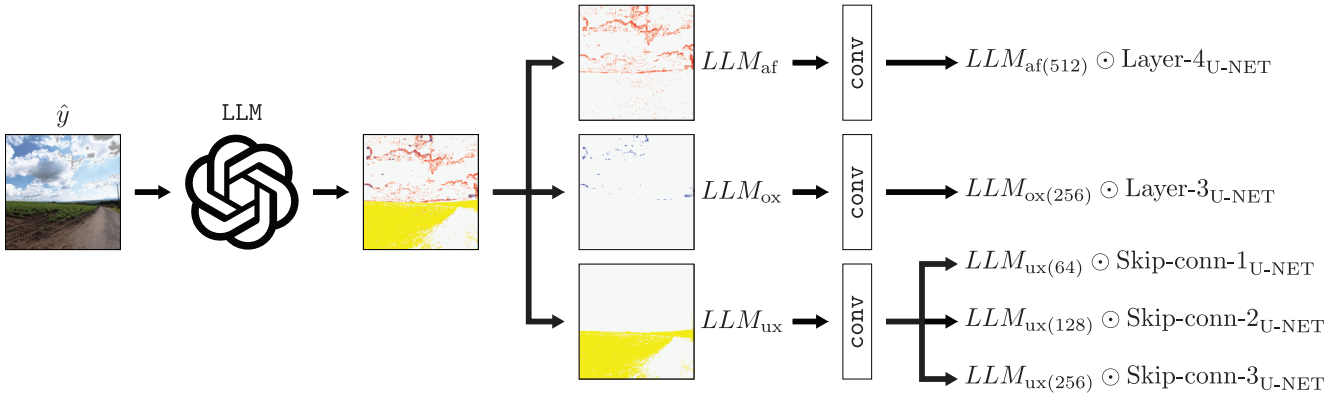


Figure S3. Overview of the proposed LLM module. Q&A sessions include: **User prompt:** "Tell me if there are any synthesis artifacts in the given scene or not. Must response with 1) Yes or No only, 2) If Yes, can you get me the saliency maps of the artifacts (in red), overexposed (in blue) and underexposed (in yellow) areas of this image?", **System:** "Outputs the saliency maps for the detected artifacts in red, overexposed pixels in blue, and underexposed pixels in yellow."

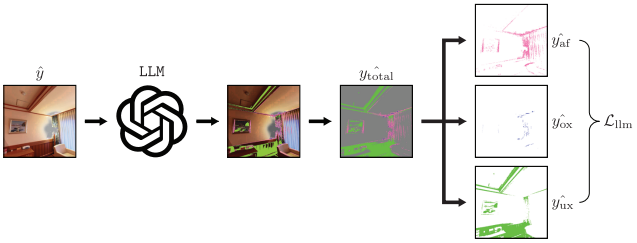


Figure S4. Depiction of the LLM-based loss $\mathcal{L}_{\text{llm}}$. Q&A sessions include: **User prompt:** "Tell me if there are any synthesis artifacts in the given scene or not. Must response with 1) Yes or No only, 2) If Yes, return the [artifact areas number of pixels, total number of pixels]. Also highlight the artifact areas with pink. Similarly find if there are any over/under exposure areas in the scene and return the [pixels in overexposure, pixels in underexposure areas] and highlight the overexposed areas in blue and underexposed areas in green.", **System:** "Outputs the number of pixels with the pixel list (in .txt files) for the detected artifacts in pink, overexposed pixels in blue, and underexposed pixels in green."

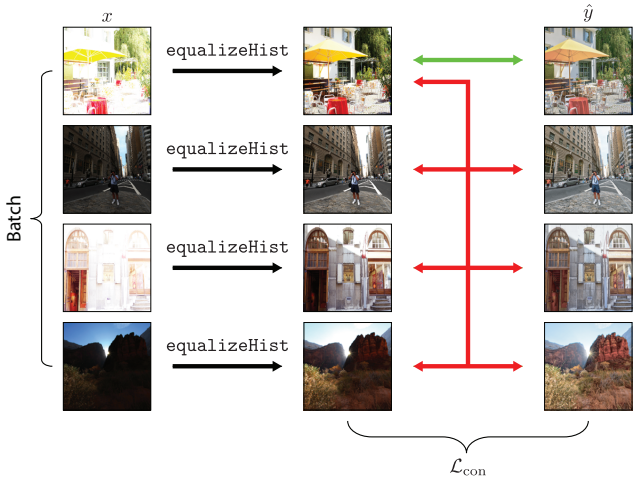


Figure S5. Depiction of the contrastive loss $\mathcal{L}_{\text{con}}$. Positive (green) and negative (red) pairs in a batch. We use a histogram-equalized version of the LDR processed using the OpenCV function `equalizeHist` [10].

Figure S6. Depiction of the semantic segmentation loss $\mathcal{L}_{\text{sem}}$. We use Segment Anything (SAM) [50] to generate segmentation classes in the histogram-equalized LDR and reconstructed tone-mapped HDR images.

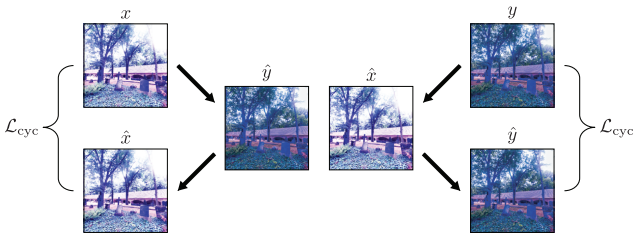


Figure S7. Depiction of the cycle consistency loss $\mathcal{L}_{\text{cyc}}$ using an image from the DrTMO [23] dataset.

Table S1. HDR reconstruction results. Comparison with unsupervised learning methods trained on mixed datasets.

Method	PSNR↑	SSIM↑	LPIPS↓	HDR-VDP-2↑
UPHDR-GAN [61]	41.98	0.975	0.046	65.54
GlowGAN-ITM [94]	32.18	0.905	0.066	61.89
LLM-HDR (Ours)	**42.01**	**0.988**	**0.023**	**68.81**

S3. Results

This section complements Sec. 5 in the main paper. It provides additional quantitative evaluation of our method for both HDR and LDR reconstruction. The section also includes an extensive qualitative and ablation results.

