GlowGAN: Unsupervised Learning of HDR Images from LDR Images in the Wild

Chao Wang\textsuperscript{1} Ana Serrano\textsuperscript{2} Xingang Pan\textsuperscript{1,3} Bin Chen\textsuperscript{1} Karol Myszkowski\textsuperscript{1} Hans-Peter Seidel\textsuperscript{1} Christian Theobalt\textsuperscript{1} Thomas Leimkühler\textsuperscript{1}
\textsuperscript{1}MPI Informatik \　\textsuperscript{2}Universidad de Zaragoza \　\textsuperscript{3}Nanyang Technological University

\{chaowang, binchen, karol, theobalt, thomas.leimkuehler\}@mpi-inf.mpg.de
anase@unizar.es xingang.pan@ntu.edu.sg hpseidel@mpi-sb.mpg.de

Figure 1: Typical image datasets have a low dynamic range (LDR), resulting in over- and underexposed pixels (a). We devise a high-dynamic-range (HDR) generator trained only on in-the-wild LDR data. Our samples (b, tonemapped) exhibit details at all brightness levels. Our model can be used for inverse tone mapping to recover details in large-scale saturated regions (c).

Abstract

Most in-the-wild images are stored in Low Dynamic Range (LDR) form, serving as a partial observation of the High Dynamic Range (HDR) visual world. Despite limited dynamic range, these LDR images are often captured with different exposures, implicitly containing information about the underlying HDR image distribution. Inspired by this intuition, in this work we present, to the best of our knowledge, the first method for learning a generative model of HDR images from in-the-wild LDR image collections in a fully unsupervised manner. The key idea is to train a generative adversarial network (GAN) to generate HDR images which, when projected to LDR under various exposures, are indistinguishable from real LDR images. The projection from HDR to LDR is achieved via a camera model that captures the stochasticity in exposure and camera response function.

Experiments show that our method GlowGAN can synthesize photorealistic HDR images in many challenging cases such as landscapes, lightning, or windows, where previous supervised generative models produce overexposed images. With the assistance of GlowGAN, we showcase the novel application of unsupervised inverse tone mapping (GlowGAN-ITM) that sets a new paradigm in this field. Unlike previous methods that gradually complete information from LDR input, GlowGAN-ITM searches the entire HDR image manifold modeled by GlowGAN for the HDR images which can be mapped back to the LDR input. GlowGAN-ITM achieves more realistic reconstruction of overexposed regions compared to state-of-the-art supervised learning models, despite not requiring HDR images or paired multi-exposure images for training.
1. Introduction

High Dynamic Range (HDR) images [57] are capable of capturing and displaying much richer appearance information than Low Dynamic Range (LDR) images, thus playing an important role in image representation and visualization. The most popular method to acquire HDR images is multiple exposure blending, which requires capturing a set of LDR images of the same scene with different exposures [13, 52, 59]. However, this is time and effort intensive and only suitable for static scenes. Due to this limitation, existing HDR image datasets only cover limited scene categories and have much fewer images than LDR datasets. Thus, supervised learning methods [16, 47, 14, 41, 45, 43, 19, 71, 70, 28] that reconstruct an HDR image from a LDR image are constrained by the HDR datasets and cannot extend to cases where no HDR training data is available, e.g., lightnings or campfires.

While HDR images are hard to collect, it is much easier to collect a large number of LDR images from the Internet. This motivates us to investigate a new unsupervised learning problem: Can we learn to reconstruct HDR images from in-the-wild LDR images? The LDR images do not need to depict the same scene, it is enough if they contain a roughly similar class of scenes (e.g., landscapes) with various exposures. This weak multi-exposure assumption is often naturally satisfied for in-the-wild LDR datasets as images can come from different camera parameters or different adjustments of the auto-exposure mode. This problem is challenging as only a single exposure is available for each scene; therefore, a way to merge the multi-exposure information spread across different scenes is required.

