Supplementary Material: Selective TransHDR: Transformer-based selective HDR Imaging using Ghost Region Mask

Jou Won Song*[1], Ye-In Park*[1], Kyeongbo Kong[2], Jaeho Kwak[1], and Suk-Ju Kang[1]

[1] Department of Electronic Engineering, Sogang University, Seoul, Korea
{wn5649,yipark06,resky1111,sjkang}@sogang.ac.kr
[2] Department of Media communication, Pukyong National University, Busan, Korea
kbkong@pknu.ac.kr

1 Additional Ablation Study

To verify the effectiveness of the transformer module, we compare the proposed model and the base model (only CNN) in the ghost region. As shown in Table 1, the proposed method with the transformer module has higher performance in the ghost region than using only the CNN.

2 Experiments for Ghost Region Mask Generation

We show more experimental results which could not be included in the main manuscript due to the lack of space.

2.1 Analysis of Ghost Region Mask Threshold

In this section, we show the change in performance according to the threshold variation of the ghost region mask mentioned in the main paper. As discussed in Section 3.1 (in the main paper), we empirically set 0.7 as our threshold for ghost region mask generation. To show the change in performance for the variations in threshold values, we choose it from 0.5 to 0.9 as our threshold on the ghost region mask. As the threshold is higher, the smaller ghost area is detected, and it means that the input image is more affected by CNN. As shown in Table 2, the proposed model shows the highest performance when the threshold is 0.7. Table 2 also shows that the performance of the proposed model decreases when the threshold is lowered and the transformer structure is used in the non-ghost area. Therefore, we plan to apply a novel deep learning model based on the end-to-end approach to increase the accuracy of ghost region detection in future work.

* equal contribution

Fig. 1. Comparison of qualitative results of the only CNN and the proposed method.

Method	PSNR-μ	PSNR-L
Base model (only CNN)	40.0271	36.8335
Proposed model	**41.6714**	**38.4885**

Table 1. Performance comparison of the only CNN and the proposed method using PSNR-μ and PSNR-L in the ghost region.

2.2 Analysis of Dilation Kernel Size

As mentioned in the ghost region mask generation in Section 3.1, to compensate for the undetected ghost regions and train various ghost masks, the kernel size of the dilation is set to the value between 11 and 17 in the opening operation by our experiments. In this additional experiment, we measure the performance by changing the kernel size of the dilation used in the test process to verify the change in performance according to the kernel size of the dilation. As shown in Table 3, when the kernel size of dilation is 15, the proposed model shows the highest performance. However, we note that this experiment does not imply that the proposed model is sensitive to the variations in kernel size. As a result of the experiment, the proposed model does not significantly drop in performance even when the kernel size is changed, and shows robust performance in the ambiguous region between the ghost and the non-ghost region.

3 Implementation details

3.1 Implementation

For training the proposed model, we chose the gradient centralized Adam optimizer [2] with the learning rate of 0.0002. The momentum parameters of $\beta 1$ and $\beta 2$ were set to 0.5 and 0.999, respectively. We trained our model with batch

Threshold	0.5	0.6	0.7	0.8	0.9
PSNR-μ	43.8131	43.8483	44.0981	44.0412	43.9831
PSNR-L	41.5114	41.5322	41.7021	41.6201	41.5912

Table 2. Performance comparison for the variations in threshold values using PSNR-μ and PSNR-L

Kernel size	11	13	15	17	19
PSNR-μ	43.9782	44.0211	44.0981	43.8761	43.8311
PSNR-L	41.5682	41.6313	41.7021	41.5345	41.5314

Table 3. Performance comparison for the variations in dilation kernel size using PSNR-μ and PSNR-L.