S3.1. HDR Reconstruction

Tab. S1 provides the results from an additional mixed data experiment. The train set used in this experiment is: 1) The HDR images from DrTMO [23], Kalantari [44], HDR-Eye [51], LDR-HDR Pair [40], and GTA-HDR [7] (1500 random samples), *i.e.*, 2854 HDR images; and 2) LDR images from GTA-HDR [7] (2000 random samples) and 50% of the LDR images in DrTMO, Kalantari, HDR-Eye, and LDR-HDR Pair, *i.e.*, 2677 LDR images. In this dataset, a total of 1177 {LDR,HDR} pairs overlap. Given that the only other methods that support this type of unpaired train data are UPHDR-GAN [61] and GlowGAN-ITM [94], we perform a direct comparison with these two approaches. The results demonstrate that our method again outperforms UPHDR-GAN and GlowGAN-ITM on all metrics.

Figs. S12 to S18 illustrate the quality of the HDR images reconstructed with our method. Our method mitigates artifacts such that the resulting HDR closely resembles the ground truth HDR. The finer textures, color hues, and shades are well preserved compared to the other methods. The granular level details in the bright and dark regions are also well reconstructed.

S3.2. LDR Reconstruction

Tab. S2 provides the results from an additional LDR reconstruction experiment. In this case, our method is trained with the data specified in the previous section. Similar to the LDR reconstruction experiment in the main paper, we compare the model with the state-of-the-art tone-mapping operators μ-law [43], Reinhard's [84], Photomatix [34], and GlowGAN-prior [94]. The results again demonstrate that our method outperforms the other approaches in terms of SSIM and LPIPS. In terms of PSNR, it is third best after Photomatix and GlowGAN-prior.

S3.3. Ablation Results

Architecture. Our method uses the cycle consistency concept of CycleGAN and introduces many new components to address the LDR $\leftrightarrow$ HDR translation task. The most significant components are the LLM module, artifact-aware feedback and exposure-aware skip connection-based U-Net generators, and the CLIP embedding encoders (Sec. 3.1). We perform a thorough ablation on each of the components to highlight the contributions. Tab. S3 summarizes the results where we add each component to the architecture as we go down. The first row lists the results from the original CycleGAN method. The second row is the CycleGAN utilizing the U-Net generator from SingleHDR(W). The third row is the CycleGAN utilizing our feedback U-Net generator. This results in a significant improvement in SSIM, LPIPS, and HDR-VDP-2 (Q-Score). The fourth row shows the results when we add the LLM module to the architecture which shows some improvement in the PSNR, LPIPS, and Q-Score. The fifth row lists the results when we remove the LLM module and add the CLIP embedding encoders to the architecture. This results in an improvement in PSNR, SSIM, and Q-Score. The sixth row depicts good improvement in almost all the metrics as we add both CLIP encoders and LLM module to the architecture. The seventh, eight, and ninth rows depict the improvements after adding

Table S2. LDR reconstruction results. Comparison with the state-of-the-art tone-mapping operators and GlowGAN-prior [94].

Method	PSNR↑	SSIM↑	LPIPS↓
μ-law operator [43]	28.68	0.891	0.311
Reinhard's operator [84]	30.12	0.903	0.342
Photomatix [34]	**32.11**	0.922	0.136
GlowGAN-prior [94]	31.91	0.931	0.112
LLM-HDR (Ours)	31.11	**0.939**	**0.095**

Table S3. Architecture ablation results with the HDRTV [16], NTIRE [79] and HDR-Synth & HDR-Real [64] datasets. **U^o:** The U-Net generator used in SingleHDR(W) [55], **U^f:** The proposed U-Net generator, **E:** CLIP embedding encoder, **LLM:** LLM module for artifact-aware and exposure-aware maps.

Variant	PSNR↑	SSIM↑	LPIPS↓	Q-Score↑
CycleGAN [117]	21.34	0.849	0.351	61.55
U^o	25.13	0.851	0.341	61.79
U^f	24.07	0.879	0.126	63.16
U^f+LLM	26.12	0.880	0.121	64.32
U^f+E	29.19	0.885	0.122	64.21
U^f+E+LLM	30.15	0.887	0.101	64.41
U^f+E+LLM+$\mathcal{L}_{con}$	36.45	0.911	0.047	66.56
U^f+E+LLM+$\mathcal{L}_{sem}$	34.21	0.897	0.094	66.12
U^f+E+LLM+$\mathcal{L}_{llm}$	36.01	0.907	0.046	66.72
U^f+E+LLM+$\mathcal{L}_{con}$+$\mathcal{L}_{sem}$	40.01	0.979	0.021	68.41
U^o+E+LLM+$\mathcal{L}_{con}$+$\mathcal{L}_{sem}$+$\mathcal{L}_{llm}$	35.15	0.953	0.042	67.18
U^f+E+LLM+$\mathcal{L}_{con}$+$\mathcal{L}_{sem}$+$\mathcal{L}_{llm}$	**40.11**	**0.979**	**0.020**	**68.51**

Table S4. Loss functions ablation results with the HDRTV [16], NTIRE [79] and HDR-Synth & HDR-Real [64] datasets.

Loss	PSNR↑	SSIM↑	LPIPS↓	HDR-VDP-2↑
$\mathcal{L}_{\text{GAN}}$+$\mathcal{L}_{\text{cyc}}$	24.55	0.878	0.125	62.11
+$\mathcal{L}_{\text{id}}$	30.15	0.887	0.101	64.41
+$\mathcal{L}_{\text{id}}$+$\mathcal{L}_{\text{con}}$	36.45	0.911	0.047	66.56
+$\mathcal{L}_{\text{id}}$+$\mathcal{L}_{\text{sem}}$	34.21	0.897	0.094	66.12
+$\mathcal{L}_{\text{id}}$+$\mathcal{L}_{\text{llm}}$	36.01	0.907	0.046	66.72
+$\mathcal{L}_{\text{id}}$+$\mathcal{L}_{\text{con}}$+$\mathcal{L}_{\text{sem}}$	40.01	0.979	0.021	68.41
+$\mathcal{L}_{\text{id}}$+$\mathcal{L}_{\text{con}}$+$\mathcal{L}_{\text{sem}}$+$\mathcal{L}_{\text{llm}}$	**40.11**	**0.979**	**0.020**	**68.51**

contrastive, semantic segmentation, and LLM-based objectives (Sec. 3.2) in the architecture, respectively. The tenth row shows the improvements when we add only the contrastive and semantic segmentation objectives in the architecture. The eleventh row depicts the results with all components in our architecture but with the original U-Net from SingleHDR(W). The last row is the full architecture with all components which achieves the best performance on all metrics.

