Hist2Style: Histogram-Guided Stylization with Bilateral Grids

Supplementary Material

3. Extended Methods

3.1. Selective Distillation of Image Editing Models

We pick FLUX.1 Kontext [dev] as the large editing model, since it is an open-weights model [8], and therefore can be run locally for cost-efficiency. We run the large language model (LLM) that generated the style names and descriptions only once. We found Google Gemini to perform well for this purpose [11].

Due to resource constraints, our initial generation phase concluded with 1.7 million edited images. To address distortions in the FLUX-edited outputs, we apply automated data filtering. Each edit is scored for cosine similarity in the VGG19 feature space; filtering out images that score below 0.5 results in a refined, final dataset of 1.1 million valid images.

3.2. Interactive Histogram Manipulation

We provide an interface for interactively editing stylizations produced by Hist2Style (see Supp. Fig. 1). After selecting a content image, users can modify the target histograms either by directly sculpting their shapes or by adjusting familiar slider controls. These edits specify **global** color adjustments, which are then fed to the model. Hist2Style interprets these global histogram modifications to produce **local**, spatially coherent edits to the content image.

3.3. Implementation

When applying a bilateral grid to images whose resolution is much higher than the resolution used for training (*e.g.*, 4096×4096 at test time vs 256×256 at training time), there may be visible grid artifacts (aliasing introduced by trilinear interpolation) [15]. Previous work ameliorates this problem via soft regularization, encouraging the grid coefficients to be smooth [15]. We find soft regularization to be hyperparameter-sensitive, and instead blur the grid spatially via a Gaussian kernel of $\sigma = \frac{5}{6}$ pixels.

A mathematically sound but less efficient alternative is to first resample the grid with a proper antialiasing filter to full spatial resolution ($8 \times H \times W$). This results in an 8-bin lookup table at each pixel which can be applied using linear interpolation (on the luminance axis).

4. Extended Results

We denote PhotoWCT^2 [5] as PWCT^2 for compactness. Content images are resized to 2MP, and style images to around 0.1MP. Supp. Fig. 2 shows metric alignment with human preference.

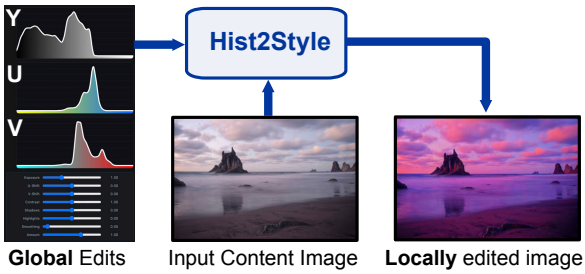

Figure 1. **Image editing.** When the user specifies **global** histogram modifications in the interactive interface, Hist2Style converts it into **local** image edits.

Table 1. Peak memory (GB) for different models across resolutions. OOM stands for out of memory.

Method	256^2	512^2	1024^2	2048^2	4096^2
Hist2Style	0.09	0.2	0.3	1	5
IDT	**0.0004**	**0.0009**	**0.003**	**0.01**	**0.05**
DLUT	0.1	0.1	*0.1*	*0.2*	*0.7*
SALUT	*0.08*	*0.09*	*0.1*	0.3	0.8
Xia *et al.*	0.2	0.2	0.2	0.6	2
ReHistoGAN	0.84	0.99	1.62	4.13	14.17
PWCT^2	1	3	5	9	30
WCT^2	0.5	1	5	20	OOM

Algorithm 1 Marginal distribution matching loss.

Input: flattened image features $X, Y \in \mathbb{R}^{HW \times C}$
Output: 1D Wasserstein loss L
for each channel $c \in \{1, ..., C\}$ **do**
    $X_c = X[:, c]$, $Y_c = Y[:, c] \in \mathbb{R}^{HW}$
    Sort: $x_c \leftarrow \text{sort}(X_c)$, $y_c \leftarrow \text{sort}(Y_c)$
    Compute squared distance: $d_c = \|x_c - y_c\|^2$
end for
Return: mean loss $L = \frac{1}{C} \sum_{c=1}^{C} d_c$

4.1. Peak Memory Usage

We report peak GPU memory usage for all methods in Supp. Tab. 1. Approaches based on global transforms, such as IDT [9, 10], are the most memory-efficient at high resolutions. Methods that rely on spatially varying transformations typically require more memory [5, 16]. Within this landscape, our method remains lightweight, requiring about 1 GB for 4 MP images and 5 GB for 16 MP, placing it well within the practical range for high-resolution stylization workloads.

Table 2. Quantitative ablation study of style transfer components. Lower is better (↓) for all metrics. The best performing method in each column is **bolded**, and the second best is *italicized*.

Method	Components						Color Matching ↓			Cycle Cons. ↓	FID ↓
	4x Larger Grid	4x Larger Image	Moment Matching	Joint Training	Robust Inference	Color Space Loss	Y'	Cb	Cr		
Hist2Style	×	×	×	×	×	×	**387.30**	*49.85*	**80.53**	**401.92**	71.44
4x Larger Grid	✓	×	×	×	×	×	591.94	51.40	93.82	693.84	82.63
4x Larger Image	×	✓	×	×	×	×	519.38	53.98	97.08	693.80	83.97
4x Larger (Image + Grid)	✓	✓	×	×	×	×	526.46	53.19	95.93	692.83	84.30
Moment Matching Loss	×	×	✓	×	×	×	494.24	**47.76**	*81.98*	*496.46*	76.91
4x Larger Grid + Moment Matching Loss	✓	×	✓	×	×	×	*478.85*	50.89	84.43	525.38	78.95
4x Larger Image + Moment Matching Loss	×	✓	✓	×	×	×	632.68	55.80	103.30	879.99	90.48
4x Larger (Image + Grid) + Moment Matching Loss	✓	✓	✓	×	×	×	542.32	53.75	94.38	816.55	92.97
Joint Training	×	×	✓	✓	×	×	534.31	56.81	99.86	651.90	80.19
Robust	×	×	✓	✓	✓	×	2344.75	165.20	260.05	1071.22	*51.69*
Color Space Loss	×	×	×	✓	×	✓	2906.77	215.14	290.54	717.59	**46.69**

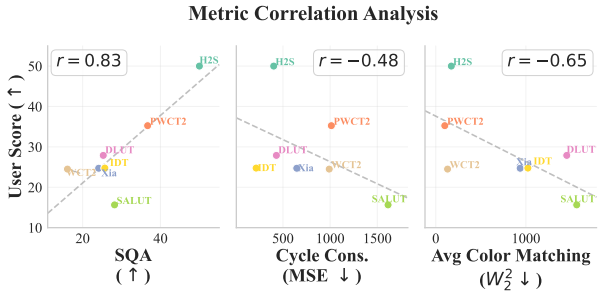

Figure 2. **Correlation plots** between user score and other metrics. Our SQA metric is more strongly aligned to human preference than the other metrics.

4.2. Runtime Measurements

To evaluate runtime performance, we test each algorithm sequentially on individual images. For each evaluation, we perform 10 warmup iterations followed by 100 timed iterations, reporting the average execution time. The recorded times exclude initial image loading when supported but include other preprocessing steps such as histogram computation.

4.3. Ablations and Extensions

For example images from our ablations, see Supp. Fig. 3.

Robust and Joint Training. An alternative method is to match the model's training conditions to test time by conditioning only on a target style image's histograms rather than ground-truth ones. Our synthetic pipeline makes this feasible because it provides coherent style variants for each content image. In practice, however, models trained this way struggle to match colors accurately, which we attribute to increased ambiguity in determining the appropriate color assignments. We also train models that have a shared joint core but different output heads for robust and standard training, but we did not observe improvements.

Bilateral Grid Size. The larger the bilateral grid, the more spatially expressive the model can be. However, similarly to Xia *et al*. [15], we find that bilateral grid size 16×16 is sufficiently expressive for photorealistic style transfer.

