Towards Photorealistic Style Transfer with Multimodal Guidance and Robustness to Content Images in Arbitrary Styles - Supplementary Material -

Ruikai Zhou*, Yating Liu*, Yi Xu†
Shanghai Jiao Tong University, Shanghai, China
{1323268276, Olivialyt, xuyi}@sjtu.edu.cn

Table 4. User study results for image-guided photorealistic style transfer task. The best and second results are highlighted in red and blue, respectively.

Method	Content Similarity↑	Style Similarity↑	Overall Quality↑
StyleID [3]	3.16	3.94	3.37
CT [13]	4.71	3.86	4.11
WCT^2 [19]	3.93	3.81	3.87
Bilateral [16]	4.31	3.69	3.79
CAP-VSTNet [15]	4.19	3.99	4.06
DNCM [6]	4.57	4.05	4.19
D-LUT [9]	4.60	3.63	4.01
Ours	4.67	4.17	4.31

Table 5. User study results for text-guided photorealistic style transfer task. The best and second results are highlighted in red and blue, respectively.

Method	Content Similarity↑	Style Similarity↑	Overall Quality↑
CLIPstyler [7]	2.37	3.39	2.55
ZeCon [18]	2.83	3.31	3.01
SpectralCLIP [17]	2.52	2.95	2.43
IP2P [1]	3.64	3.24	3.27
SuperEdit [11]	3.28	3.22	3.15
CLIPtone [8]	4.61	3.53	3.76
Ours	4.76	3.67	3.88

1. User Study

Given the subjective nature of photorealistic style transfer, we further evaluate the performance through user studies. Considering that our method supports flexible switching between image and text guidance, we conduct user studies for both tasks. Specifically, for each task, we invite 20 participants to evaluate 20 subgroups of images, where the input images are randomly selected from our test set. In the case of image-guided photorealistic style transfer, each subgroup includes a content image, a reference style image, and the stylized results generated by different methods for comparison. Similarly, for text-guided photorealistic style transfer, each subgroup contains a content image, a text description, and the stylized results from various methods. For each subgroup, users are asked to evaluate each method based on three evaluation criteria: content similarity, style similarity, and overall quality. We use a widely-adopted five-grade categorical scale [12] for quality assessment, with scores ranging from 1 to 5 ("Bad", "Poor", "Fair", "Good", "Excellent"). Finally, 400 scores (20 × 20) are collected for each criterion, and the average of these scores is then calculated.

Table 4 and Table 5 present the results, from which we can draw the following conclusions:

(1) For image-guided photorealistic style transfer, our method achieves the highest overall quality and style similarity scores. While our content similarity score is slightly lower than that of CT [13], CT performs poorly in terms of style similarity and overall quality.

(2) For text-guided photorealistic style transfer, our method outperforms other approaches across all three evaluation criteria. These results demonstrate that our method not only ensures high content fidelity but also achieves superior stylization and overall visual quality.

2. Efficiency Comparison

In Table 6 and Table 7, we evaluate the inference efficiency of different methods in terms of runtime and GPU memory usage. Bilateral [16] is absent due to the lack of publicly available code. Zecon [18] encounters out-of-memory (OOM) errors at resolutions above 256×256. Therefore, the qualitative results of Zecon presented in the main paper are obtained by first downsampling the input to 256×256 and then upsampling the output. D-LUT [9] is significantly slower, as it requires training a separate model for each style

*These authors contributed equally.
†Corresponding author.

Table 6. Efficiency comparison on image-guided photorealistic style transfer. Bilateral [16] is absent due to the unavailability of its code. "OOM" denotes out-of-memory.