Loss Functions. We also study the loss functions and their contribution towards the overall results keeping the architectural components constant. Tab. S4 summarizes the results. The first row is the results from the standard adversarial and cycle consistency loss. In the second row we add the identity loss which improves all metrics. In the third, fourth, and fifth rows we add the contrastive, semantic segmentation, and LLM-based objectives. This results in an improvement in all metrics. The sixth row shows the results when we add contrastive and semantic losses together which gives significant improvement in the metrics. Finally, adding all loss functions (seventh row) achieves the best results on all metrics.

Fig. S9 visually validates the contribution of the identity loss $\mathcal{L}_{\text{id}}$ illustrating that it helps in recovering the hue and shades of original image in the reconstruction. Figs. S10 and S11 demonstrate similar results for the contrastive and semantic segmentation objectives. The contrastive loss $\mathcal{L}_{\text{con}}$ can recover high-level attributes such as color and texture information more accurately compared to our method without this loss component. We also observe that the semantic segmentation loss $\mathcal{L}_{\text{sem}}$ recovers the low-level attributes such as object boundaries and edges more vividly than our method without this loss component. These high/low-level attributes are of vital importance to downstream applications like object recognition.

We also perform separate tests on contrastive and semantic segmentation loss to examine the contribution of histogram-equalized LDR compared to original LDR images. We summarize these results in Tabs. S5 and S6. In both cases the loss calculated between the histogram-equalized versions of the LDR (instead of original the LDR) and the recnstructed HDR is better. This is due to the fact that histogram-equalized LDR can reveal semantic information in extremely over/underexposed areas of the original LDR images which in turn improves the extracted CLIP embeddings for $\mathcal{L}_{\text{con}}$ and segmentation maps for $\mathcal{L}_{\text{sem}}$.

Table S5. Semantic segmentation loss ablation results with the HDRTV [16], NTIRE [79] and HDR-Synth & HDR-Real [64] datasets.

Variant	PSNR↑	SSIM↑	LPIPS↓	HDR-VDP-2↑
LDR	38.89	0.956	0.033	66.92
LDR^{Hist}	**40.11**	**0.979**	**0.020**	**68.51**

Table S6. Contrastive loss ablation results with the HDRTV [16], NTIRE [79] and HDR-Synth & HDR-Real [64] datasets.

Variant	PSNR↑	SSIM↑	LPIPS↓	HDR-VDP-2↑
LDR	39.12	0.966	0.030	67.11
LDR^{Hist}	**40.11**	**0.979**	**0.020**	**68.51**

Finally, we experiment with the weights for the loss components with respect to PSNR. We perform tests on the temperature parameter τ of the contrastive loss and the weight parameters α and β for the loss components $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{sem}}$. Fig. S8 shows the results. The temperature parameter τ gives the optimal PSNR with $\tau = 0.08$ while the result degrades if we increase the value further. If we decrease the value, it gives a sub-optimal PSNR at $\tau = 0.07$, so we fix the value to $\tau = 0.08$ in all experiments. The values of α and β give optimal PSNR for $\alpha = 2$ and $\beta = 2$. The weights for the LLM-based loss are set to 3, 2, and 1.5 for the artifact, overexposed, and underexposed components, respectively. We conducted ablation experiments on these weights considering values from 1 to 5 and found that the selected combination of weights yield the best PSNR.

S4. Limitations and Future Work

Our method is designed for LDR $\leftrightarrow$ HDR bi-directional translation for images and can be further extended to video-based LDR $\leftrightarrow$ HDR translation with some modifications in the generators (*e.g.*, using special ViT [32] or Timesformer [9]) models. Diffusion-based cycle consistency approach can also be explored for video-based LDR/HDR translation. Further research can also explore multi-task architectures along with dynamic range translation such as deblurring, denoising, super-resolution, and demosaicing tasks. The pre-processing step in our loss calculations for contrastive and semantic segmentation loss employs histogram equalization, which can be further investigated with methods like Gamma correction [58] or his-

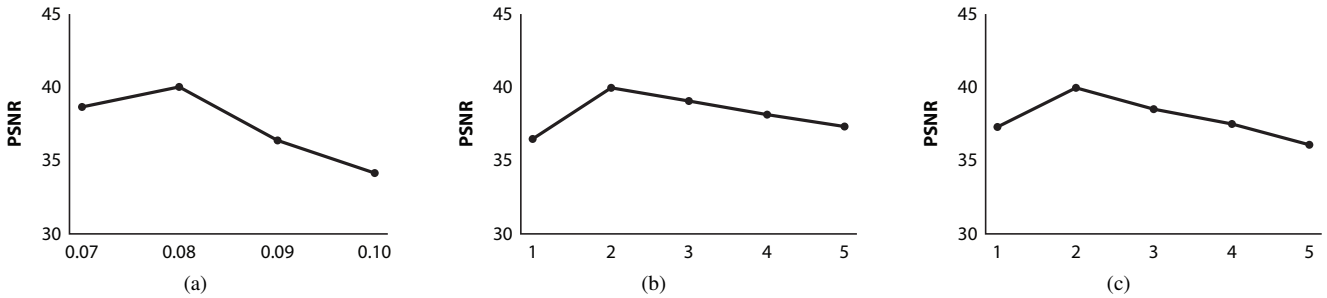


Figure S8. Ablation results for the temperature parameter τ (a), and loss component weights α (b) and β (c).

togram matching [18]. Another direction can be the integration of no-reference quality assessment [4] into the framework which will enable parallel training and prediction of the reconstructed LDR/HDR along with its quality score. More advanced LLM-based formulation such as explainability of the generation process and object's material based specular reflections can also be introduced in the future. Moreover, the Differentiable Ray Tracing technology [24] used to generate realistic HDR images/videos in video games [7] to simulate the physics behind light, specular reflections and shadows can also be explored using deep learning concepts like Nvdiffrast [53] and Neural Radiance Fields (NeRFs) [73] in 2D domain.

Figure S9. Ablation results for $\mathcal{L}_{\text{id}}$ with images from the HDR-Synth & HDR-Real dataset [64] and LDR-HDR pair datasets [40].