In this work, we address this challenge via GlowGAN, which, to our knowledge, is the first method to learn an HDR generative model from in-the-wild LDR image collections in a fully unsupervised manner. GlowGAN uses adversarial training of an HDR generator with a discriminator that operates merely in LDR. Specifically, the generator produces an HDR image, which is projected to LDR via a camera model and is then sent to the discriminator as a fake image for adversarial training. The camera model consists of multiplying the HDR sample with an exposure value, clipping the dynamic range, and applying a camera response function (CRF). Importantly, during training, we use a randomly sampled exposure from a prior Gaussian distribution when projecting HDR to LDR. This requires the generated HDR images to be realistic under any possible exposure, thus only valid HDR samples would satisfy this “multi-exposure constraint”. Furthermore, we also model the stochasticity in the non-linear camera response by randomly sampling CRFs according to a well-established parametric distribution [16, 21]. During inference, we can disable the camera model so that the generator produces HDR imagery directly.

We conduct extensive experiments on several datasets collected from the Internet, including landscapes, windows, lightning, fireplaces, and fireworks. By training on these LDR images, GlowGAN successfully learns to generate high-quality HDR images that capture rich appearance information. These images can be presented on HDR displays, or via suitable tone mapping to create appealing imagery. In contrast, previous unconditional GANs tend to miss information in over- or under-exposed regions.

By modeling a distribution of HDR images, GlowGAN paves the way for new applications such as unsupervised inverse tone mapping (GlowGAN-ITM). ITM aims to reconstruct an HDR image from a single-exposure LDR input, where a key challenge is to restore the flat-white overexposed regions [16, 19]. We can use a pre-trained GlowGAN as a prior and apply GAN inversion to optimize latent code and exposure, making the model generate the corresponding HDR image for the input LDR image. An exciting result is that our method can, without using any HDR imagery or paired multi-exposure data, reconstruct starkly more plausible information for large overexposed regions than other supervised learning approaches trained on such data. Our contributions are summarized as follows:

- We are the first to present unsupervised learning of HDR images from in-the-wild LDR images. This gets rid of the reliance on ground truth HDR images that are much harder to collect.
- To achieve this, we propose a novel GlowGAN, which bridges HDR space and LDR space via a camera model. GlowGAN can synthesize diverse high-quality images with a much higher dynamic range than vanilla GANs, opening up new avenues for getting cheap abundant HDR data.
- Using GlowGAN as a prior, we design an unsupervised inverse tone mapping method GlowGAN-ITM, which reconstructs large overexposed regions significantly better than the state-of-the-art fully-supervised approaches.

Our code, pre-trained models, and datasets are publicly available under https://glowgan.mpi-inf.mpg.de.

2. Related Work

2.1. High Dynamic Range Imaging

HDR imaging [57] is crucial for creating immersive viewing experiences that portray the vast dynamic range of the real world. While HDR capture is cumbersome and HDR displays are not yet commonplace in most environments [64], the representation of visual information free from the limitations of typical LDR encodings has evolved significantly in recent years [68]. Working with HDR content is crucial for rendering [12] and has been shown to be beneficial for 3D scene reconstruction: An HDR representation can naturally handle multi-exposure [29, 27] and raw data [51]. Recon-
A particular interest has evolved around the conversion between HDR and LDR content. Tone mapping, the transformation from HDR to LDR with as little information loss as possible, is a mature field with well-understood trade-offs [57, 66]. In contrast, inverse tone mapping (ITM) [5, 58], the recovery of HDR content from LDR imagery, remains a challenging inverse problem. It typically involves multiple steps, including linearization, dynamic range expansion, reconstruction of over- and underexposed regions, artifact reduction, and color correction [3]. Among these, the reconstruction of saturated pixels is considered the most challenging [16, 19], as it requires the hallucination of content [67].

Early ITM works considered expansion curves using either global [40, 48, 2, 49] or content-driven operators [50, 5, 4, 6, 58, 58], without explicitly reconstructing saturated regions. More recent learning-based solutions can be categorized into two streams: A neural network either predicts the HDR image directly (Direct ITM) [16, 47, 19, 45, 71, 70, 10], or it predicts multiple LDR images with different exposures [14, 41, 42, 43, 28], which are subsequently merged into an HDR image (Indirect ITM) [13, 52, 59]. ITM and exposure fusion can be combined with adversarial training [71, 42, 54]. Also, the extension to video ITM has been explored extensively, leveraging inter-frame consistency [30, 37, 38, 7, 23, 24].