Image Processing Size. In principle, providing the image encoder of the model with higher resolution enables access to finer details about the image. However, in our experiments, we did not observe much improvement beyond 64×64.

Moment Matching Loss. We compare using a moment-matching loss (MSE on mean and standard deviation) versus using our chosen distribution loss from Algorithm 1. There is little difference between the two, but Algorithm 1 shows a slight improvement in color matching, which we attribute to it being more information dense. We also observe that reducing the weight of the distribution loss degrades performance, showing the importance of this loss.

Color Space Loss. Since our training data is mined from a large image editing model [8], it is susceptible to imperfect pixel alignment. Therefore, we find that using a perceptual loss often improves the content and style fidelity of the results over using a loss in image color space.

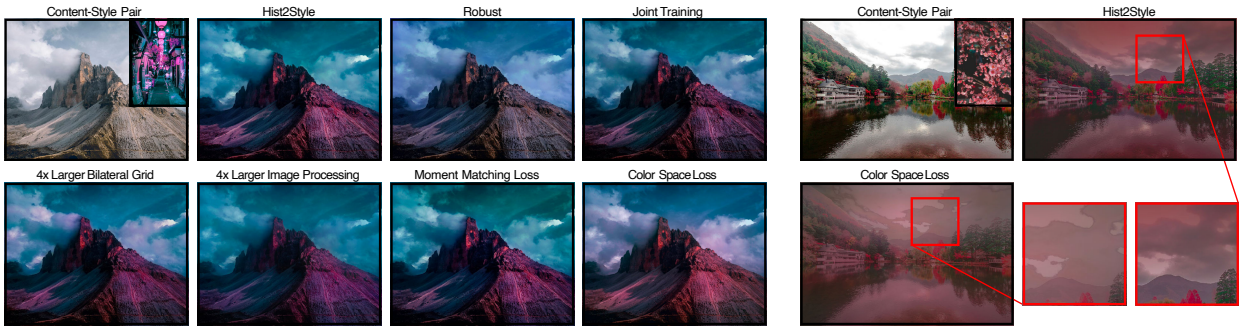

Figure 3. **Ablation studies and extensions.** We explore different architectural and training configurations. "Color Space Loss" refers to the output from a model where the loss is applied to raw color space during training, instead of perceptual space. Compared to this model, Hist2Style expresses a richer color transfer and better maintains image contrast. "Robust" refers to the output from a model trained not with the true histogram of the target image, but with that of another image in the same style. Compared to the robust model, Hist2Style displays similar contrast but improved color richness and accuracy.

5. Limitations and Future Work

Smarter data synthesis. We use the same prompts to generate styles for all content images for efficiency. In future work, we could use vision-language models (VLMs) [11, 14] for generating stylization prompts that are targeted for each content image.

Intuitive Color Space. In this work, we discuss manipulating histograms in Y'CbCr color space. However, potentially other intuitive spaces like HSV/HSL are worth experimenting with in future work.

Model Photorealistic Constraints. We constrain our model to be photorealistic by ensuring it is locally affine in bilateral space. However, the concepts in this paper naturally extend to other transforms such as LUTs, as well as bilateral grids that store features more expressive than affine transforms. Another possibility is to make the bilateral space higher-dimensional [1, 3, 4], *e.g.* from the current (Y, H, W) to (Y_1, Y_2, Y_3, H, W) .

Controllability. We provide fine user control via histogram manipulation. Marginal histograms are used by photographers in various image editing software programs [2]. We also allow users to vary the strength of the applied bilateral grid, to attenuate or strengthen it. In future work, one could use different style embedding representations based on desired user controls. For example, introducing masks to spatially mix different styles.

Nondestructive Content Editing. For some styles it may be necessary to add effects such as film grain, speckle, or bokeh. Future work would combine Hist2Style's nondestructive color adjustments with effect layers to reproduce these looks while preserving photorealism.

Video and 3D Assets. Our approach could be extended to other domains such as video [9, 13, 16] and 3D representations [12]. Future work would include enforcing temporal consistency for video sequences and integrating the method with radiance field based scene representations [12].

General Style Transfer. In this work, we focus on photorealistic stylization but this procedure naturally extends to non-photorealistic artistic objectives.

Tradeoffs with ReHistoGAN. ReHistoGAN [6] presents a notable alternative to our approach. Rather than employing a lightweight model that operates natively in bilateral space, they utilize a high-capacity model and rely on post-processing to constrain the edit to be locally affine in bilateral space. This reliance on a heavier architecture incurs higher computational latency (see main Tab. 2) and introduces the risk of content hallucinations, although the latter is mitigated by their post-processing step. Nevertheless, this strategy offers greater potential for complex spatial reasoning. As such high-capacity models become more computationally efficient, exploring this tradeoff remains a promising direction for future work.

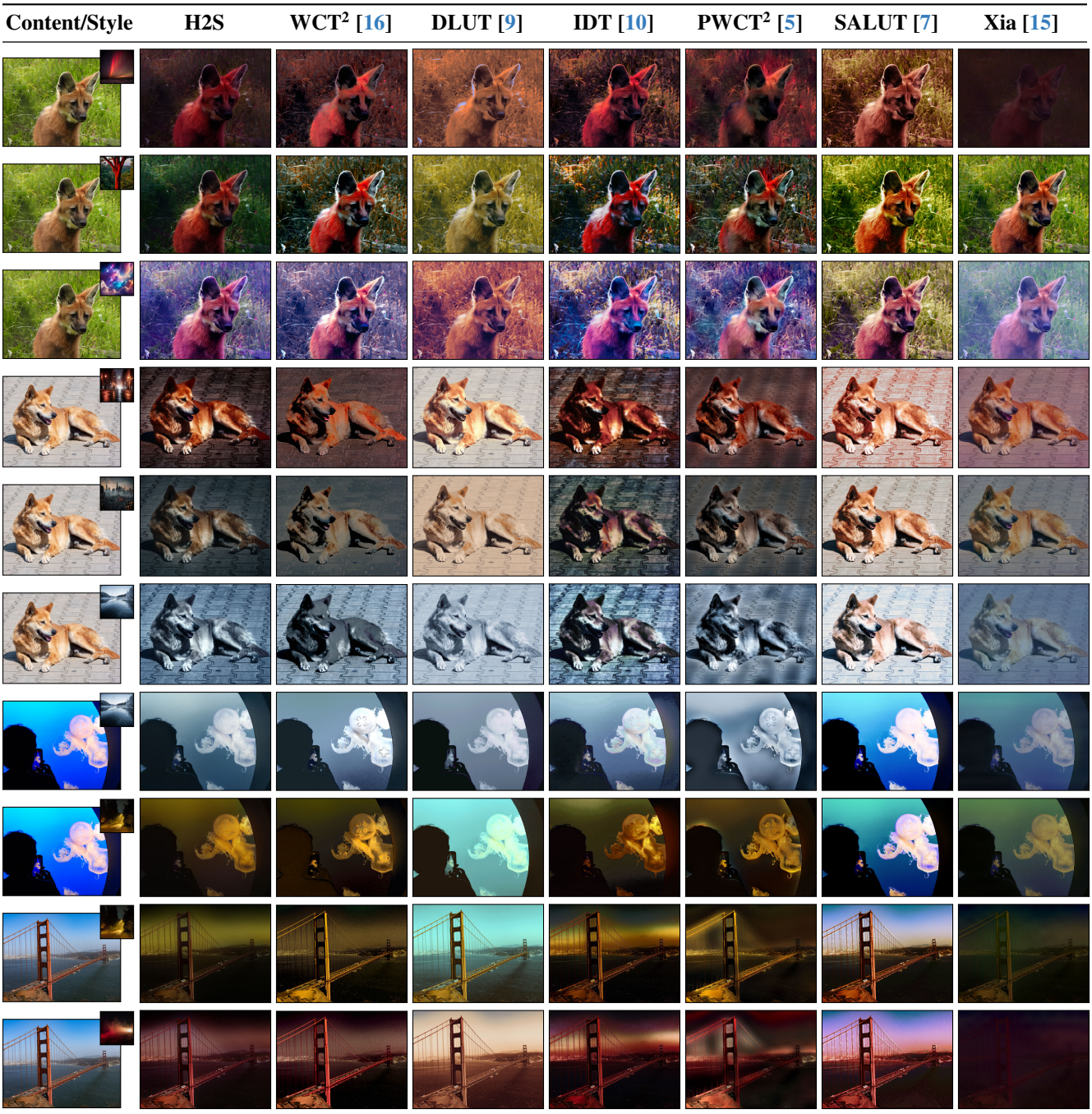

Figure 4. Qualitative comparison on the non-public evaluation set. Best viewed zoomed in.