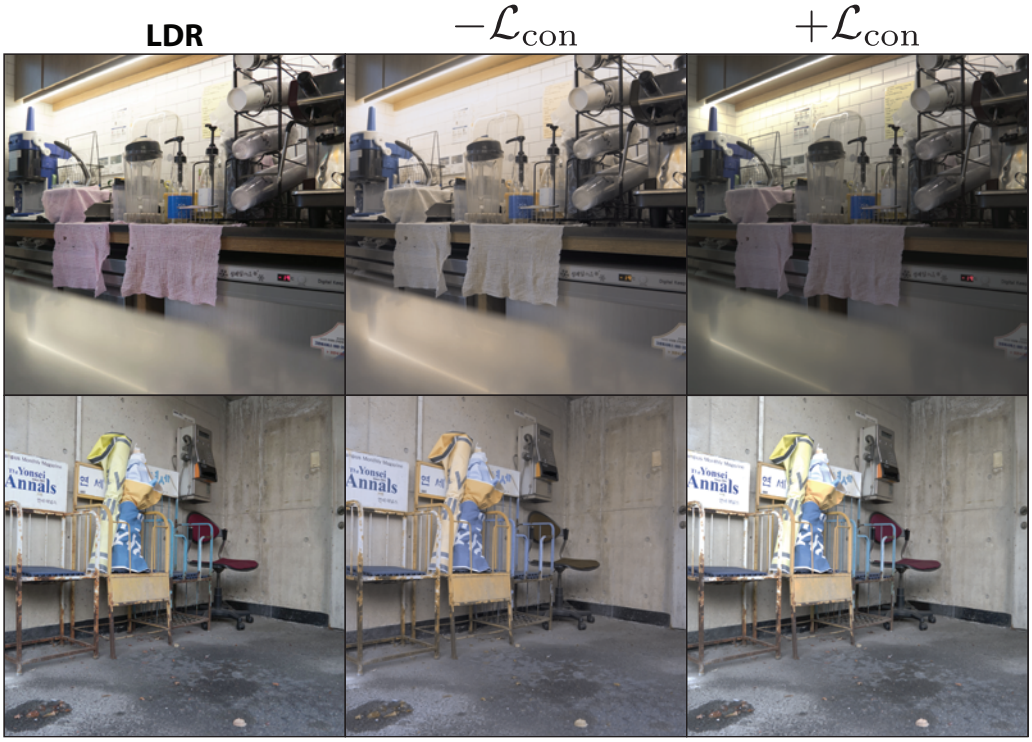


Figure S10. Ablation results for $\mathcal{L}_{\text{con}}$ with images from the LDR-HDR pair [40] dataset.

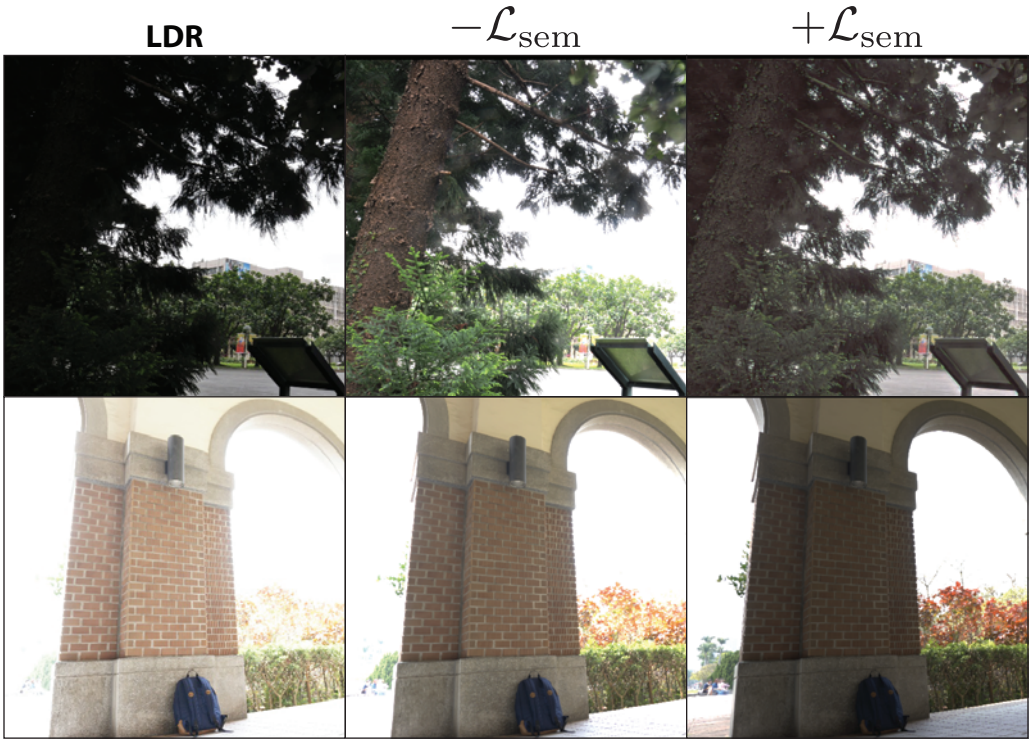


Figure S11. Ablation results for $\mathcal{L}_{\text{sem}}$ with images from the HDR-Synth & HDR-Real dataset [64] dataset.

