Zero-Shot Structure-Preserving Diffusion Model for High Dynamic Range Tone Mapping (Supplementary Material)

Ruoxi Zhu[1,*], Shusong Xu[2,3,*], Peiye Liu[2,3], Sicheng Li[2,3], Yanheng Lu[2,3],
Dimin Niu[2,3], Zihao Liu[2,3], Zihao Meng[2,3], Zhiyong Li[2,3], Xinhua Chen[1], Yibo Fan[1,†]
[1]Fudan University, [2]DAMO Academy, Alibaba Group, [3]Hupan Lab

In this supplementary material, we first provide information about the dataset we used and our training procedure in Sec. 1. Then, we detail our inference and post-processing pipeline in Sec. 2. The proposed structure refinement operation is discussed under the framework of DDIM in Sec. 3. Finally, more experimental results are presented in Sec. 4.

1. Datasets and Training Details

We trained our model on Flickr2K [6] dataset. It contains 2650 high-quality LDR images of 2K resolution taken in different scenes. Before training, we pre-calculate the MSCN maps and Gaussian-blurred luma maps of these images. When calculating the MSCN maps, we use the default settings recommended by [7]. To prepare the training patches during the training phase, the MSCN maps, Gaussian-blurred luma maps and target ground truth of the same image are cropped into 640 × 640 patches in a randomly selected position.

In the first training stage that contains 200 epochs, the structure refinement operation is not included in the diffusion process, and we only train the implicit control branch using the loss function same as [9, 13], which is formulated as:

$$\mathcal{L} = \mathbb{E}_{z_0,t,c,\varepsilon\sim\mathcal{N}(0,1)} \|\varepsilon - \varepsilon_\theta(z_t, t, c)\|_2^2. \tag{1}$$

The batch size is 8 and the learning rate is $1e^{-5}$. In the second training stage that contains another 200 epochs, the decoding-encoding process is included in the pipeline, and the encoder is trained together with the implicit control branch using the same loss function as Eqn.1. We expect to fine-tine the implicit control branch as well as acquire an encoder that can convert an image into the latent space with as little loss as possible in the second training stage.

2. Inference Details

We test our model on HDRPS dataset [2], a benchmark testset consisting of 105 HDR images taken in different scenes and under various exposure conditions. We first calculate the MSCN maps and Gaussian-blurred luma maps using the approach introduced in the main text. Afterward, these maps are divided into 640 × 640 overlap patches with a stride of 320. Patches at different positions are fed to the diffusion model respectively to generate different results. The overlapped parts in the output image are calculated using a weighted-sum strategy. For example, the first patch lies between 0 and 640 in the x direction and the second patch lies between 320 and 960. Hence, for pixels lying between 320 and 640, their values are calculated as :

$$\boldsymbol{I}(i,j) = \frac{640-i}{320}\boldsymbol{I}_1(i,j) + \frac{i-320}{320}\boldsymbol{I}_2(i,j-320), \tag{2}$$

where $\boldsymbol{I}$, $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ denote the resulting image, the first patch and the second patch respectively, i and j denote the indexes of coordinates. Patches in other positions are weighted and summed in a similar way.

After the whole image is generated, we reconstruct the color by:

$$\boldsymbol{C}_{out}^{(i)} = \boldsymbol{Y}_{pred}(\boldsymbol{C}_{in}^{(i)}/\boldsymbol{Y})^s, \tag{3}$$

where $i \in \{R, G, B\}$ denotes the index of the color channel, $\boldsymbol{Y}$ and $\boldsymbol{Y}_{pred}$ are the Y channel of the original HDR image and the output of the diffusion model respectively. The parameter s is set as 0.5 according to [11].

This work was supported in part by the National Key R&D Program of China (2023YFB4502802), in part by the National Natural Science Foundation of China (62031009), in part by the "Ling Yan" Program for Tackling Key Problems in Zhejiang Province (No.2022C01098), in part by Alibaba Innovative Research (AIR) Program, in part by Alibaba Research Fellow (ARF) Program, in part by the Fudan-ZTE Joint Lab, in part by CCF-Alibaba Innovative Research Fund For Young Scholars. [*]contribute equally. [†]Yibo Fan is the corresponding author.