The majority of learning-based techniques discussed above require supervision from paired LDR–HDR data or exposure stacks. A notable exception is the work of Banterle et al. [7], which recovers details from non-saturated frames in a video sequence and applies them to saturated frames. The need for training data constitutes a fundamental problem: HDR image data is hard to obtain and therefore naturally scarce. Further, most HDR capture techniques require static content, which significantly restricts the applicability of learning-based methods to arbitrary scenes. In contrast, we are the first to train an unsupervised HDR image generator from a fully unstructured set of LDR images. We believe this is an important step towards solving two main problems in HDR imaging: First, our generator can synthesize an abundance of HDR samples, alleviating the scarcity of HDR content. Second, our system allows to perform ITM with a significant improvement in the reconstruction of overexposed regions. We summarize the difference between our method and previous methods in Table 1 and Fig. 2.

### Table 1: Contributions of GlowGAN in comparison to other method classes.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Training Dataset</th>
<th>HDR¹</th>
<th>Unsupervised²</th>
<th>Generative³</th>
<th>Hallucination⁴</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla GAN [20, 62, 36]</td>
<td>Unstructured LDRs</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Direct ITM [16, 47, 19, 45, 71, 70, 10]</td>
<td>Paired LDR-HDR images</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>Weak</td>
</tr>
<tr>
<td>Indirect ITM [14, 41, 42, 43, 28]</td>
<td>Paired LDR exposure stacks</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>Weak</td>
</tr>
<tr>
<td>GlowGAN (Ours)</td>
<td>Unstructured LDRs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Strong</td>
</tr>
</tbody>
</table>

¹ Capable of generating HDR images or performing inverse tone mapping.
² Capable of learning to produce HDR images from unstructured and unpaired LDR data.
³ Capable of generating new images.
⁴ Capable of hallucinating overexposed areas with plausible content, which is one of the most challenging problems in ITM.

![Figure 2: Vanilla GANs are trained on unstructured LDR data and can only generate LDR. Direct ITM methods perform image-to-image (I2I) translation, but need to be trained on paired LDR–HDR data. Indirect ITM methods use I2I to produce an LDR exposure stack to be fused into an HDR result and require supervision in the form of LDR exposure stacks. In contrast, our approach allows to generate HDR images (left to right), using a stochastic camera model to allow supervision via unpaired in-the-wild LDR images. ITM (right to left) is performed using GAN inversion. Only our method yields HDR from unstructured LDR training data.](image-url)

### 2.2. Lossy Generative Adversarial Networks (GAN)

GANs [20] are very successful in modelling distributions of images with high visual fidelity. The StyleGAN family [35, 36, 33, 34] marks the current state of the art, scaling to high resolutions, while the recent extension StyleGAN-XL [62] allows for unprecedented diversity in the generated content. To this date, (unconditional) GANs operate in LDR, since high-quality HDR data at the scale required for successful training is difficult to obtain. We propose a simple modification to the GAN training pipeline, which allows to train an HDR generator from available LDR data only.

AmbientGAN [15] has demonstrated that it is feasible to train a generative model from lossy measurements, i.e., a GAN can be trained from degraded samples, as long as the stochastic properties of the degradation are known. This concept has been used to learn a generator for clean images from noisy data [32], or for all-in-focus images from data.
that idea contains shallow depth of field [31]. Most prominently, the idea has been applied to learn 3D generators from 2D images by explicitly modeling the projection from 3D to 2D using a distribution of extrinsic camera parameters [25, 17, 44, 53, 63, 65]. We follow the paradigm of injecting a degradation model into the GAN pipeline, by devising a novel model of the processing steps in a digital camera, converting HDR radiance into LDR pixel intensities.

3. Method

We train a GAN [20] to capture the distribution of HDR images in a domain (e.g., landscapes) by combining it with a stochastic camera model that transforms the generated HDR images into their LDR counterparts. The discriminator is only fed LDR images, which allows the system to be trained on an easily accessible in-the-wild LDR image dataset. Our camera model can be inserted into any GAN model to yield HDR outputs as long as the LDR training dataset consists of images exhibiting different exposures across samples. We consider this a mild assumption, which in particular in-the-wild photo datasets easily satisfy. See Fig. 3 for an overview of our system. In the following, we describe our pipeline in detail (Sec. 3.1), before turning to our main application of unsupervised inverse tone mapping GlowGAN-ITM (Sec. 3.2).