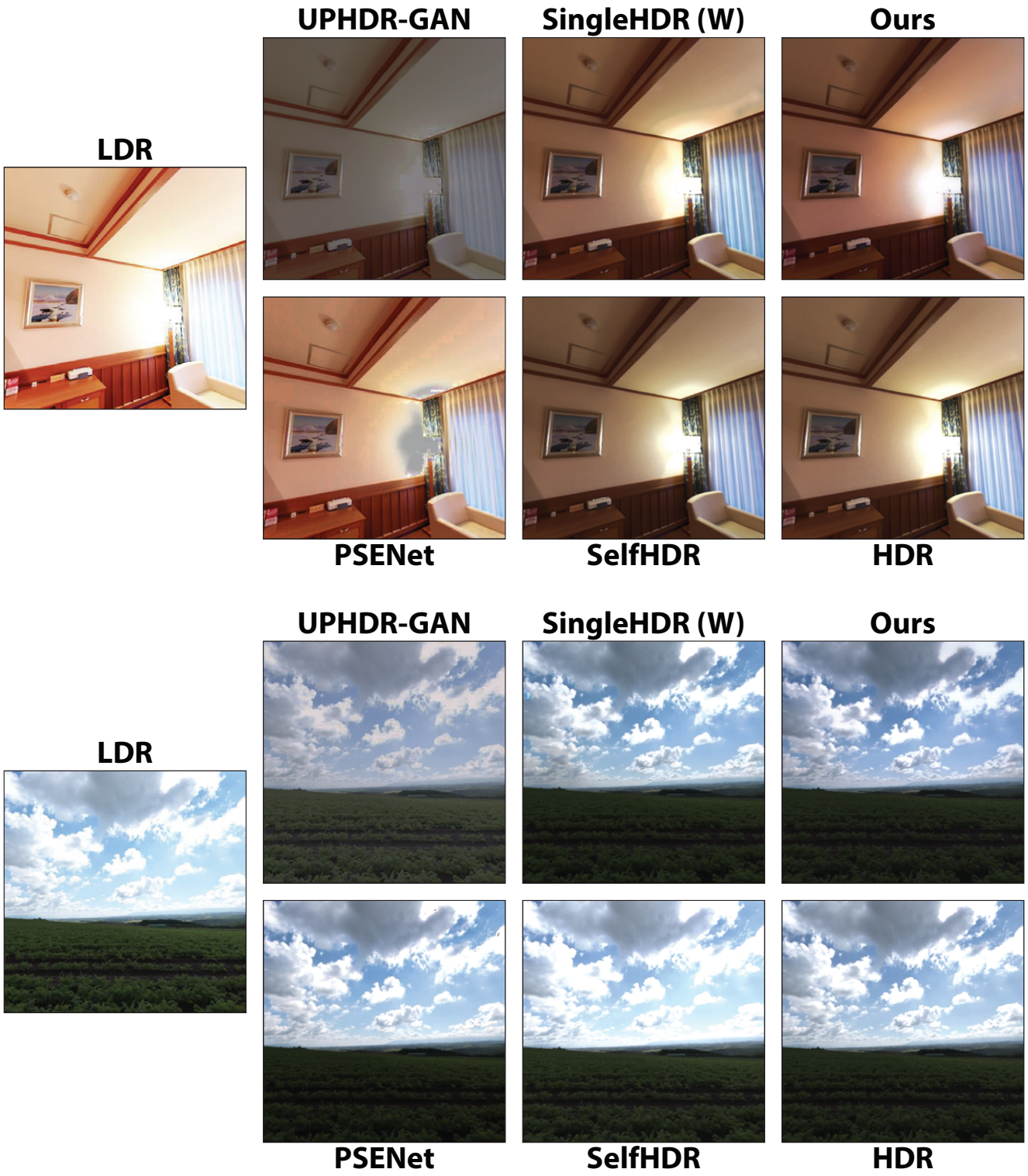


Figure S12. Examples of HDR images reconstructed with our method and recent state-of-the-art.

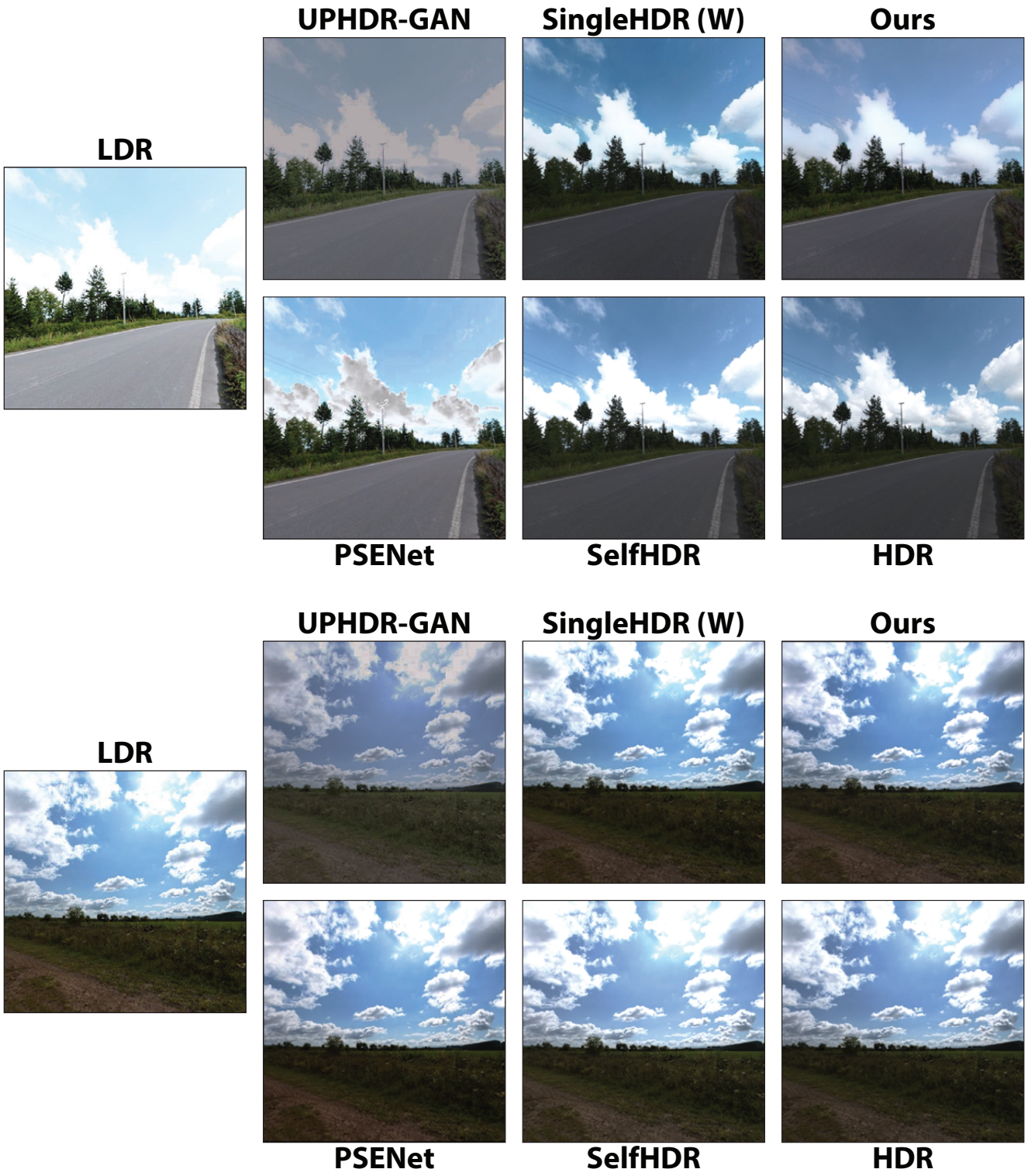


Figure S13. Examples of HDR images reconstructed with our method and recent state-of-the-art.