Algorithm 1 Structure-Preserving Sampling Process Based on DDIM

```
$\boldsymbol{\mu}_{ori}, \boldsymbol{\sigma}_{ori}, \hat{\boldsymbol{I}}_{ori} = TSD(\boldsymbol{I}_{ori})$
for $t = T, T-1, ..., 1$ do
    $z'^{(t-1)}_{pred0} = \frac{z_t - \sqrt{1-\overline{\alpha}_t}\varepsilon_\theta(z_t,t,c)}{\sqrt{\overline{\alpha}_t}}$
    if $t > t_0$ then
        $\boldsymbol{\mu}_{t-1}, \boldsymbol{\sigma}_{t-1}, \hat{\boldsymbol{I}}_{t-1} = TSD(\mathcal{D}(z'^{(t-1)}_{pred0}))$
        $z^{(t-1)}_{pred0} = \mathcal{E}(\boldsymbol{\mu}_{t-1} + \gamma\boldsymbol{\sigma}_{t-1}\hat{\boldsymbol{I}}_{ori})$
    else
        $z^{(t-1)}_{pred0} = z'^{(t-1)}_{pred0}$
    end if
    $\varepsilon = \frac{z_t - \sqrt{\overline{\alpha}_t} z^{(t-1)}_{pred0}}{\sqrt{1-\overline{\alpha}_t}}$
    $z_{t-1} = \sqrt{\overline{\alpha}_{t-1}} z^{(t-1)}_{pred0} + \sqrt{1-\overline{\alpha}_{t-1}}\varepsilon$
end for
$\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, \hat{\boldsymbol{I}}_0 = TSD(\mathcal{D}(z_0))$
$\boldsymbol{I}_{pred} = \boldsymbol{\mu}_0 + \gamma\boldsymbol{\sigma}_0\hat{\boldsymbol{I}}_{ori}$
return $\boldsymbol{I}_{pred}$
```

3. Structure Refinement Operation with DDIM

In the main text, we describe the structure refinement operation based on DDPM reverse sampling process [4] for simplicity. Whereas we recommend to adopt DDIM [10] in the real practice in that it significantly reduces the iteration steps for better efficiency. In this case, our structure refinement operation is slightly modified to adapt to the formula of DDIM sampling, which is given in Alg.1. Empirically, the hyperparameter t_0 is set to be 10, the total step T is set to be 20 and γ is set to be 2 for edge enhancement before t reaches t_0.

4. Additional Results

Due to the limited space of the main paper, here we illustrate more experimental results on the datasets mentioned in the main text.

4.1. Results on HDRPS Dataset

Fig.1 shows the comparisons of our model with previous SOTA algorithms on HDRPS dataset [2]. It can be seen that our model generates more details than other methods, especially in bright regions, producing visually pleasing images.

4.2. Results on HDR+ Dataset

We also test our algorithm and previous methods on some difficult cases in HDR+ [3] dataset, which contain extremely high exposure contrast. The results are shown in Fig.2. Other methods tend to produce obvious banding artifacts, while our method is still able to yield naturally-look images.

4.3. Results on VIS-NIR Scene Dataset

As discussed in the main text, our model can be applied to the task of VIS-NIR image fusion without training again. More results on VIS-NIR Scene dataset [1] are illustrated in Fig.3, which shows that our model enhances the image details better than other methods.

4.4. User Study

Considering that there is no ground truth for tone mapping, we also conducted a user study to get the mean opinion scores (MOS) of different algorithms. Twenty images from HDRPS dataset [2] are selected and tone-mapped by four algorithms respectively. We invited 30 volunteers for the user study. They were asked to give each tone-mapped image a score according to its image quality (0 represents the poorest quality and 5 represents the best quality). As shown in Table 1, the proposed method achieves a higher score than previous methods.

	LA-Net	DeepTMO	Vinker *et al.*	Proposed
MOS (Max: 5)	2.8	3.9	3.0	4.2

Table 1. Mean of scores of different methods in user study.