3.1. GlowGAN

We seek to capture the unknown true distribution of HDR images $p_{\text{HDR}}$ from samples of the distribution of LDR images $p_{\text{LDR}}$. Similar to the standard GAN setup, we achieve this by training a generator $G$ which turns a random latent vector $z \in \mathbb{R}^k$ into an HDR sample $r \in \mathbb{R}^{H \times W \times 3}$, an RGB image with $H \times W$ pixels and no upper restriction on the value range. To train $G$, we inject a camera model $C \in \mathbb{R}_{\geq 0}^{H \times W \times 3} \rightarrow [0, 1]^{H \times W \times 3}$ into the adversarial training pipeline, turning the unbounded HDR image into an LDR image with values between 0 and 1. $C$ captures the distribution $p_{\text{cam}}$ of pixel-wise image transformations typically applied in a digital camera to convert incoming radiance to pixel intensities, including varying exposures, clipping, and varying non-linearities arising from the camera response function (CRF). The result of this process is an LDR image $I = C(r) = C(G(z))$. The discriminator $D$ is tasked with differentiating the fake samples $I$ from samples from the distribution of true LDR images $p_{\text{LDR}}$. Since the samples $r$ undergo stochastic projections from HDR to LDR, $G$ is forced to produce valid HDR images, resulting in the distribution $p_{\text{HDR}}$ of generated HDR images approaching the true distribution $p_{\text{HDR}}$ [15].

As our generator and discriminator backbone, we choose the current state-of-the-art model StyleGAN-XL [62]. This model has been demonstrated to yield excellent image quality on diverse datasets. The generator model consists of two stages (left block in Fig. 3): First, a mapping network $M_{\theta_M}$ in the form of an MLP with parameters $\theta_M$ turns the initial random vector $z \in \mathbb{R}^k$ into a more disentangled latent feature representation $w \in W \subset \mathbb{R}^k$. Second, $w$ is fed into a synthesis CNN $S_{\theta_S}$ with parameters $\theta_S$ to yield the final HDR output $r = S_{\theta_S}(M_{\theta_M}(z))$. Notice that the particular choice of the generator and discriminator networks is orthogonal to our approach. Except for the camera model introduced in the next paragraphs, we use StyleGAN-XL without any modifications in architecture or training hyperparameters. We verify this in Sec. 4.2.

At the core of our method is the stochastic camera model $C$ (central block in Fig. 3) which projects the HDR image $r$ onto an LDR counterpart $I$. It is designed to model the distribution of typical processing steps in a digital camera:

$$C(r) = \text{CRF}_{\beta, \gamma} \left( \min(2^r \cdot r, 1) \right).$$

In the first step, we multiply each pixel of $r$ with a single global exposure value. The exposure is parameterized by the random variable $e$, capturing the exposure distribution of typical cameras arising from the combined effect of aperture, shutter speed, and sensor sensitivity (ISO). We do not have access to the true distribution of exposure parameters $e$, as images from in-the-wild image collections frequently do not have EXIF headers that would contain this information. Since the exposure is a combined effect of multiple factors (aperture, shutter, ISO, time of day, etc.), we choose to model $e$ using a normal distribution, i.e., $e \sim N(0, \sigma_e^2)$. The exposure variance $\sigma_e^2$ is the only hyper-parameter in our system and is analyzed in Sec. 4.2.

After applying a random exposure, the $\min$ operator in Eq. 1 clips large radiance values to 1, effectively removing all contrast in overexposed regions. Notice that this loss of information is precisely the reason why $C$ is not invertible for individual images: By selecting an exposure, we only observe a bracket of radiance values. In contrast, our method seeks to invert $C$ over the distribution of images and camera models [15].