size of 8. In our implementation, all convolution layers were used with the ReLU except for the last layer for HDR image reconstruction. DRDBs had a growth rate of 64. The last convolution layer in each DRDB compressed the feature map to the size of the input size by applying one 1×1 convolution. In addition, the patch size of the ghost region mask-guided transformer module is randomly set to a value between 10×10 and 20×20, and the ghost region mask threshold is set randomly to a value between 0.6 and 0.9. Through this, the proposed model that selectively applies the CNN and the transformer module can obtain more robust performance in various cases. We performed training for 800 epochs and reduced them to half at 200, 400, 600, and 750 epochs. We implemented our model using PyTorch [4] on NVIDIA GeForce RTX 3090 GPU. Furthermore, the edge map extraction operation ($M(\cdot)$) was obtained by computing the difference between adjacent pixels:

$$I_x = I(x+1, y) - I(x-1, y), \tag{1}$$

$$I_y = I(x, y+1) - I(x, y-1), \tag{2}$$

$$\Delta I = (I_x, I_y), \tag{3}$$

$$M(I) = (\Delta I)^2, \tag{4}$$

where I stands for the reference image. x and y denote the horizontal and vertical indices, respectively.

4 Details on the Proposed Network Architecture

Table 4 shows sub-network structure of the proposed method. Each network is described by a list of layers including an output shape, a kernel size, a padding size, and a stride. In addition, whether the activation function is applied is described. A description of the entire network architecture using the subnetworks can be found in Table 5.

5 Visualization of Proposed Ghost Region Mask

In Figs. 2 and 3, we visualize the proposed ghost region mask images. It can be seen that the proposed ghost mask succeeds in detecting mismatched regions of the image. In addition, the ghost region masks also show that misaligned regions are caused by a variety of causes, including camera angles and object motion. Therefore, it shows that the proposed method can be well applied to various cases, not just motions.

6 Additional Qualitative Results

In this section, we present additional qualitative results that we did not show in the main paper due to the limited space of paper. In Figs. 4, 5, and 6 show visual results for various motion cases in Kalantari et al.'s [1] dataset and Tursun et al.'s [5] dataset. It can be confirmed that the proposed method achieves better visual results in terms of ghost and stain artifacts and sharpness compared to AHDR [6] and HDRGAN [3].

References

1. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph. **36**(4), 144–1 (2017)
2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
3. Niu, Y., Wu, J., Liu, W., Guo, W., Lau, R.W.: Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. In: IEEE Transactions on Image Processing. vol. 30, pp. 3885–3896 (2021)
4. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
5. Tursun, O.T., Akyüz, A.O., Erdem, A., Erdem, E.: An objective deghosting quality metric for hdr images. In: Computer Graphics Forum. vol. 35, pp. 139–152. Wiley Online Library (2016)
6. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1751–1760 (2019)

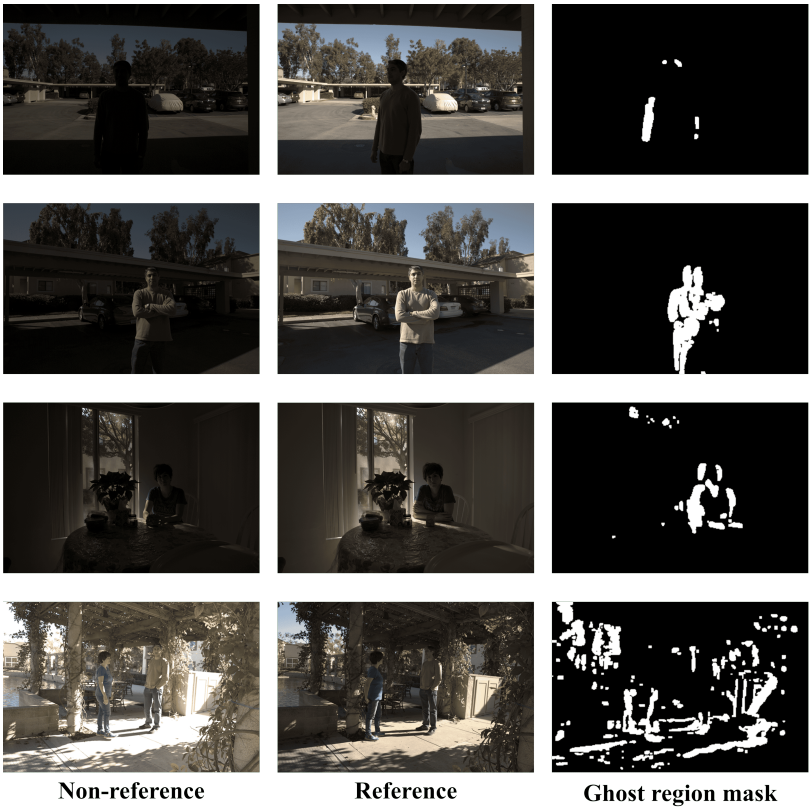

Fig. 2. Visualization of ghost region mask for Karantari et al's [1] dataset.