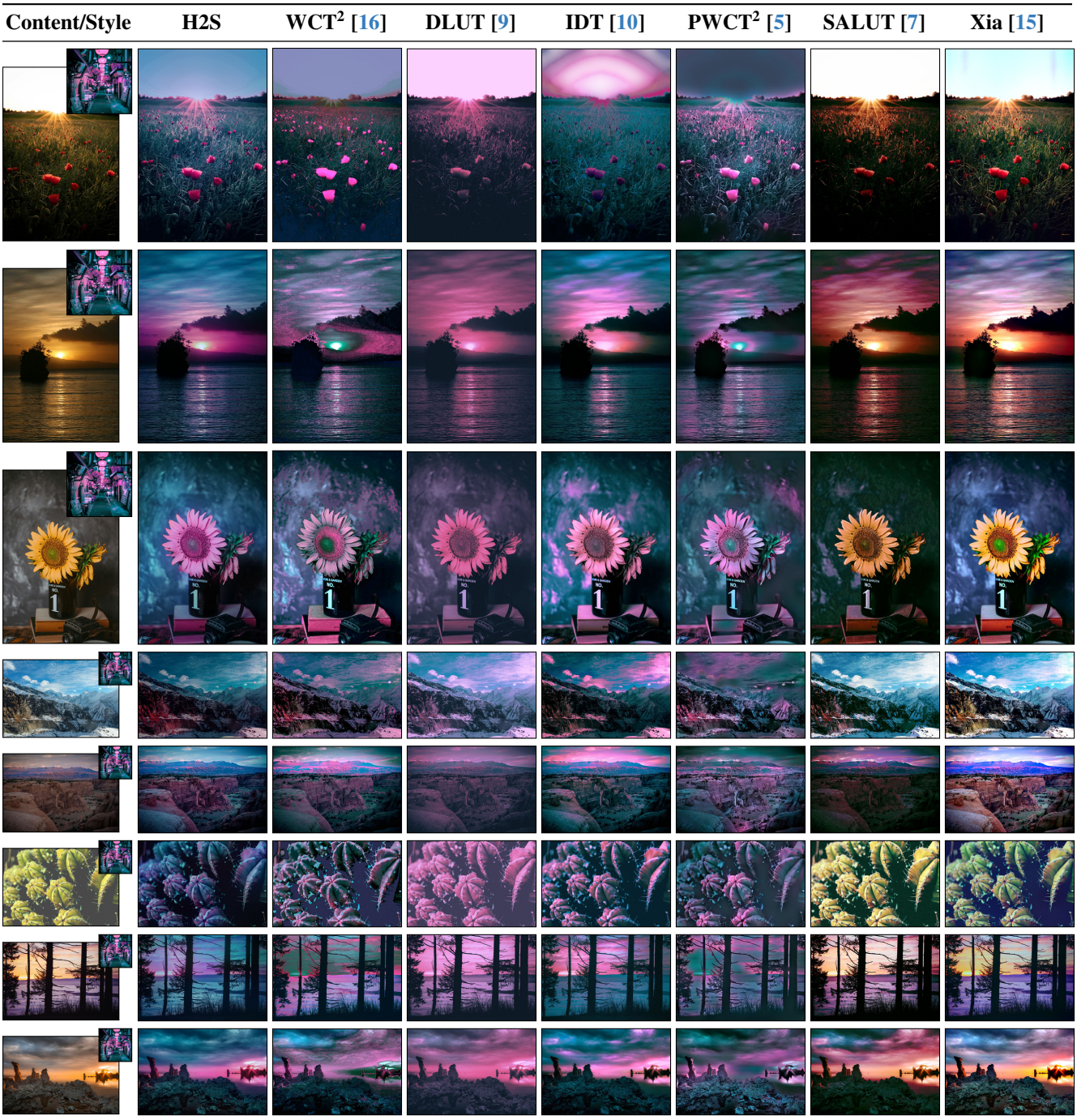

Figure 5. Different content images with a single style. Qualitative comparison on the user-study evaluation set. Best viewed zoomed in.

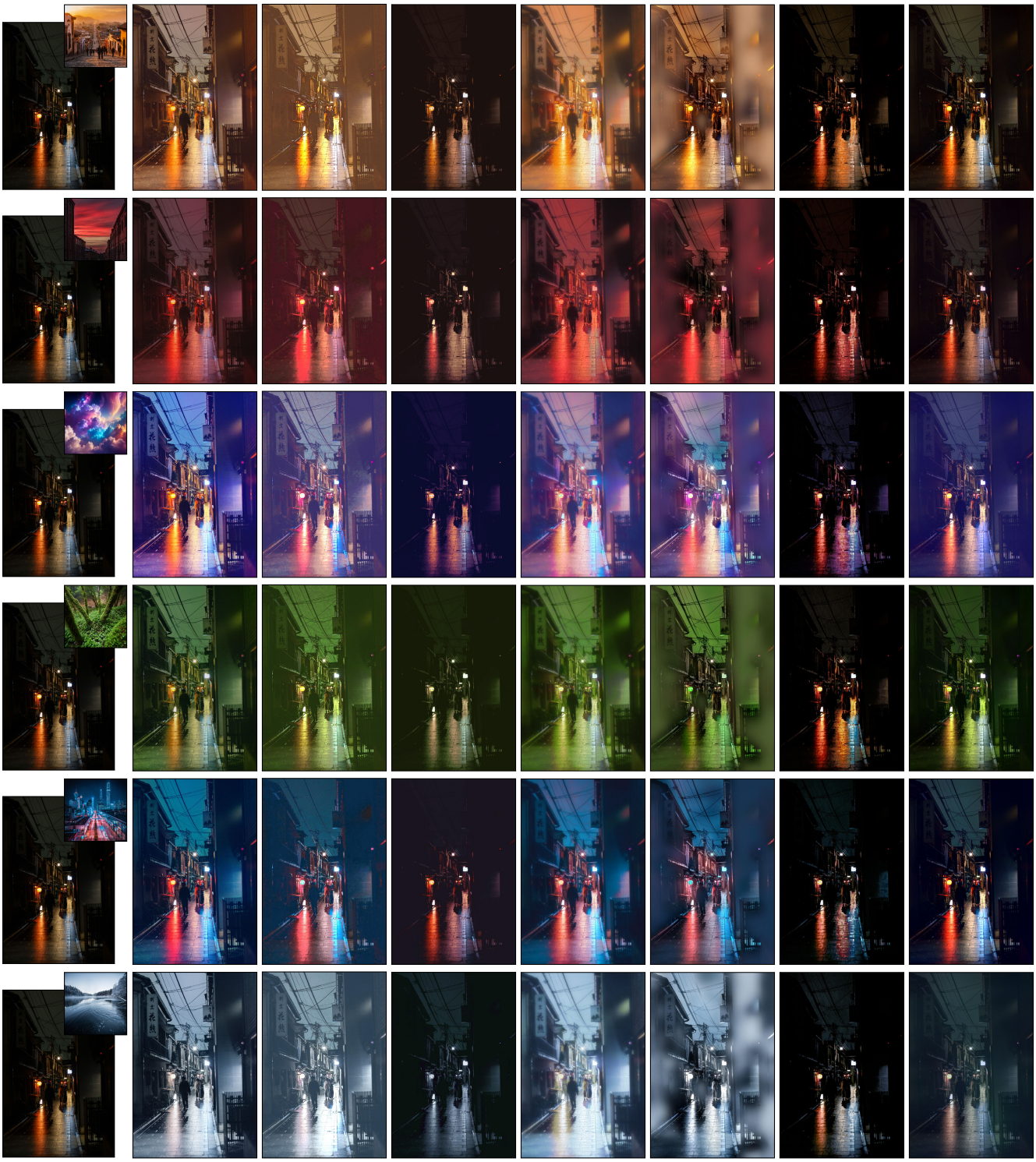

Figure 6. Single content image with different styles. Qualitative comparison on the non-public evaluation set. Best viewed zoomed in.