The final component of our model in Eq. 1 is the non-linear sensor response. The CRF describes the mapping from radiance values arriving at the sensor to pixel intensities stored in the final image. We follow the established distribution of Eilersen et al. [16]:

$$\text{CRF}_{\beta, \gamma}(x) = \frac{(1 + \beta)x^\gamma}{\beta + x^\gamma},$$

with $\beta \sim N(0.6, 0.1)$ and $\gamma \sim N(0.9, 0.1)$ as obtained from the analysis of a large dataset [21]. We visualize the distribution of CRFs arising from this model in Fig. 4.
With the full stochastic camera model in place, our system can be trained in an adversarial fashion from scratch without further modifications to the standard GAN pipeline to yield an HDR generator trained only on an LDR image dataset.

### 3.2. GlowGAN-ITM

In addition to producing HDR image samples, a trained GlowGAN can be used to perform unsupervised inverse tone mapping. While existing approaches typically rely on supervision from LDR–HDR image pairs or exposure stacks, our method enables the recovery of HDR imagery from LDR inputs without HDR data using GAN inversion [72, 69]. This allows us to hallucinate plausible content in over-exposed image regions, which is one of the most challenging problems in ITM and, hence, a particular focus of our approach.

To obtain high-quality ITM results and to facilitate multimodality, we choose a per-image optimization-based approach to perform the inversion [11, 1]. Given an LDR image \( \hat{I} \), we consider the following optimization objective:

\[
[e^*, w^*, \theta^*_S] = \arg \min_{c, w, \theta_S} \Phi \left( C(S_{\theta_S}(w)), \hat{I} \right). \tag{2}
\]

Here, we jointly optimize over exposure \( e \) and the latent code \( w \) while fine-tuning the synthesis network parameters \( \theta_S \) to obtain a faithful match between the target LDR image \( \hat{I} \) and its reconstruction. \( \Phi \) denotes the discrepancy measure between the two images. Using the Adam optimizer [39] with standard parameters, we proceed in two stages: In the first stage, we exclude the generator weights \( \theta_S \) from the optimization and measure image discrepancy \( \Phi \) using the LPIPS perceptual distance [18]. In the second stage, we only optimize (fine-tune) \( \theta_S \) using the pivotal tuning technique of Roich et al. [60], with \( \Phi \) being the sum of a pixel-wise \( \ell_2 \) loss and the LPIPS perceptual distance. We relax \( w \) to explore the extended latent space \( W^+ \) [1]. We did not find it necessary to optimize over the CRF parameters \( \beta, \gamma \) for high-quality results and consequently fix them to their mean values. More details on the optimization can be found in the supplemental. Upon completion of the optimization, we obtain the HDR version of \( \hat{I} \) via \( r^* = S_{\theta^*_S}(w^*) \).

Following the lossy projection in Eq. 1, the mapping from LDR to HDR images is not unique: Overexposed regions in the LDR image can be explained by many different HDR images. Our system allows capturing this multi-modality by running the optimization of Eq. 2 multiple times with different parameter initializations for \( w \) and \( e \). Specifically, we initialize each optimization run with \( w = M_{\theta_S}(z) \), where \( z \) is a normally distributed random vector. This allows us to obtain multiple plausible HDR solutions, which almost exclusively differ in the overexposed regions.

Obtaining accurate GAN inversion results is challenging [9, 1]. Fortunately, in most cases, we are only interested in hallucinating content in saturated regions, while well-exposed pixels can be re-used after linearization. Therefore, optionally, we diminish potential distortions arising from the inversion by blending the linearized original LDR image \( \hat{I} \) with the reconstructed HDR result \( r^* \) [16, 19]. Referring to the camera model in Fig. 3, this blending happens in HDR space between \( \text{Mult.} \) and \( \text{Clip.} \) and is computed as follows:

\[
r^*_{\text{blend}} = e^* \cdot (m \odot r^*) + (1 - m) \odot \text{CRF}^{-1}(\hat{I}).
\]

Here, \( \odot \) denotes the Hadamard product and \( m \) is a soft mask, indicating saturated pixels in \( \hat{I} \), which we compute for each pixel \( i \) following Eilertsen et al. [16]:

\[
m_i = \max \left( \frac{\hat{I}_{i,c} - \tau}{1 - \tau}, 0 \right).
\]

where \( \hat{I}_{i,c} \) denotes the LDR image with pixel index \( i \) and color channel \( c \). We set the threshold \( \tau = 0.97 \) in all our experiments, resulting in a short ramp toward saturation.