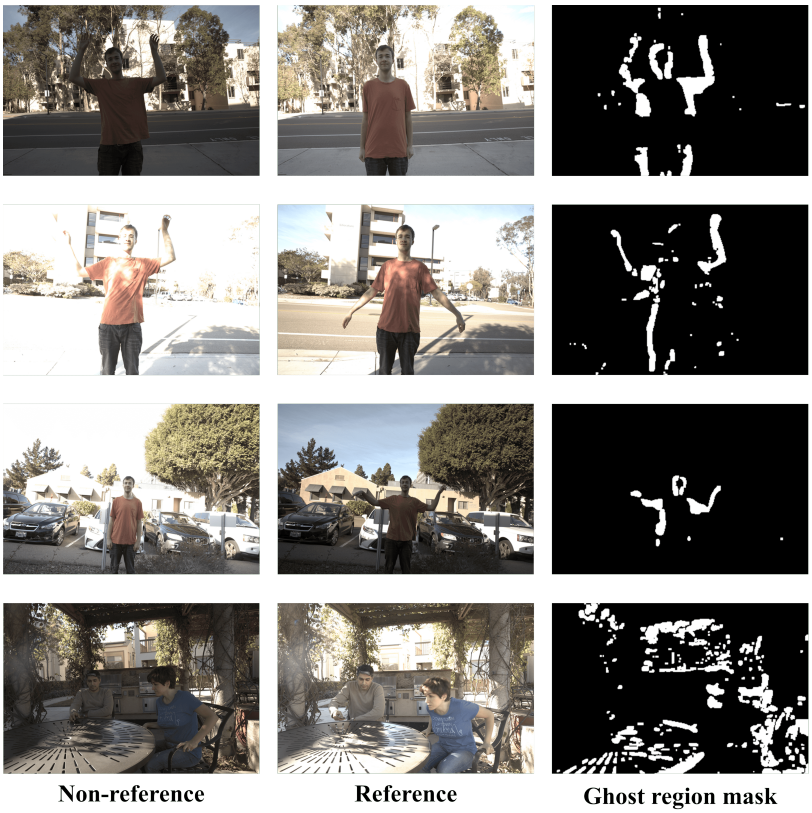

Fig. 3. Visualization of ghost region mask for Karantari et al's [1] dataset.

Network	Layer (activation function)	Output size	Kernel	Stride	Pad
Multi-scale CNN	Conv 1 (ReLU)	W x H x D	1 x 1	1	0
	Conv 2 (ReLU)	W x H x D	3 x 3	1	1
	Conv 3 (ReLU)	W x H x D	5 x 5	1	2
	Concatenate	W x H x 3D	-	-	-
	Conv 4 (ReLU)	W x H x D	3 x 3	1	1
Encoder	Conv 1 (ReLU)	W x H x D	3 x 3	1	1
	Multi-scale CNN	W x H x D	-	-	-
	Multi-scale CNN	W x H x D	-	-	-
	Multi-scale CNN	W x H x D	-	-	-
TM (ghost path)	Conv Q (ReLU)	W x H x D	1 x 1	1	0
	Conv K (ReLU)	W x H x D	1 x 1	1	0
	Conv V (ReLU)	W x H x D	1 x 1	1	0
	Conv 1 (ReLU)	W x H x D	1 x 1	1	0
	Conv 2 (ReLU)	W x H x D	1 x 1	1	0
TM (non-ghost path)	Conv 1 (ReLU)	W x H x D	3 x 3	1	1
	Multi-scale CNN	W x H x D	-	-	-
	Multi-scale CNN	W x H x D	-	-	-
	Multi-scale CNN	W x H x D	-	-	-
TM (output feature)	Conv 1 (ReLU)	W x H x D	3 x 3	1	1
	Multi-scale CNN	W x H x D	-	-	-
Down-sampling	Max pooling	W//2 x H//2 x D	2 x 2	1	0
	Conv 1 (ReLU)	W//2 x H//2 x 2D	3 x 3	1	1
	Multi-scale CNN	W x H x 2D	-	-	-
	Multi-scale CNN	W x H x 2D	-	-	-
Up-sampling	ConvTr	W x H x D	2 x 2	2	0
	Conv 1 (ReLU)	W x H x D	3 x 3	1	1
	Multi-scale CNN	W x H x D	-	-	-
	Multi-scale CNN	W x H x D	-	-	-