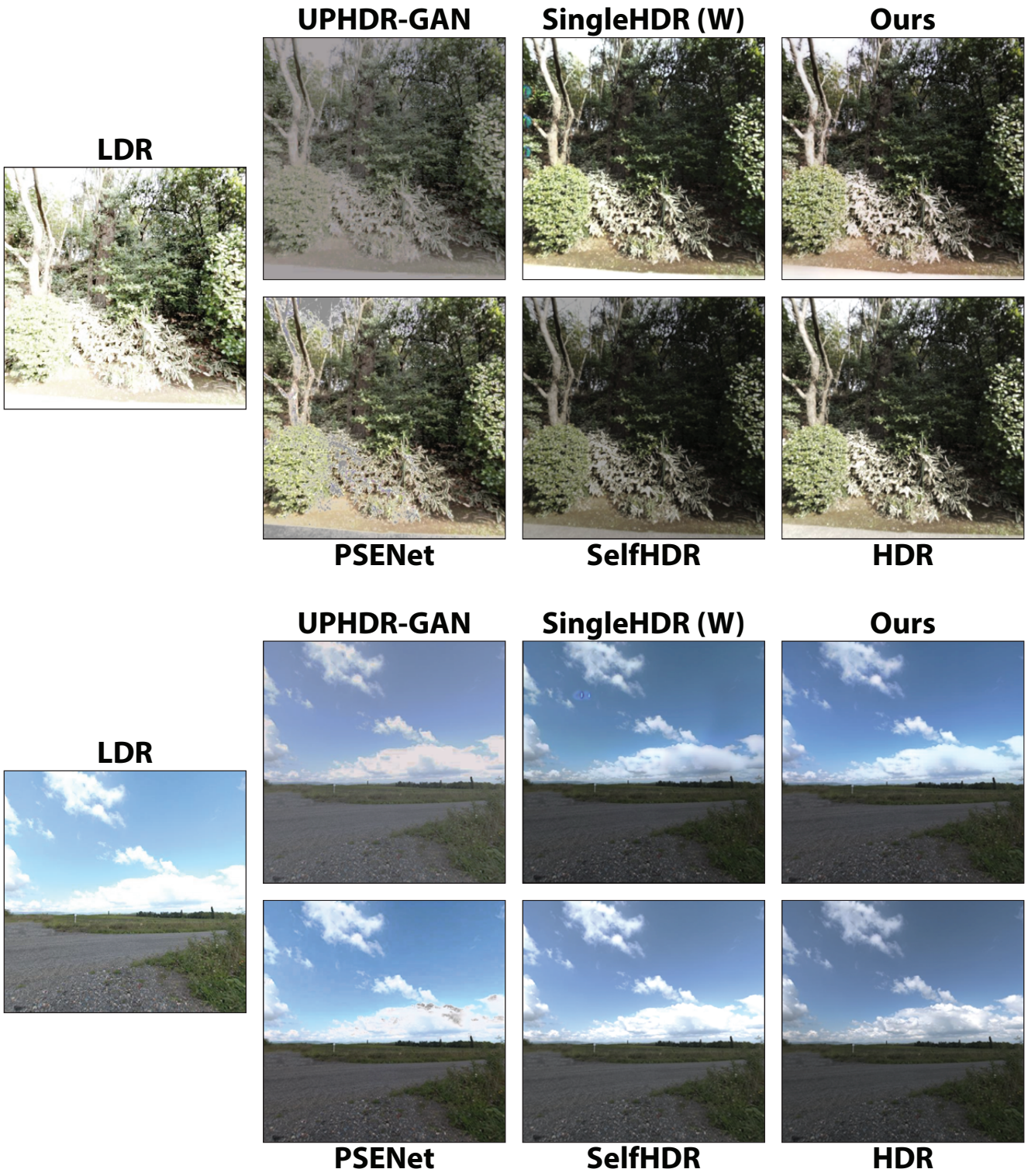


Figure S14. Examples of HDR images reconstructed with our method and recent state-of-the-art.