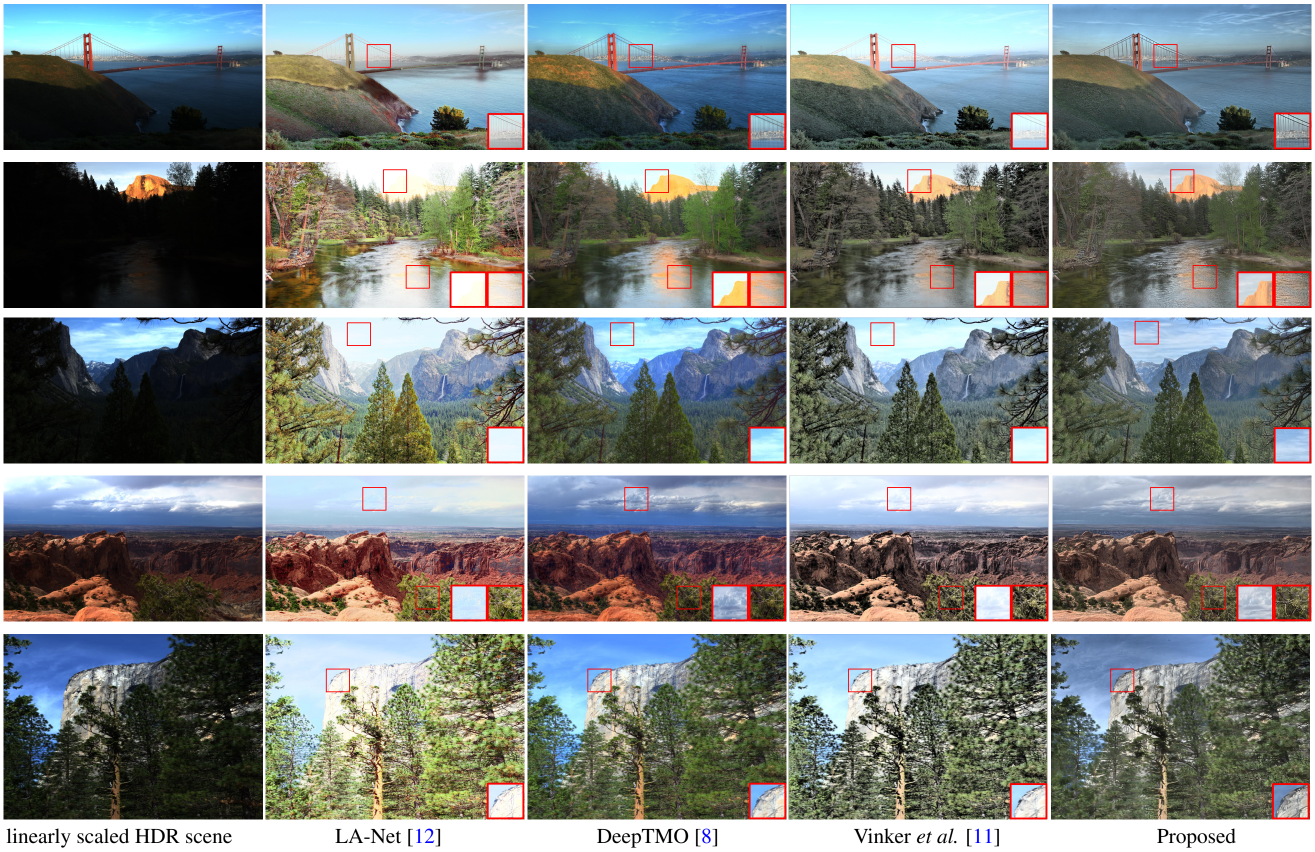

Figure 1. Visual comparisons on HDRPS dataset [2]. Our method performs best in both bright regions and dark regions.