Table 4. Architectural details of subnetworks. ConvTr denotes transposed convolution layer, Conv denotes convolution layer, and TM denotes transformer module.

Network	Layer (input name)	Output	Output size	Kernel	Stride	Pad
FE	Encoder 1 (low)	non 1	256 x 256 x 64	3 x 3	1	1
	Encoder 1 (middle)	ref	256 x 256 x 64	3 x 3	1	1
	Encoder 1 (high)	non 2	256 x 256 x 64	3 x 3	1	1
	TM 1 (non 1, ref)	Over	256 x 256 x 64	-	-	-
	TM 2 (non 2, ref)	Under	256 x 256 x 64	-	-	-
	Concatenate (Over, ref, Under)	F1	256 x 256 x 192	-	-	-
	Conv 1 (F1)	F2	256 x 256 x 64	3 x 3	1	1
	Multi-scale CNN 1 (F2)	F3	256 x 256 x 64	-	-	-
	Multi-scale CNN 2 (F3)	F4	256 x 256 x 64	-	-	-
	Multi-scale CNN 3 (F4)	F5	256 x 256 x 64	-	-	-
FS	Down sampling (F5)	F6	128 x 128 x 128	-	-	-
	DRDB 1 (F6)	F7	128 x 128 x 128	-	-	-
	DRDB 2 (F7)	F8	128 x 128 x 128	-	-	-
	DRDB 3 (F8)	F9	128 x 128 x 128	-	-	-
	Up sampling (F9)	F10	256 x 256 x 64	-	-	-
	Conv 1 (middle)	H1	256 x 256 x 64	3 x 3	1	1
	Multi-scale CNN 3 (H1)	H2	256 x 256 x 64	-	-	-
	Multi-scale CNN 4 (H2)	H3	256 x 256 x 64	-	-	-
	Multi-scale CNN 5 (H3)	H4	256 x 256 x 64	-	-	-
	Concatenate (F5, F10, H4)	F11	256 x 256 x 192	-	-	-
	Conv 3 (F11)	F12	256 x 256 x 64	3 x 3	1	1
	Multi-scale CNN 1 (F12)	F13	256 x 256 x 64	-	-	-
	Multi-scale CNN 2 (F13)	F14	256 x 256 x 64	-	-	-
	Multi-scale CNN 3 (F14)	F15	256 x 256 x 64	-	-	-
	Conv 4 (F12)	F13	256 x 256 x 64	3 x 3	1	1
	Conv 5 (F13)	F14	256 x 256 x 3	3 x 3	1	1
	F14 + middle	output	256 x 256 x 3	3 x 3	1	1

Table 5. Architectural details of the proposed method. FE denotes feature extraction network. TM and FS denote transformer module and fusion network, respectively.