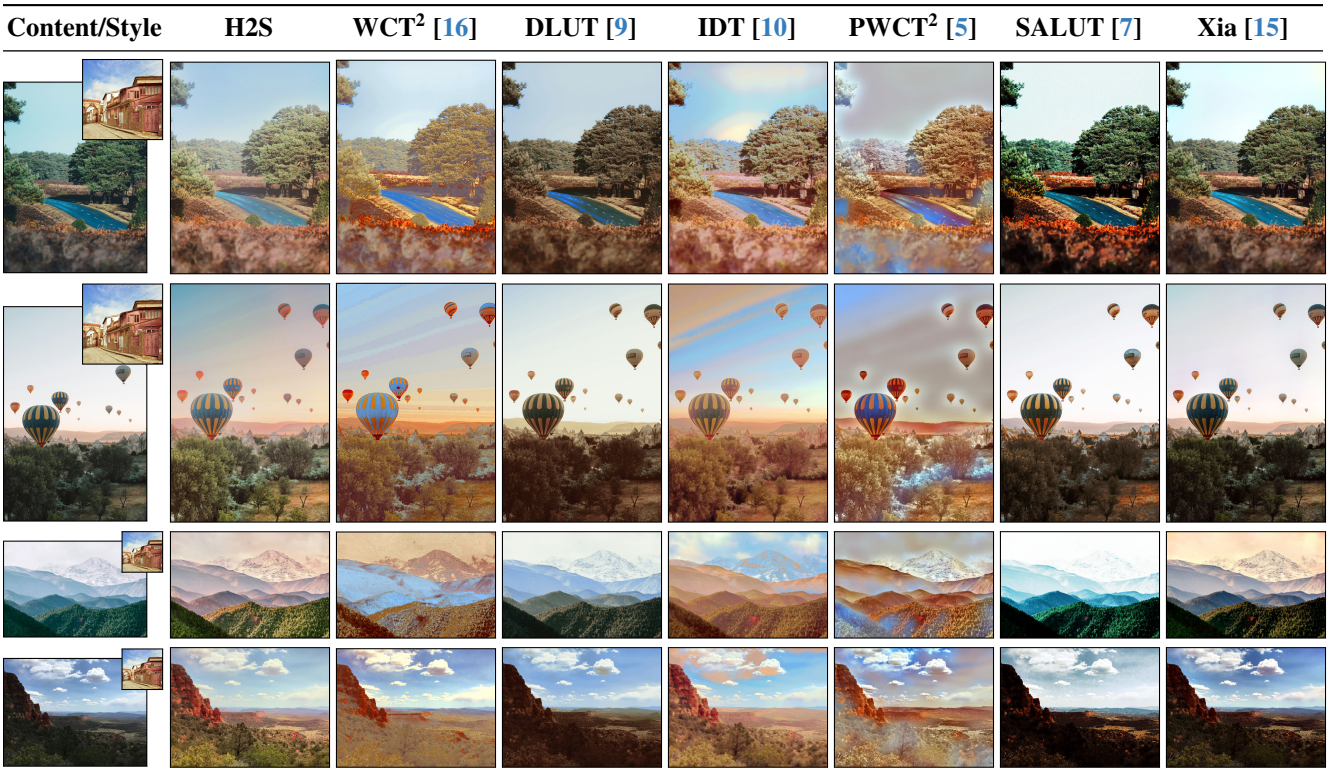

Figure 7. Qualitative comparison on the non-public evaluation set. Best viewed zoomed in.

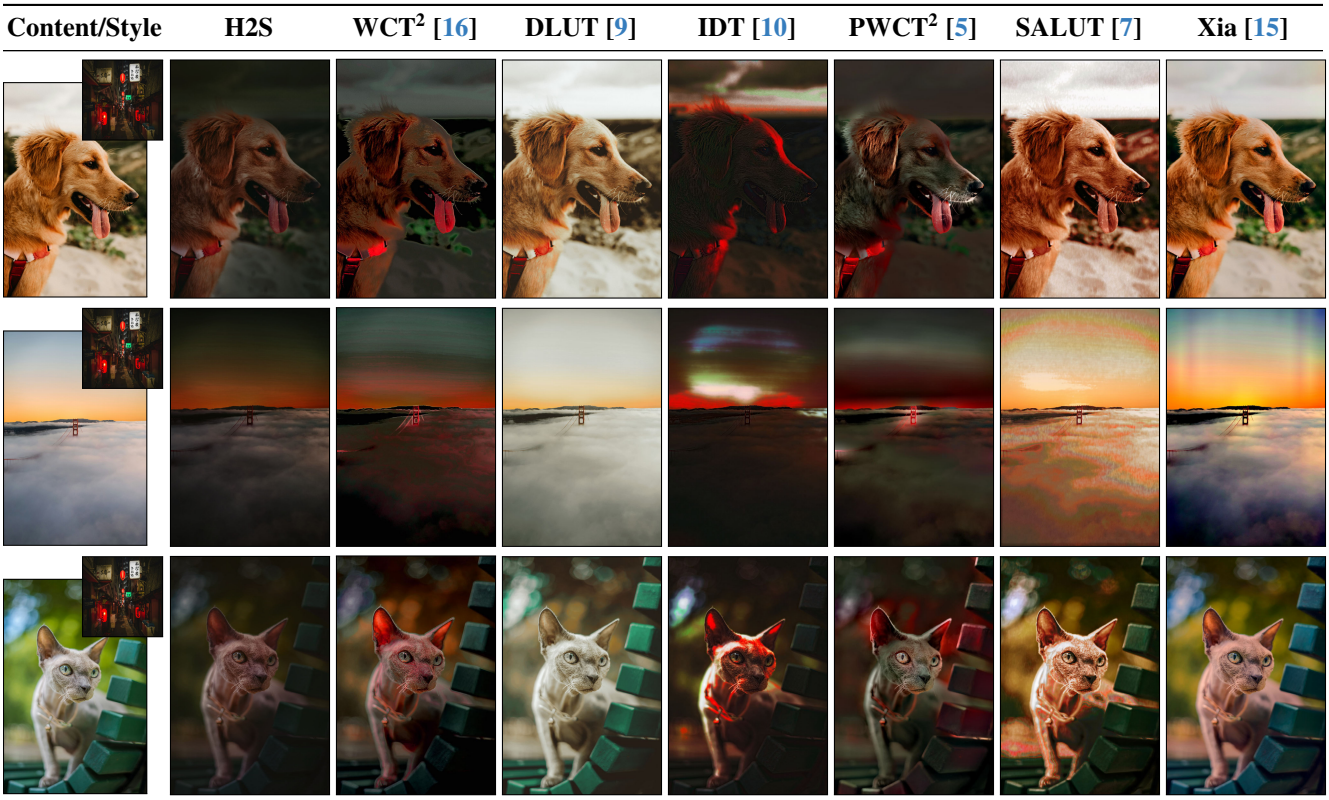

Figure 8. Qualitative comparison on the user-study evaluation set. Best viewed zoomed in.