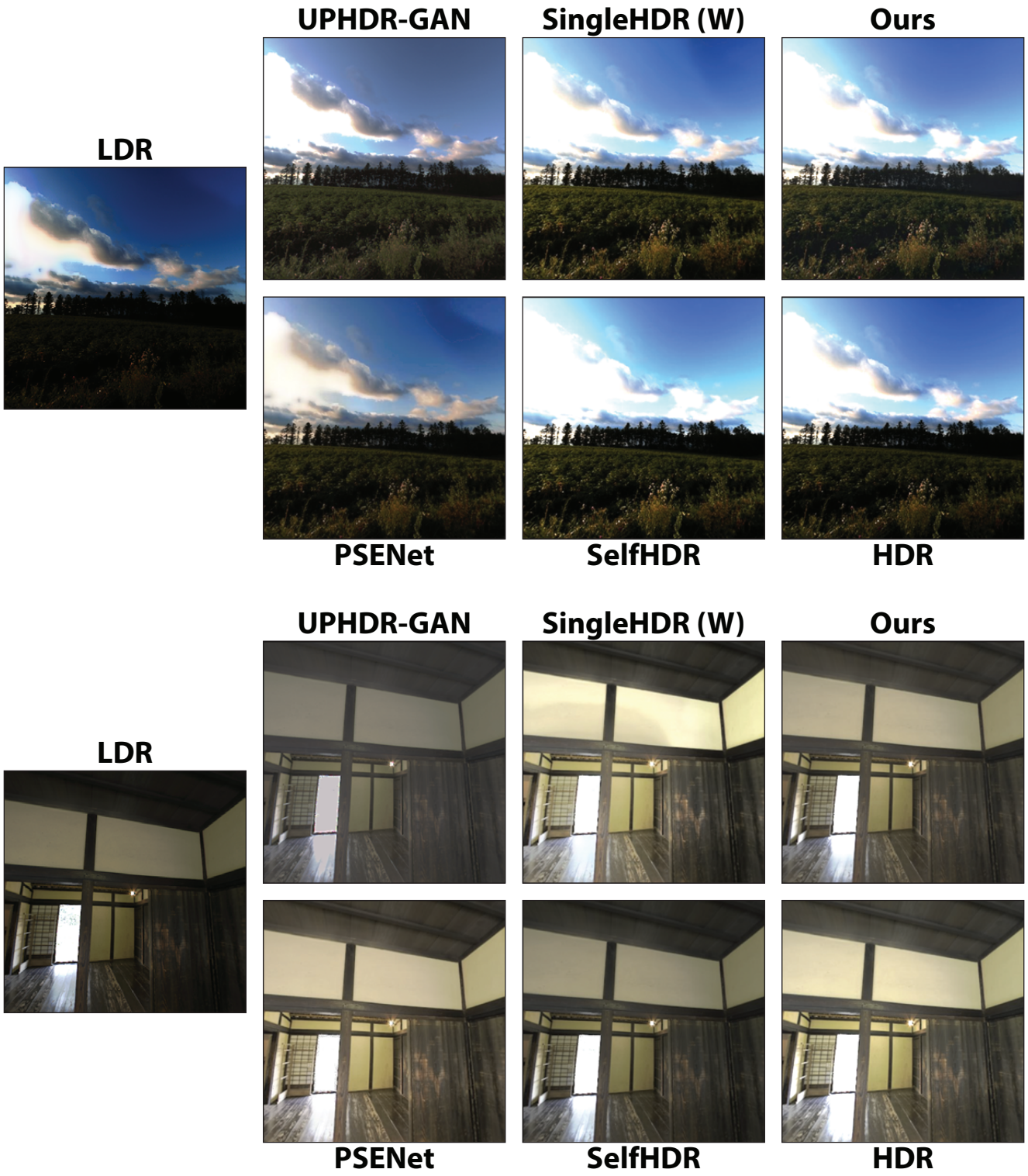


Figure S15. Examples of HDR images reconstructed with our method and recent state-of-the-art.

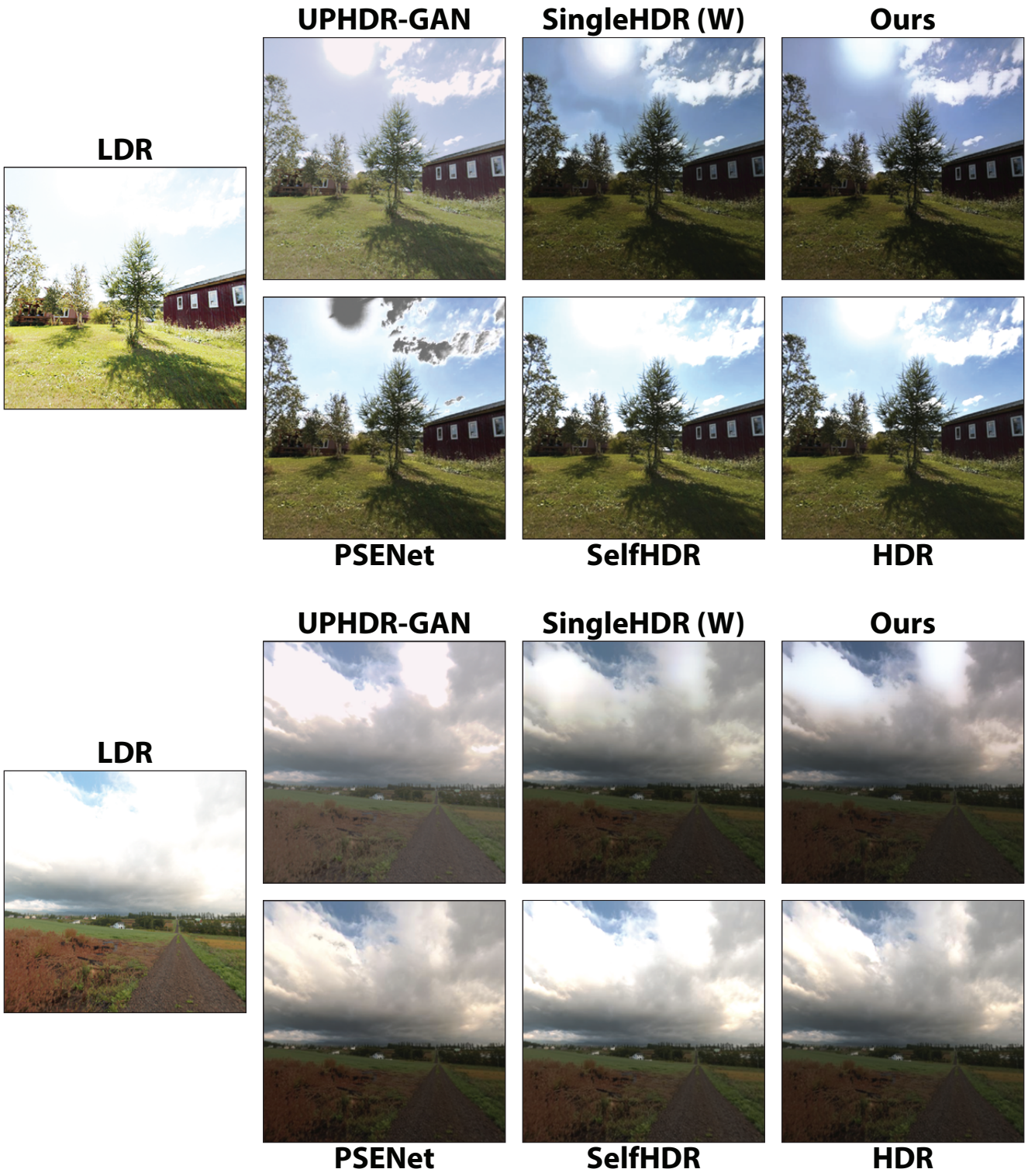


Figure S16. Examples of HDR images reconstructed with our method and recent state-of-the-art.