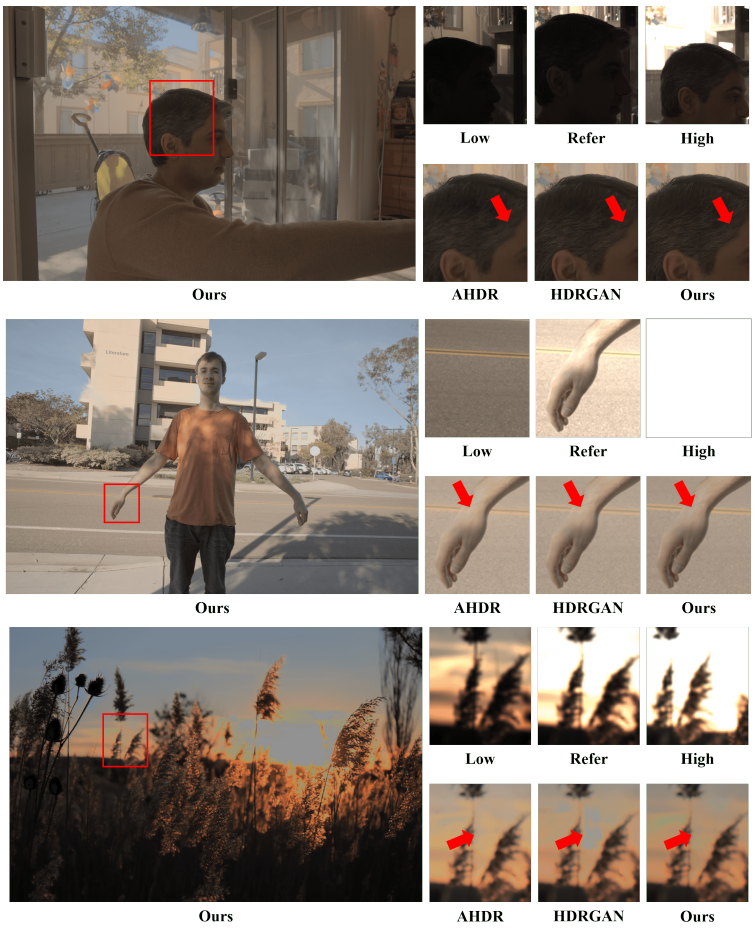

Fig. 4. Comparision of qualitative results for various motion cases on the Kalantati et al.'s [1] dataset (top and middle images) and Tursun et al.'s [5] dataset (bottom image).

Fig. 5. Comparision of qualitative results for various motion cases on the Kalantati et al.'s [1] dataset (top and middle images) and Tursun et al.'s [5] dataset (bottom image).

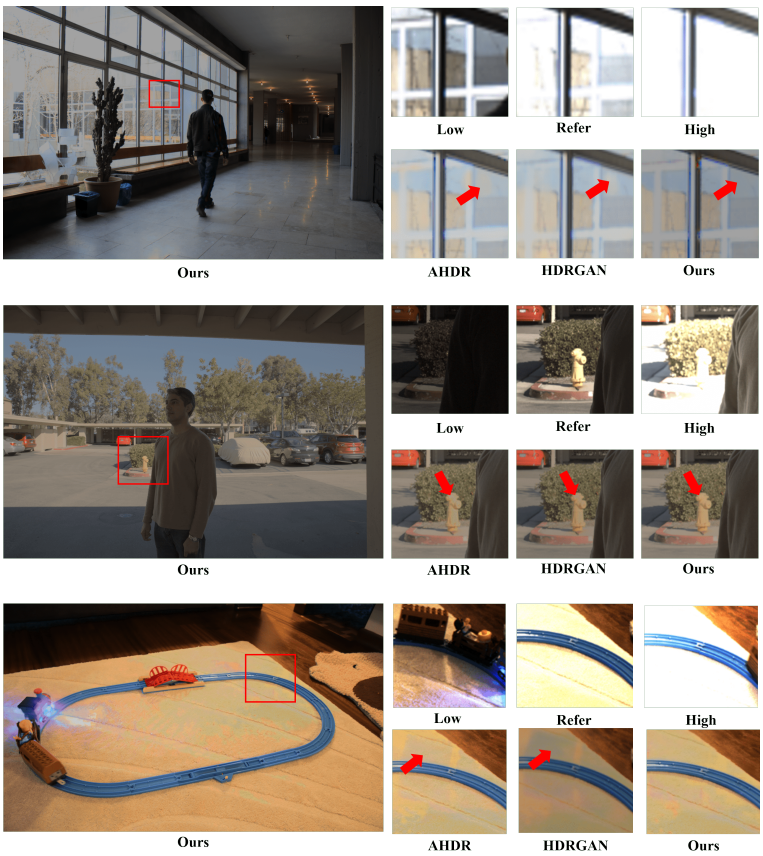

Fig. 6. Comparision of qualitative results for various motion cases on the Kalantati et al.'s [1] dataset (top and middle images) and Tursun et al.'s [5] dataset (bottom image).