### 4. Experiments

We have conducted the following experiments to demonstrate the effectiveness of our approach. Sec. 4.1 shows the generated HDR data and compares it with an LDR equivalent; Sec. 4.2 explores the influences of the exposure distribution, model backbone, and sampled camera response curve; and Sec. 4.3 presents the results of our GlowGAN-ITM. We use the tone mapper of Mantiuk et al. [46] to display HDR content in this paper. We refer readers to the supplemental for more results and full HDR data.

**Implementation details.** We implement our method on top of the official StyleGAN-XL [62] implementation under the
Figure 5: (a) Dynamic range comparison between our GlowGAN (top) and a vanilla GAN (bottom) when varying the exposure values (EV) of the image. (b) HDR samples generated by our GlowGAN for the five tested datasets (please see Fig. 1 for more samples). Our method generates high-quality images with much higher dynamic range and without overexposure.

PyTorch [56] environment. Training takes six days with four RTX 8000 GPUs for 256×256 resolution dataset – roughly the same time as training a vanilla StyleGAN-XL model.

Datasets. We collect five datasets of LDR images from the Internet: Landscapes (∼7700 images), Lightning (∼7000 images), Windows (∼4200 images), Fireplaces (∼2600 images), and Fireworks (∼5600 images). We randomly crop and resize each image to the target resolution. Please refer to Fig. 1 and Fig. 5 for examples of generated images from models trained on these datasets. We collect our datasets from several websites: Flickr, Pexels, Instagram, and 500PX. We provide more dataset statistics in the supplemental.

4.1. Generation of HDR Images

We show in Fig. 5 a variety of samples generated with GlowGAN and a comparison with a vanilla StyleGAN-XL. Generated samples from the vanilla GAN often bear overexposed regions similar to those in the LDR training images. In contrast, samples from GlowGAN preserve detailed appearance information even for bright objects such as the sun, as they have more extensive dynamic ranges than those from the vanilla GAN.

To analyze dynamic range, we plot the image brightness histogram of the two models in Fig. 6, where each histogram is computed from 500 randomly sampled images. It can be seen that the histogram of the vanilla GAN exhibits an early cutoff, while GlowGAN clearly avoids pixel intensity clamping (Fig. 6a) and has a much wider histogram, also in the dark regions (Fig. 6b). These results satisfy our expectation of learning HDR information from LDR data. In the supplemental, we compare LDR samples of our model with their nearest neighbors in the training dataset. Our samples are clearly different, strongly indicating that our model does not memorize the dataset. GlowGAN thus opens up a new avenue for getting cheap abundant HDR data that can be used in several applications, e.g., for augmenting training data for ITM methods or for creating environment maps for image-based lighting (IBL) [12], as demonstrated in the supplemental.

### Table 2: The effect of generator backbone and exposure variance $\sigma_e^2$ on image quality and dynamic range.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma_e^2$</th>
<th>FID↓</th>
<th>KID($\times10^4$)↓</th>
<th>DR50</th>
<th>DR90</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG-XL$^1$</td>
<td>–</td>
<td>3.48</td>
<td>2.69</td>
<td>11.9</td>
<td>12.3</td>
</tr>
<tr>
<td>Ours</td>
<td>0.1</td>
<td>3.54</td>
<td>3.03</td>
<td>15.3</td>
<td>19.9</td>
</tr>
<tr>
<td>Ours</td>
<td>1.0</td>
<td>3.61</td>
<td>3.07</td>
<td>15.4</td>
<td>20.2</td>
</tr>
<tr>
<td>Ours</td>
<td>3.0</td>
<td>3.87</td>
<td>4.04</td>
<td>16.2</td>
<td>20.7</td>
</tr>
<tr>
<td>Ours</td>
<td>5.0</td>
<td>4.00</td>
<td>4.78</td>
<td>16.5</td>
<td>20.8</td>
</tr>
<tr>
<td>SG2$^2$</td>
<td>–</td>
<td>9.2</td>
<td>23.37</td>
<td>12.9</td>
<td>16.9</td>
</tr>
<tr>
<td>Ours w/ SG2$^3$</td>
<td>1.0</td>
<td>10.02</td>
<td>28.41</td>
<td>16.3</td>
<td>22.9</td>
</tr>
</tbody>
</table>

$^1$ Refers to a vanilla StyleGAN-XL model.
$^2$ Refers to a vanilla StyleGAN2-ADA model.
$^3$ Refers to our approach with a StyleGAN2-ADA backbone.

