SSH: A Self-Supervised Framework for Image Harmonization Supplementary Material

1. Experiment Setting and Network Architecture

We use 256×256 resolution for full-reference metrics evaluation (PSNR, MSE, SSIM, and LPIPS) and use the original resolution when showing the visual comparison for better view. Our network consists of a content network, a reference network, and a fusion network. We use a *ConvBlock* to denote the convolutional-LeakyReLu-BatchNorm block. The detailed structures of the content network, reference network and fusion network are shown in Table. 1, 3, 2.

Layer	Kernel	Stride	C_{in}	C_{out}
ConvBlock	3×3	1	4	32
ConvBlock	3×3	1	32	32
MaxPooling	2×2	2	-	-
ConvBlock	3×3	1	32	64
ConvBlock	3×3	1	64	64
MaxPooling	2×2	2	-	-
ConvBlock	3×3	1	64	128
ConvBlock	3×3	1	128	128
MaxPooling	2×2	2	-	-
ConvBlock	3×3	1	128	256
ConvBlock	3×3	1	256	256
MaxPooling	2×2	2	-	-
ConvBlock	3×3	1	256	512

Table 1: Details of the content network.

Layer	Kernel	Stride	C_{in}	C_{out}
ConvBlock	7×7	1	3	32
ConvBlock	3×3	2	32	64
ConvBlock	3×3	2	64	128
ConvBlock	3×3	2	128	256
ConvBlock	3×3	2	256	512

Table 2: Details of the reference network.

Layer	Kernel	Stride	C_{in}	C_{out}
ConvBlock	3×3	1	1024	512
Upsampling	3×3	2	-	-
ConvBlock	3×3	1	512	256
ConvBlock	3×3	1	256	256
Upsampling	3×3	2	-	-
ConvBlock	3×3	1	256	128
ConvBlock	3×3	1	128	128
Upsampling	3×3	2	-	-
ConvBlock	3×3	1	128	64
ConvBlock	3×3	1	64	64
Upsampling	3×3	2	-	-
ConvBlock	3×3	1	64	32
ConvBlock	3×3	1	32	32
Conv	3×3	1	32	3

Table 3: Details of the fusion network.

2. Visual Comparison

We compare our method with several recent competing methods on the general foreground objects: the photorealistic style transfer method [5], learning-based image harmonization approaches DIH [4], S^2AM [2], DoveNet [1] and show the visual results to demonstrate the effectiveness of our method. Examples are shown in the Fig. 3. The outputs of [5] looks unsatisfied. DIH [4] mostly generates an output image that shows subtle appearance change compared to the unprocessed input. S^2AM [2] and DoveNet [1] easily capture the incorrect appearance information (brightness, color, and contrast) due to their weak data augmentation strategy. Benefited from the strong data augmentation and the superiority of the self-supervised framework, the outputs of the proposed method SSH are consistently better than others and are closer to the human annotated results.

3. Human Subjects Evaluation

We conduct human subjects evaluation to compare SSH with other state-of-the-art methods. We randomly select 15 foreground and background pairs from the **RealHM** benchmark. Each image is first processed by five methods

WCT^2 [5], DIH [4], S^2AM [2], DoveNet [1], and SSH), and then displayed on a screen for comparing. The 15 examples are shown in the Fig. 4 5.

4. Extension Task

Besides image harmonization, we also try to adopt the proposed SSH method on other related task. Here we show the results on the sky replacement application, where the object is considered as the foreground image and the sky is adopted as the background reference. As shown in Fig. 1, our method generates pleasing results when the sky is replaced, due to the robust generalibility of the self-supervised framework.

DC SSH

Figure 1: **Visual Results of Sky Replacement.** The first column represents the direct composite results and the second column shows the results generated from SSH by using the replaced sky as the background reference. Best viewed zoomed in.

5. Visual Comparison for Ablation Study

Besides the basic harmonization loss L_{harm}, we design the reconstruction loss L_{recon} and disentanglement loss L_{dis} to further improve the performance of the proposed methods, as discussed in our main manuscript. The reconstruction loss L_{recon} helps the reference network to capture the correct appearance of the content input itself and the disentanglement loss L_{dis} help to disentangle the appearance representation and content representation from the given input. During the experiments, we discover that the reconstruction loss and disentanglement loss help the framework to extract the correct appearance from the images and also stabilize the testing performance. We show the typical failure case when these two loss is separately removed, as in the Fig. 6. Either removing the reconstruction loss or the disentanglement loss causes instability when there is a drastic difference in appearance between the background and the foreground.