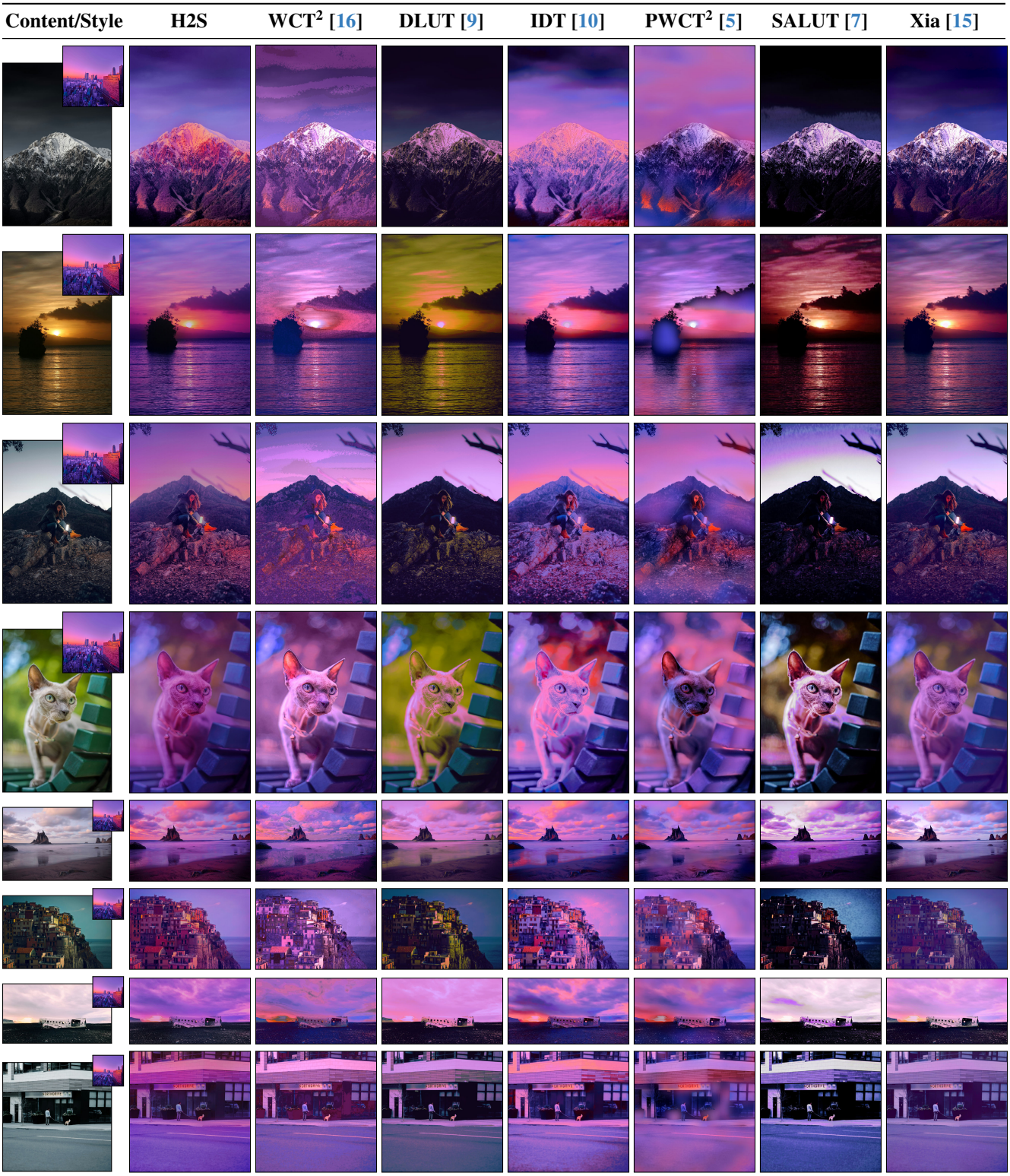

Figure 9. Different content images with a single style. Qualitative comparison on the user-study evaluation set. Best viewed zoomed in.

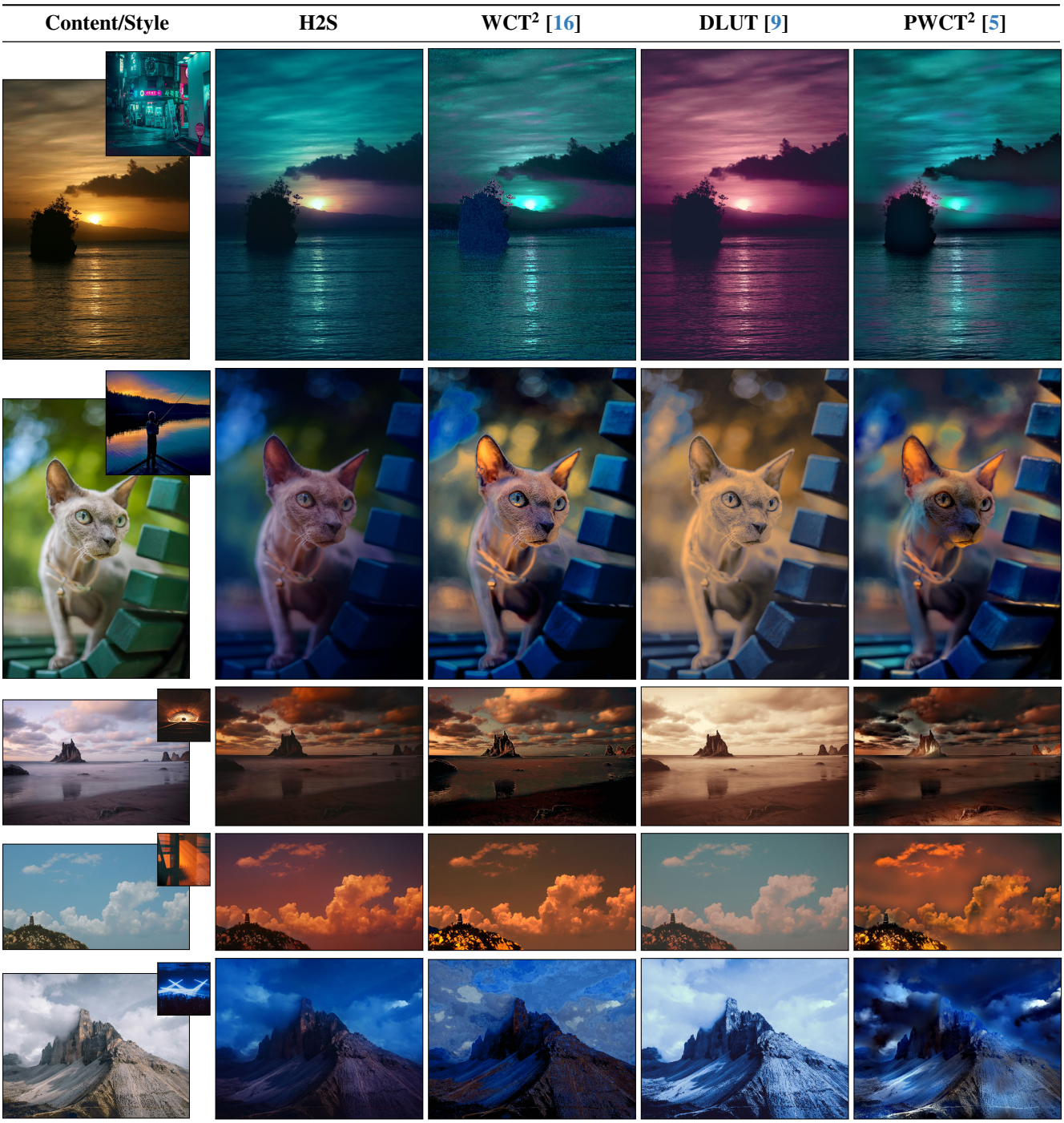

Figure 10. Qualitative comparison with WCT^2 [16], DLUT [9], and PhotoWCT^2 [5] on the user-study evaluation set.