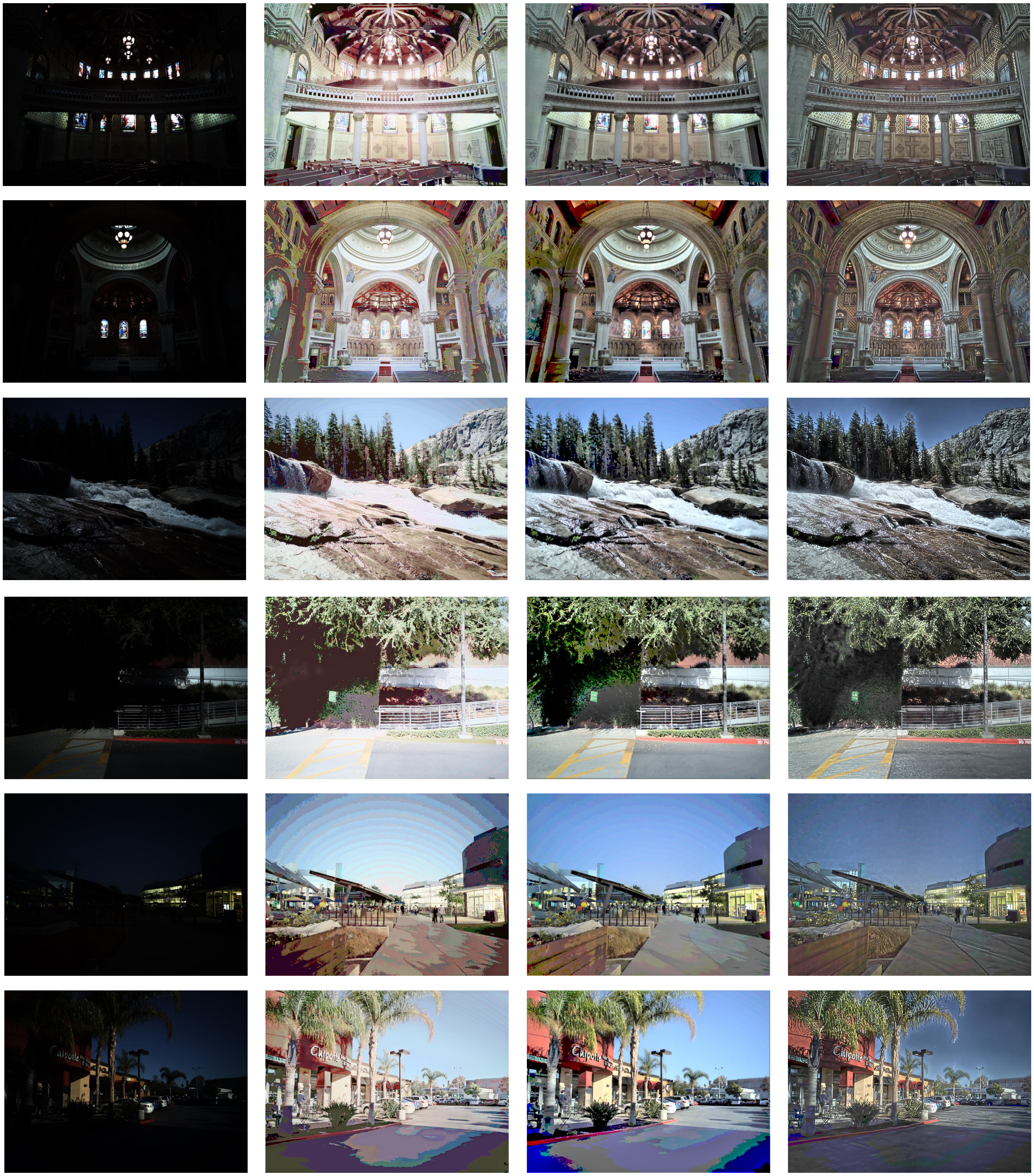

Figure 2. Visual comparisons on HDR+ dataset. Other methods produce much more banding artifacts than ours.