Figure 2: **Visualization of t-SNE**.

6. Understanding the Representation Learned from Self-Supervision

To further understand the representation learned by the proposed SSH method, we visualized the feature embedding learned by the reference network. Specifically, we download the real-world images form the Internet and then resize them to 256×256 resolution. We use a pre-trained reference network to extract the appearance feature of these examples. After that, the t-SNE [3] algorithm is adopted to analyze the distance between these feature representations by visualizing the clustering. The visual output is shown in Fig. 2. As we can see, images with similar color/brightness are well clustered, which further demonstrate the effectiveness and interpretability of our method.

References

[1] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8394–8403, 2020. 1, 2, 3, 4, 5

[2] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, 29:4759–4771, 2020. 1, 2, 3, 4, 5

[3] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 2

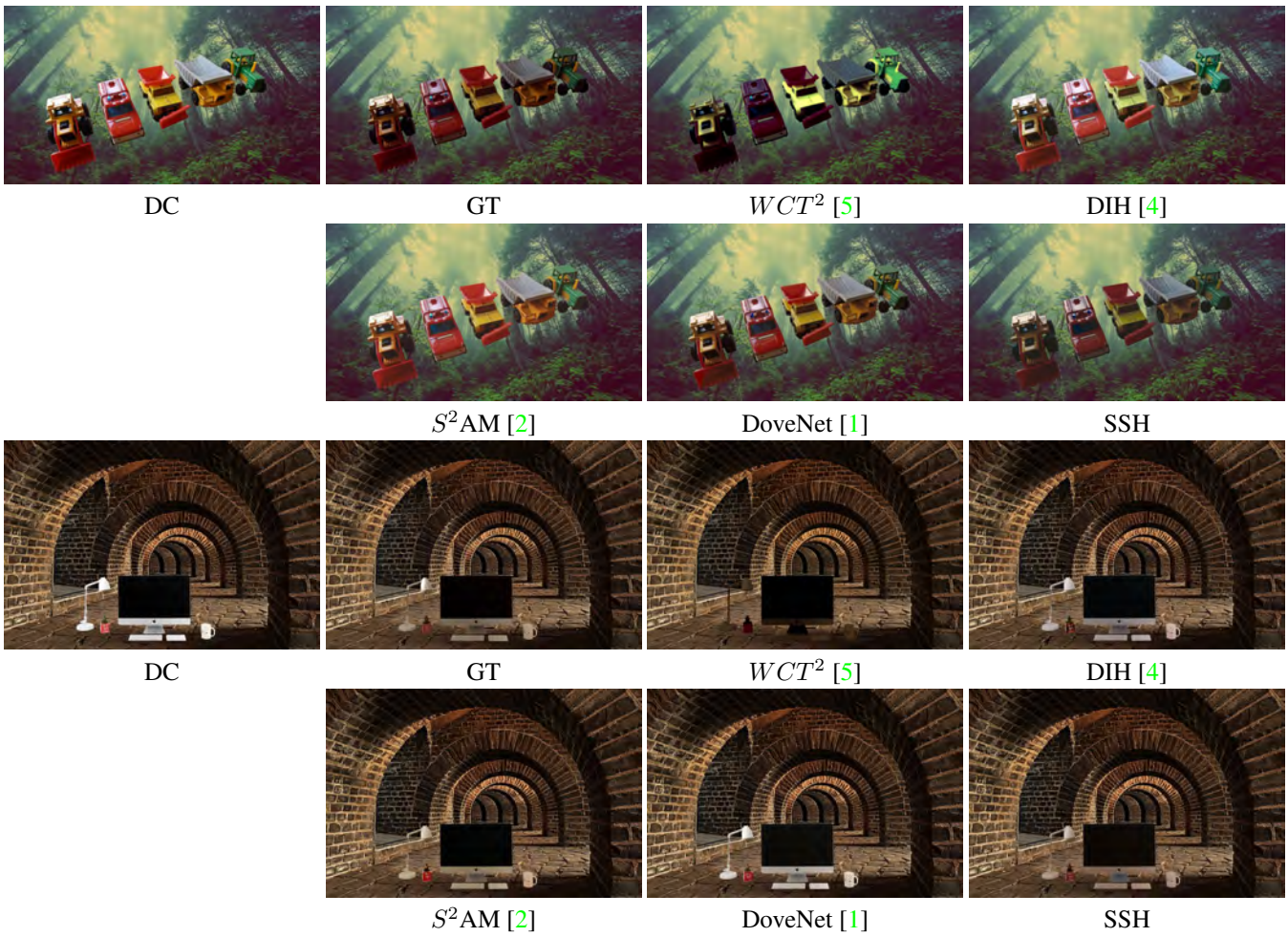

Figure 3: **Comparison with the State-of-the-art Methods.** Best viewed zoomed in.