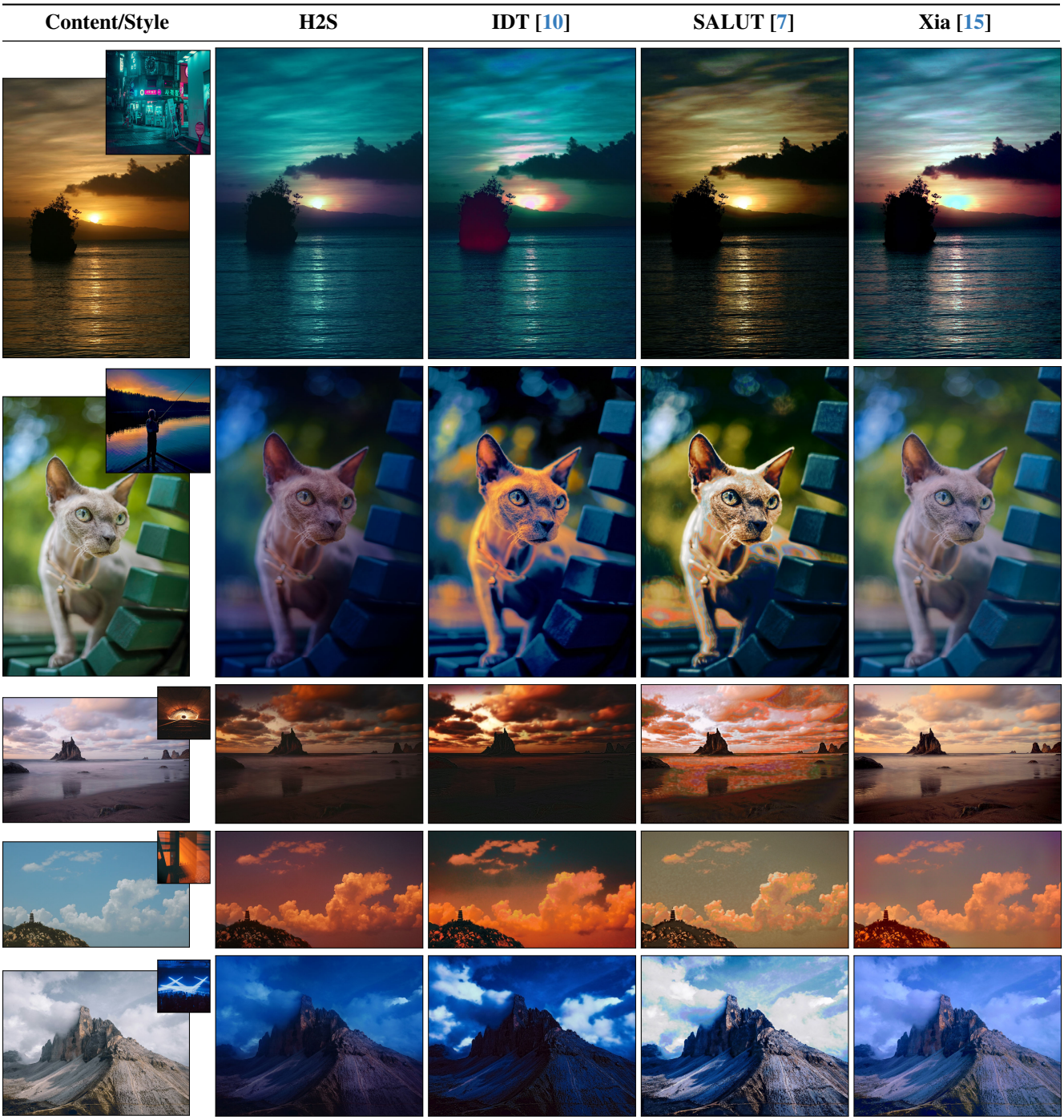

Figure 11. Qualitative comparison with IDT [10], SALUT [7], and Xia [15] on the user-study evaluation set.

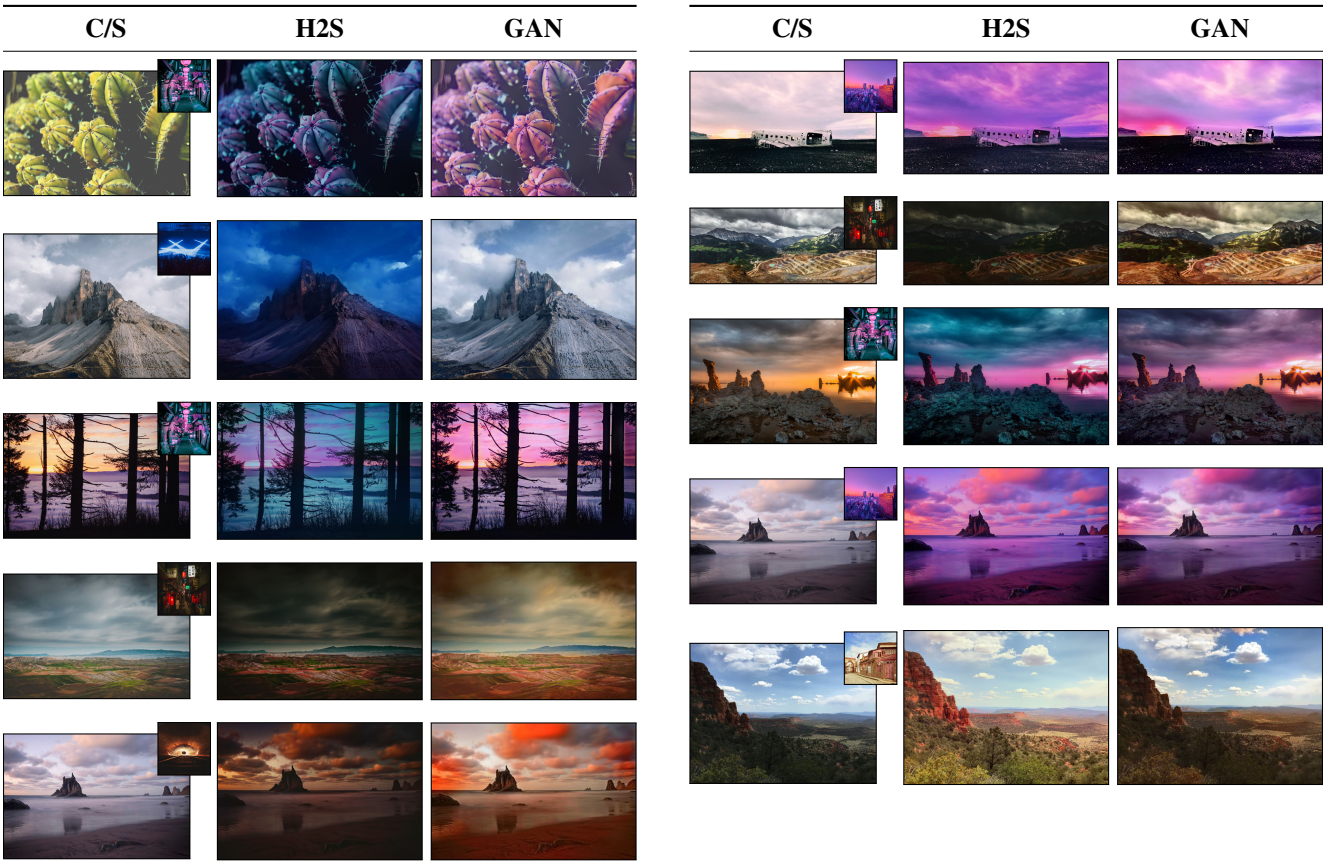

Figure 12. Qualitative comparison of Hist2Style (H2S) against ReHistoGAN (GAN) on the user-study evaluation set.