4.2. Ablation Study

Exposure Distribution. Our method assumes the exposure $e$ follows a Gaussian distribution with variance $\sigma_e^2$. Here we study how $\sigma_e^2$ impacts the generated image quality and the dynamic range. We employ the commonly used FID [26]...
Figure 7: Results for the ITM application and comparisons to six state-of-the-art methods for different datasets. Given a single LDR image as input, our method can produce an HDR image that contains plausible and realistic content in the previously overexposed regions, while previous methods tend to produce blurred results or even noticeable artifacts in such regions. In the first row (Landscapes dataset), it can be seen that recovering the ground truth content present in the original HDR scene is not possible in fully saturated regions, however, our method is able to produce plausible results that are consistent with the scene.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CRF</th>
<th>FID ↓</th>
<th>KID (×10^4) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landscapes</td>
<td>Fixed</td>
<td>3.89</td>
<td>3.80</td>
</tr>
<tr>
<td></td>
<td>Stochastic</td>
<td>3.61</td>
<td>3.07</td>
</tr>
<tr>
<td>Lightning</td>
<td>Fixed</td>
<td>3.40</td>
<td>4.77</td>
</tr>
<tr>
<td></td>
<td>Stochastic</td>
<td>3.29</td>
<td>4.57</td>
</tr>
</tbody>
</table>

Table 3: Comparing quality with fixed and stochastic CRFs.

Figure 8: Our method can generate different but plausible HDR images from a single overexposed LDR input image.

and KID [8] scores to evaluate image quality on the Landscapes dataset. As we do not have ground truth HDR images, the scores are computed between the generated LDR images (i.e., the output of the camera model) and the LDR training images. We also compute the dynamic range (DR) for each generated HDR image \( r \) as \( DR = \log_2 \left( \frac{r_{\text{max}}}{r_{\text{min}}} \right) \), where \( r_{\text{max}} \) and \( r_{\text{min}} \) are the max and min values of the image, respectively. We report the median and the 90th percentile of DR computed over 50k images, referred to as DR50 and DR90, respectively. From Table 2, we can observe a trade-off between image quality and the dynamic range, i.e., increasing \( \sigma_e^2 \) leads to a higher dynamic range (with diminishing returns for high \( \sigma_e^2 \)) but slightly worse FID and KID scores. To understand the positive correlation between \( \sigma_e^2 \) and DR, suppose that \( r \) has a low dynamic range, then with a very small or large exposure \( e \) (which is more likely to happen for large \( \sigma_e^2 \)), it would produce an out-of-distribution LDR image \( l \) that is overly dark or bright. In other words, only a valid high dynamic range \( r \) can yield realistic LDR images when processed with different exposures. On the other hand, as \( \sigma_e^2 \) increases, the camera model interferes more with the image generation process, which may increase the
training difficulty as the Gaussian distribution is only an approximation to the underlying exposure distribution. In practice, users can choose a suitable $\sigma_e^2$ depending on their goal. In most of our experiments, we use $\sigma_e^2 = 1$ as it already removes overexposure while featuring good quality, and it also obtains better scores in the inverse tone mapping application. We provide more results on the effect of $\sigma_e^2$ in the supplemental. From Table 2, we can further see that a vanilla StyleGAN-XL exhibits slightly higher image quality, but a substantially smaller (log-scale) dynamic range. Numerical results on other datasets are provided in the supplemental.

**Generator backbone.** We further test our method based on StyleGAN2-ADA [33]. As Table 2 shows, our method also successfully synthesizes images with high dynamic range, albeit the final image quality directly depends on the baseline generative model.