[4] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797, 2017. 1, 2, 3, 4, 5

[5] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019. 1, 2, 3, 4, 5

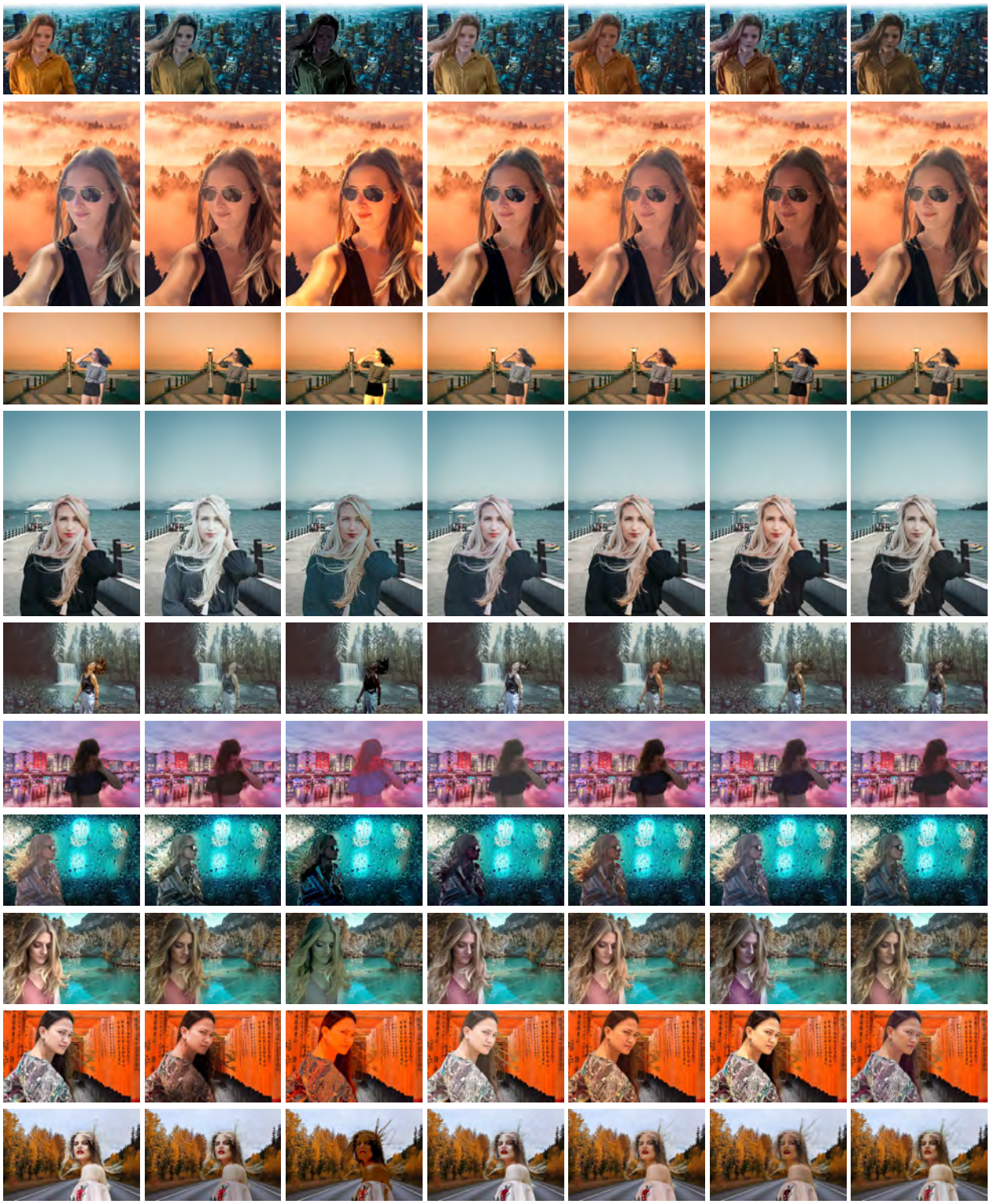