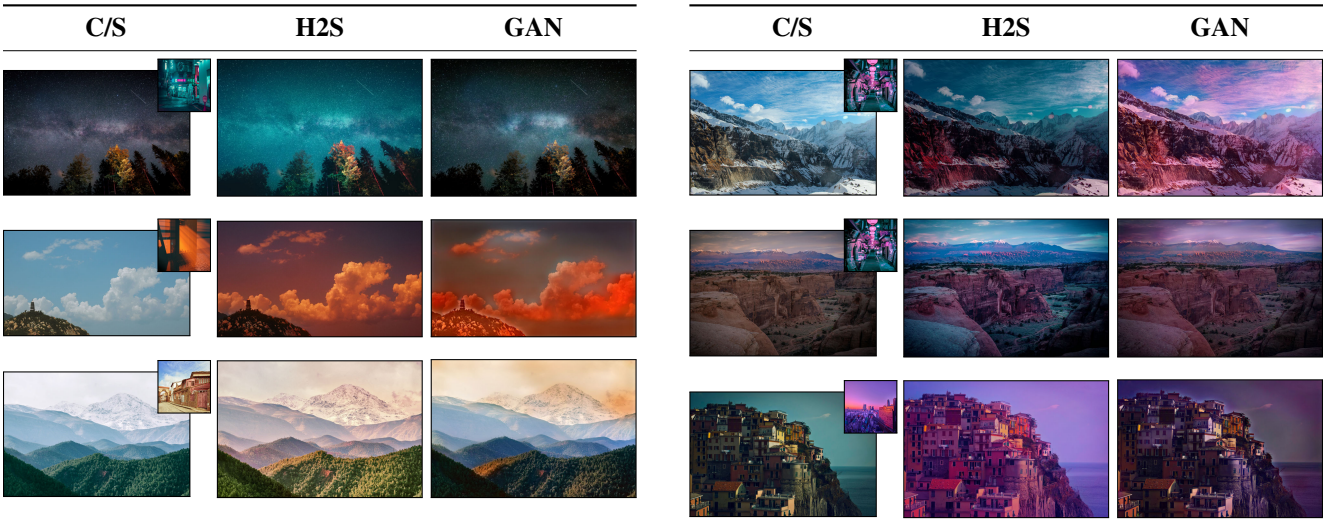

Figure 13. Qualitative comparison of Hist2Style (H2S) against ReHistoGAN (GAN) on the user-study evaluation set.

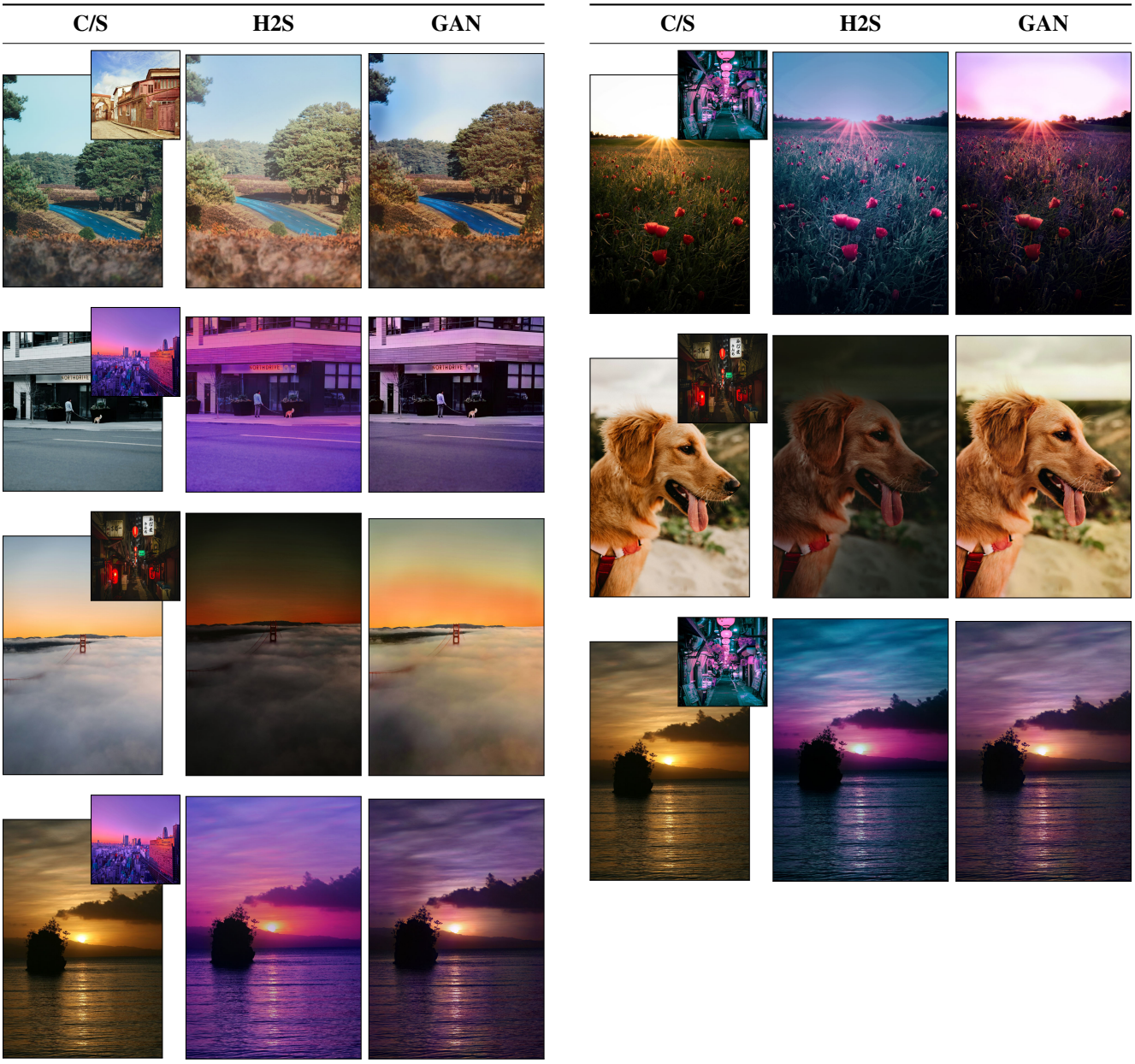

Figure 14. Qualitative comparison of Hist2Style (H2S) against ReHistoGAN (GAN) on the user-study evaluation set.