**Stochastic CRF.** We further study the effects of the stochastic CRF sampling process in our model. Table 3 compares our stochastic CRF sampling with a fixed CRF, using $\beta = 0.6$ and $\gamma = 0.9$. Modelling the stochasticity leads to a clear improvement in FID and KID scores. This is because the in-the-wild LDR images used for training are captured with different cameras with diverse CRFs, which can be better modelled via stochastic CRF sampling.

**4.3. Application: Inverse Tone Mapping**

A potential application of our approach is GlowGAN-ITM. We show both quantitatively and through a user study that our method outperforms previous approaches in hallucinating content in large overexposed regions, effectively recovering HDR content from a single LDR image.

**Objective comparisons.** We compare our GlowGAN-ITM to six state-of-the-art fully-supervised ITM methods, which we abbreviate as HDRCNN [16], MaskHDR [19], SingleHDR [45], ExpandNet [47], ReHDR [41], and LANet [70]. Following the work of Hanji et al. on quality assessment of single image HDR reconstruction methods [22], we select their three recommended full-reference metrics (PU21-PSNR, PU21-VSI, and HDR-VDP3) and their recommended non-reference metric (PU21-PIQE). As test set for the reference metrics we use as HDR ground truth a set of 62 images collected from existing datasets [22, 55] and generate the corresponding LDR input images following the pipeline proposed by Eilertsen et al. [16]. For the non-reference metric, we use an extended set of 100 LDR images obtained from the Internet which we use directly as input with 15% to 45% of the pixels saturated. For fairness in these comparisons, both sets are composed of landscape images, since fully-supervised approaches are typically trained with datasets mainly containing this type of content. We show in Table 4 the results of these metrics and in Fig. 7 visual comparisons for our five datasets. Note that, since our method focuses on hallucinating content in completely saturated regions, it is highly unlikely that this content fully matches that of the original ground truth image, therefore full-reference metrics are not well suited for assessing the quality of our reconstructions. Nevertheless, GlowGAN-ITM is on par with fully-supervised approaches for the reference metrics, while it excels in the non-reference metric, showing that our hallucinated content is more plausible in terms of naturalness. Additionally, previous methods can generate only one potential reconstruction given an input LDR image, while our approach allows the generation of multiple results with different but plausible semantic information, as Fig. 8 shows.

**User Study.** We perform a subjective study to further assess the quality of our generated results for the ITM application. We include 20 scenes (four for each of our five datasets). For each scene, HDR results obtained with each of the seven methods were shown on a single screen, and participants were asked to rank the seven images from 1 (most preferred) to 7 (least preferred). The presentation order of the scenes and methods was randomized. The images were displayed in an HDR display Dell UP3221Q (3840×2160 resolution) in a standard office room with natural illumination, and participants sat at a distance of 0.5 meters from the display. A total of 24 participants (38% female, aged 22 to 37 years old with normal or corrected-to-normal vision) participated in the study. We show in Fig. 9 the preference rankings for each method, aggregated for all scenes and participants. For the analysis, we use pairwise Kruskal-Wallis tests adjusted by Bonferroni correction for multiple comparisons since the rankings do not follow a normal distribution. Results reveal that our method was ranked significantly higher than all others ($p < 0.001$), and it was selected as the top performing method in over 80% of the trials.

**5. Conclusion & Discussion**

We have introduced GlowGAN, a novel paradigm for learning HDR imagery from LDR data. Our method is orthogonal to other advances in generative adversarial
While this assumption is naturally satisfied for in-the-wild photo datasets, the dynamic range we can obtain is tightly linked to the exposure variance in the dataset. We hope that our approach inspires future work on learning rich models from casually captured images.

Acknowledgements

Ana Serrano received funding from the Spanish Agencia Estatal de Investigación (project PID2022-141539NB-I00). This work was supported by an academic gift from Meta.

References


Table 4: Evaluation of the ITM application. We achieve on-par quality with the best-performing supervised methods in reference metrics while outperforming other approaches in the non-reference metric that evaluates the naturalness and quality of the image. Note that reference metrics are not best suited for this evaluation, since the strength of our method lies in the reconstruction of overexposed regions in which the hallucinated content does not match that of the original HDR scene.

10518