Figure 4: **Examples for Human Subjects Evaluation.**

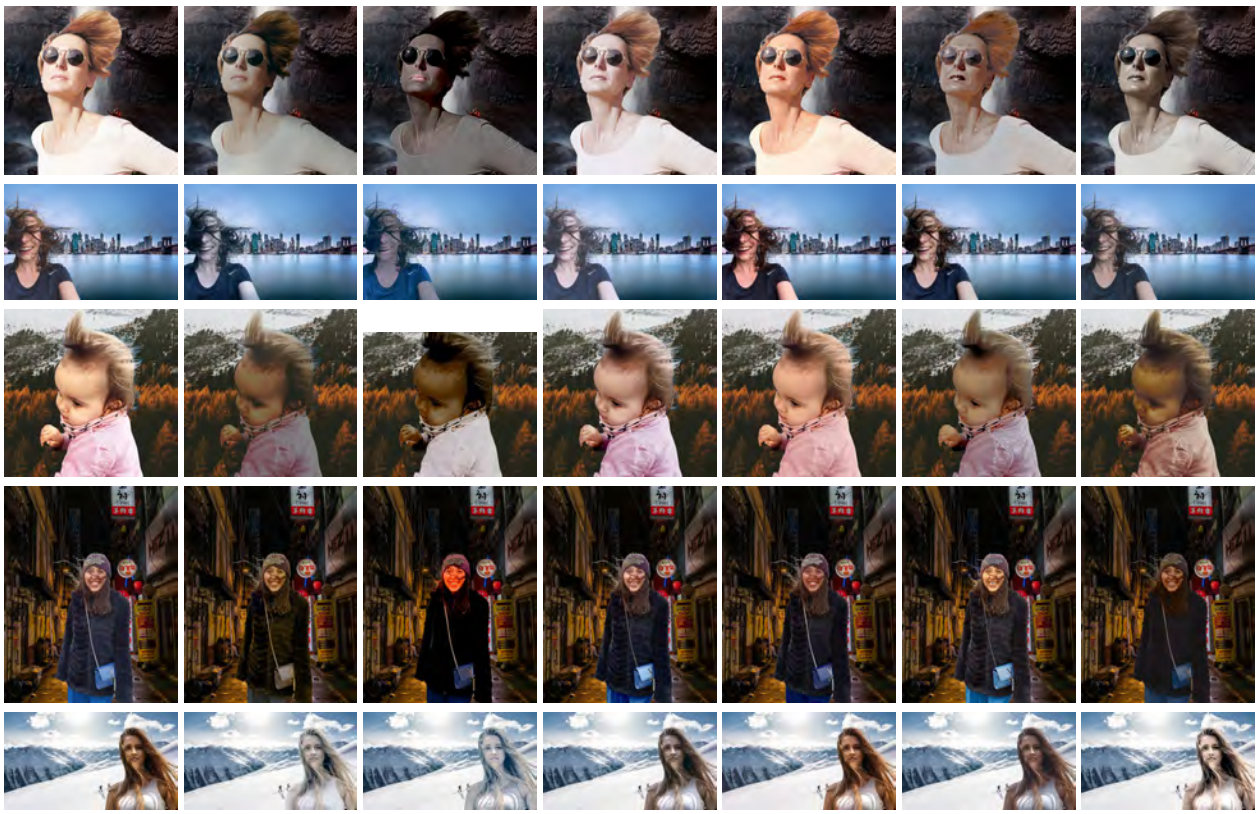

Figure 5: **Examples for Human Subjects Evaluation.**

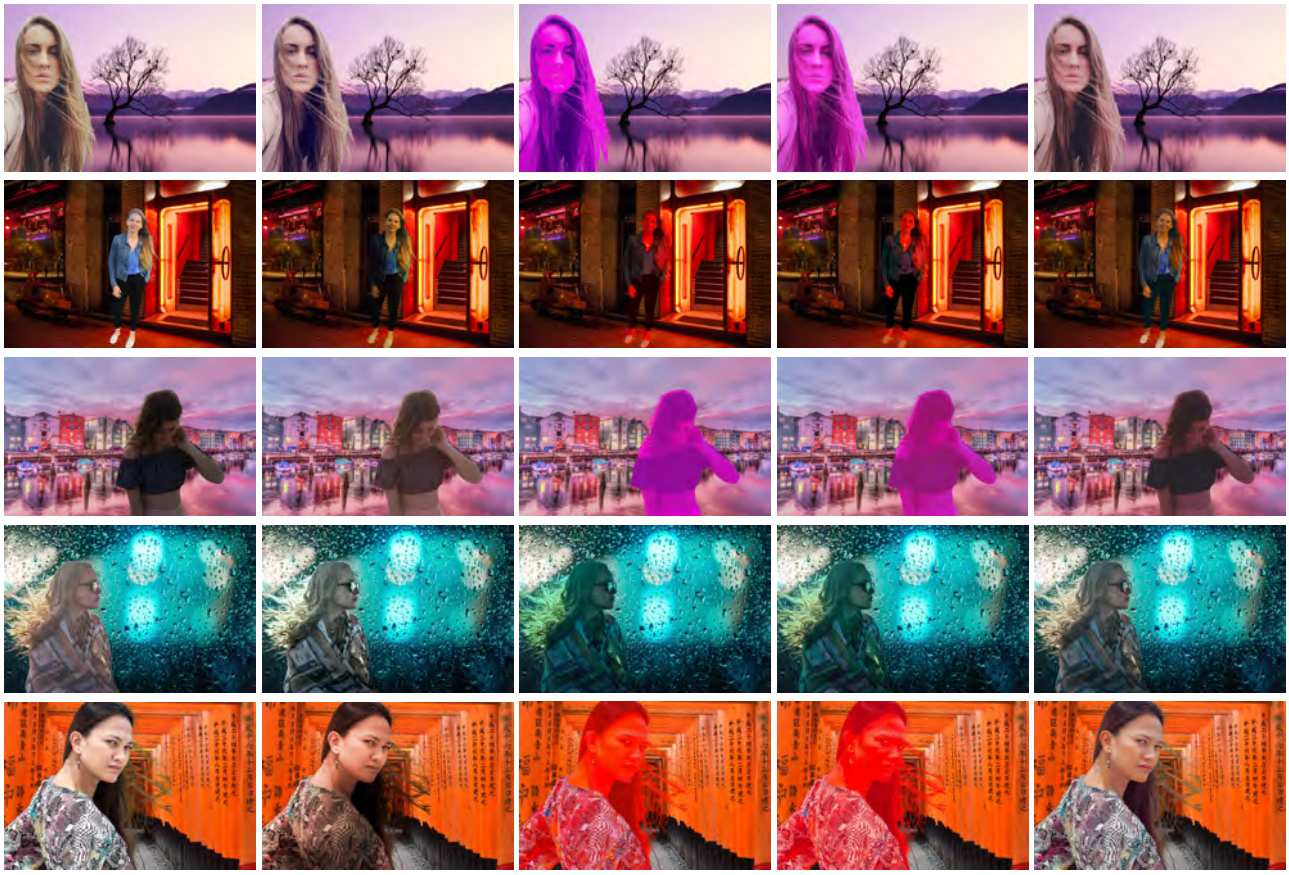

Figure 6: **Visual Results of Ablation Study.** "w/o recon" and "w/o disentangle" represents the results generated by SSH without reconstruction loss and SSH without disentanglement loss respectively. Best viewed zoomed in.