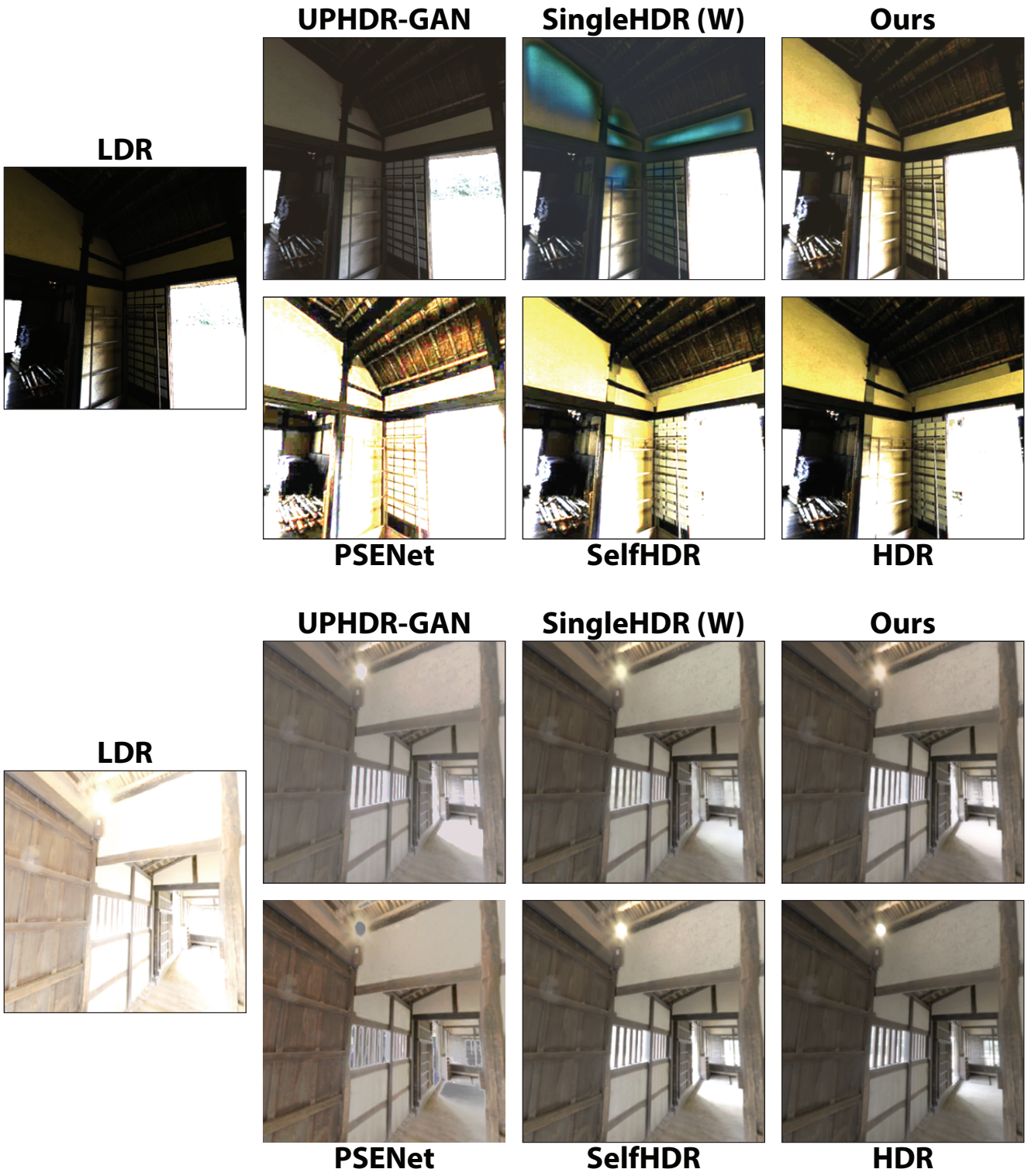


Figure S17. Examples of HDR images reconstructed with our method and recent state-of-the-art.

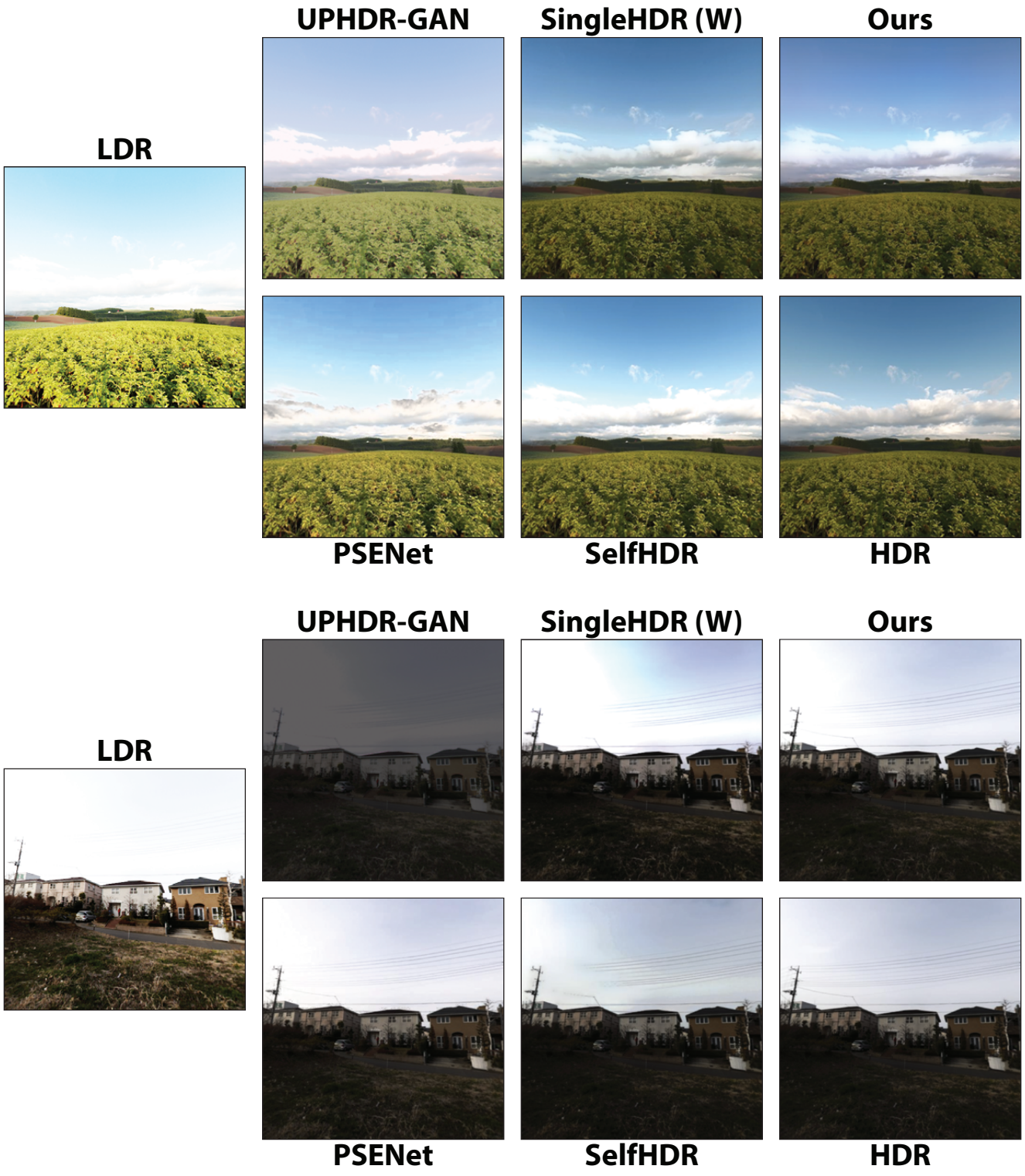


Figure S18. Examples of HDR images reconstructed with our method and recent state-of-the-art.