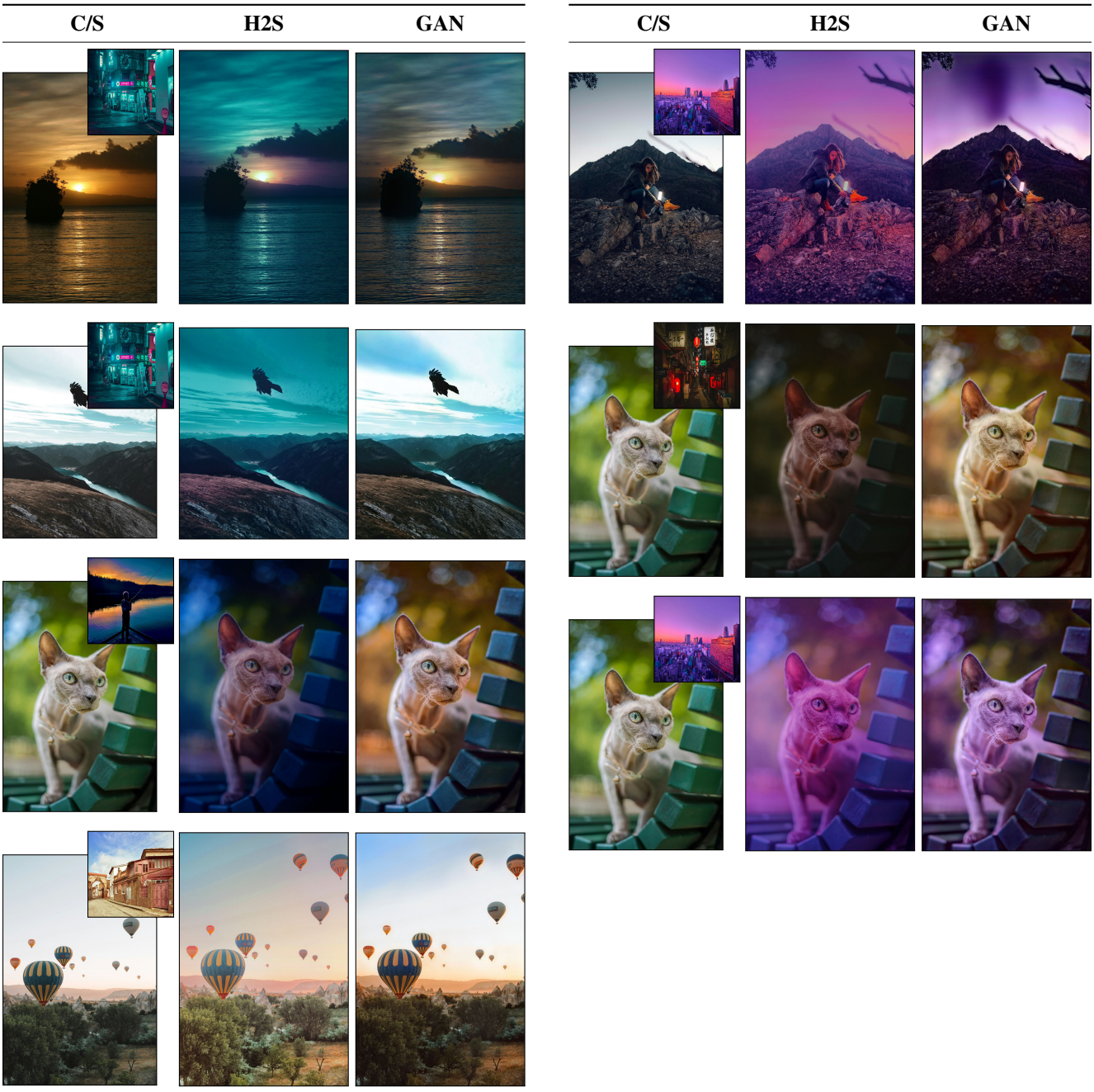

Figure 15. Qualitative comparison of Hist2Style (H2S) against ReHistoGAN (GAN) on the user-study evaluation set.

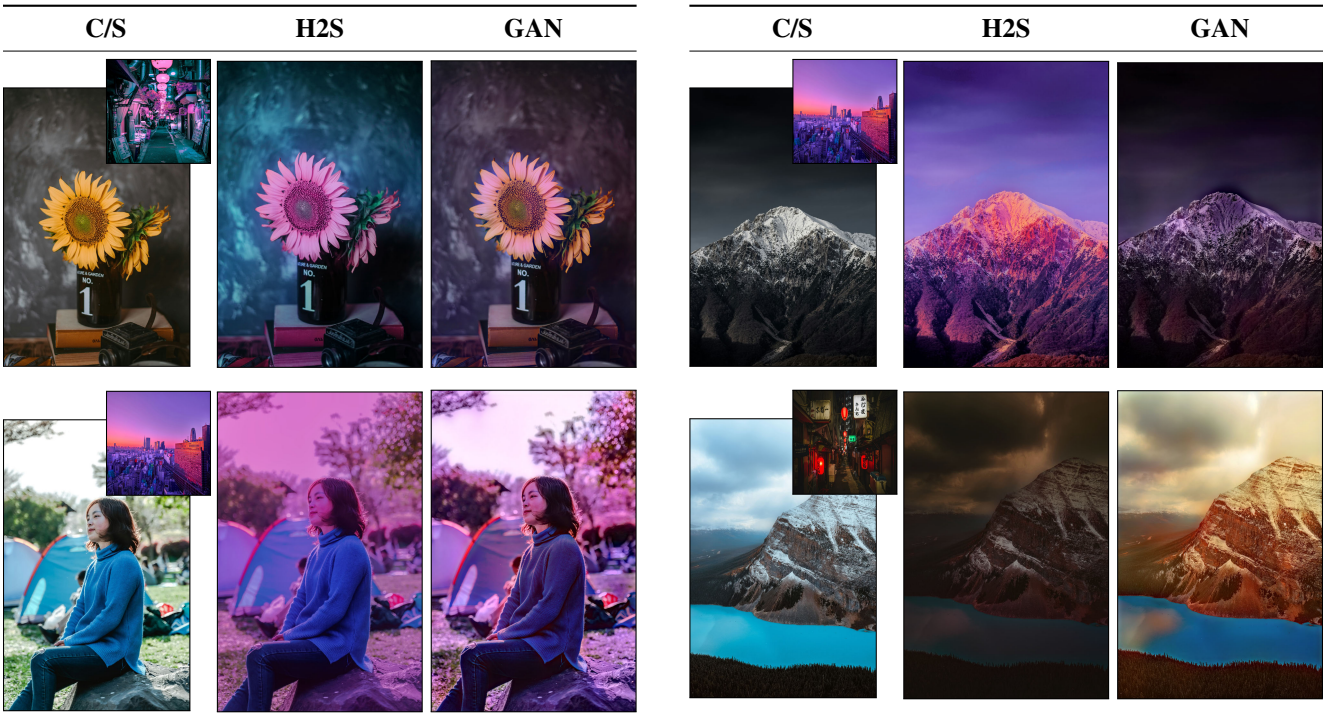

Figure 16. Qualitative comparison of Hist2Style (H2S) against ReHistoGAN (GAN) on the user-study evaluation set.

References

[1] Andrew Adams, Jongmin Baek, and Abe Davis. Fast high-dimensional filtering using the permutohedral lattice. *Eurographics*, 2010. 3

[2] Adobe Systems Incorporated. Adobe photoshop lightroom, 2007. 3

[3] Jonathan T Barron and Ben Poole. The fast bilateral solver. *ECCV*, 2016. 3

[4] Jonathan T Barron, Andrew Adams, YiChang Shih, and Carlos Hernández. Fast bilateral-space stereo for synthetic defocus. *CVPR*, 2015. 3

[5] Tai-Yin Chiu and Danna Gurari. Photowct^2: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals, 2021. 1, 4, 5, 6, 7, 8, 9

[6] Afifi et al. Histogan: Controlling colors of gan-generated and real images via color histograms. In *CVPR*, 2021. 3

[7] Zerui Gong, Zhonghua Wu, Qingyi Tao, Qinyue Li, and Chen Change Loy. Sa-lut: Spatial adaptive 4d look-up table for photorealistic style transfer, 2025. 4, 5, 6, 7, 8, 10

[8] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 1, 2

[9] Mujing Li, Guanjie Wang, Xingguang Zhang, Qifeng Liao, and Chenxi Xiao. D-LUT: Photorealistic Style Transfer via Diffusion Process . In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9206–9214, Los Alamitos, CA, USA, 2025. IEEE Computer Society. 1, 3, 4, 5, 6, 7, 8, 9

[10] P. F. Pitié, A. C. Kokaram, and R. Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 2007. 1, 4, 5, 6, 7, 8, 10

[11] Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. Gemini: A family of highly capable multimodal models, 2025. 1, 3

[12] Yuehao Wang, Chaoyi Wang, Bingchen Gong, and Tianfan Xue. Bilateral guided radiance field processing, 2024. 3

[13] Cheng-Yu Wei, N. Dimitrova, and Shih-Fu Chang. Color-mood analysis of films based on syntactic and psychological models. In *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, pages 831–834 Vol.2, 2004. 3

[14] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 3

[15] Xide Xia, Meng Zhang, Tianfan Xue, Zheng Sun, Hui Fang, Brian Kulis, and Jiawen Chen. Joint bilateral learning for real-time universal photorealistic style transfer. *CoRR*, abs/2004.10955, 2020. 1, 2, 4, 5, 6, 7, 8, 10

[16] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 4, 5, 6, 7, 8, 9