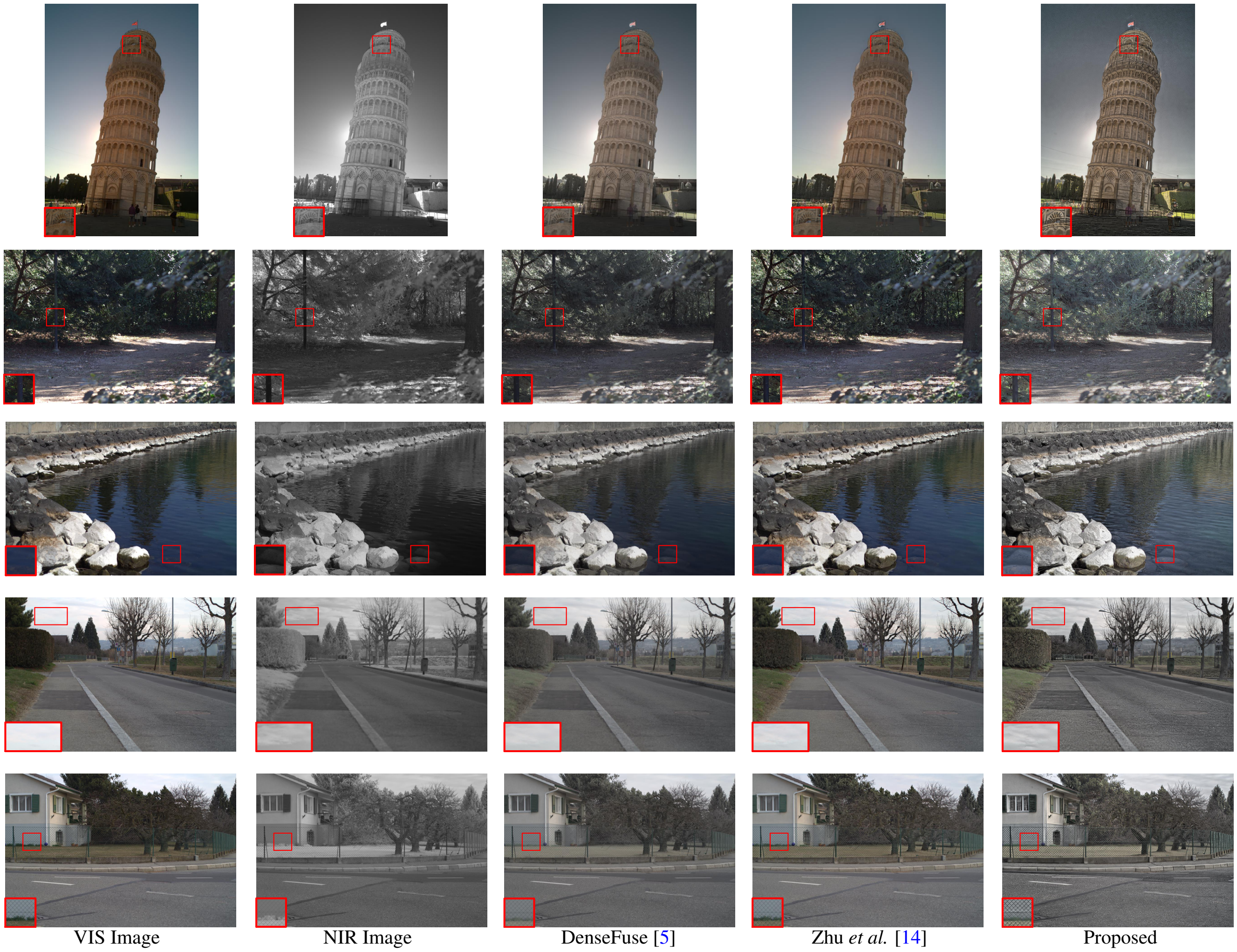

Figure 3. Comparisons of different VIS-NIR fusion methods on VIS-NIR Scene dataset [1]. Our method produces highly detailed images.

References

[1] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR) 2011*, pages 177–184, 2011. 2, 5

[2] M.D. Fairchild. The hdr photographic survey. *Color and Imaging Conference*, 15:233–238, 2007. 1, 2, 3

[3] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 35(6), 2016. 2

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[5] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. 5

[6] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017. 1

[7] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21 (12):4695–4708, 2012. 1

[8] Aakanksha Rana, Praveer Singh, Giuseppe Valenzise, Frederic Dufaux, Nikos Komodakis, and Aljosa Smolic. Deep tone mapping operator for high dynamic range images. *IEEE Transactions on Image Processing*, 29:1285–1298, 2020. 3

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2022. 2

[11] Yael Vinker, Inbar Huberman-Spiegelglas, and Raanan Fattal. Unpaired learning for high dynamic range image tone mapping. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14637–14646, 2021. 1, 3, 4

[12] Kai-Fu Yang, Cheng Cheng, Shi-Xuan Zhao, Hong-Mei Yan, Xian-Shi Zhang, and Yong-Jie Li. Learning to adapt to light. *Int. J. Comput. Vision*, 131(4):1022–1041, 2023. 3, 4

[13] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[14] Ruoxi Zhu, Yi Ling, Xiankui Xiong, Dong Xu, Xuanpeng Zhu, and Yibo Fan. Luminance-preserving visible and near-infrared image fusion network with edge guidance. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1155–1159, 2023. 5