Resolution	Metric	StyleID [3]	CT [13]	WCT2 [19]	CAP-VSTNet [15]	DNCM [6]	D-LUT [9]	Ours
1080p	Inference Time (ms)	OOM	0.56	1725	679.6	20.80	73883	21.74
	Memory (GB)		1.53	16.2	7.13	1.71	3.29	1.75
2K	Inference Time (ms)	OOM	0.91	2622	1732	21.44	75027	21.92
	Memory (GB)		1.65	24.0	11.38	1.80	3.29	1.85
4K	Inference Time (ms)	OOM	2.14	OOM	2578	21.86	75700	22.62
	Memory (GB)		2.53		23.60	2.00	3.29	2.10
8K	Inference Time (ms)	OOM	7.64	OOM	OOM	24.79	78337	28.16
	Memory (GB)		3.67			3.14	4.78	3.53

Table 7. Efficiency comparison on text-guided photorealistic style transfer. Zecon [18] runs out of memory at resolutions higher than 256×256. "OOM" denotes out-of-memory.

Resolution	Metric	CLIPstyler [7]	ZeCon [18]	SpectralCLIP [17]	IP2P [1]	SuperEdit [11]	CLIPtone [8]	Ours
1080p	Inference Time (ms)	105636	OOM	110016	44542	44013	17.66	26.06
	Memory (GB)	15.14		15.27	6.02	6.02	2.22	2.00
2K	Inference Time (ms)	170322	OOM	172706	116396	114038	18.40	27.18
	Memory (GB)	23.74		24.24	8.80	8.80	2.26	2.10
4K	Inference Time (ms)	OOM	OOM	OOM	514945	502925	19.18	27.74
	Memory (GB)				16.72	16.72	2.49	2.36
8K	Inference Time (ms)	OOM	OOM	OOM	OOM	OOM	24.09	32.38
	Memory (GB)						3.62	3.78

Figure 10. Comparison of style removal strategies. Grayscale conversion loses color information, while high-frequency extraction loses content details.

image during inference. As shown in Table 6 and Table 7, our method achieves competitive or superior efficiency in both image-guided and text-guided tasks.

3. Additional Ablation Studies

3.1. Style Removal Module

To further evaluate the effectiveness of the proposed Style Removal Module, we compare it with two naive alternatives: (1) the gray-scaled version [5] of the content image; (2) the high-frequency components [19] (wavelet coefficients) of the content image. Figure 10 presents the qualitative results. Grayscale conversion removes inherent color information, restricting style transfer to coarse global adjustments and causing the image to lose its rich colors. High-frequency components, on the other hand, preserve only edge details and fail to recover pixel-level content. In contrast, our Style Removal Module effectively disentangles content from style, producing a style-less image that fully preserves original content details, enabling accurate and coherent style injection in the subsequent stage.

3.2. Design of Equation 11

In the main paper (Eq. 11), we obtain the text-guided mapping-adaptive parameters r^{Text} for iDRA-MLP by

Table 8. Comparison of Equation 11 (ours) with the direct multiplication baseline (Eq. 24). Best results are highlighted in **bold**.

Setting	Grayscale SSIM ↑	CLIP Image Similarity ↑	CLIP Text-Image Directional Similarity ↑
Direct Multiplication	0.907	0.955	0.085
Ours	**0.919**	**0.962**	**0.087**

Figure 11. Qualitative ablation results for the remaining loss terms.

Table 9. Ablation study on loss weights. We vary each hyperparameter in a reasonable range. "Ours" denotes our default setting with λ_{nor}, λ_{CLIP}, λ_{ct}, λ_{wt}, and λ_{mm} set to 1.0, 1.0, 1.0, 0.05, and 1.0, respectively.

Metric	Ours	λ_{nor} (0.5 / 2.0)	λ_{CLIP} (0.5 / 2.0)	λ_{ct} (0.5 / 2.0)	λ_{wt} (0.01 / 0.1)	λ_{mm} (0.5 / 2.0)
Content Similarity ↑	0.762	0.758 / 0.766	0.762 / 0.763	0.764 / 0.764	0.763 / 0.762	0.763 / 0.763
Style Similarity ↑	0.848	0.850 / 0.843	0.848 / 0.846	0.846 / 0.850	0.844 / 0.845	0.845 / 0.850
Grayscale SSIM ↑	0.919	0.914 / 0.924	0.929 / 0.908	0.909 / 0.928	0.919 / 0.918	0.923 / 0.909
CLIP Image Similarity ↑	0.962	0.963 / 0.961	0.967 / 0.958	0.959 / 0.967	0.960 / 0.964	0.961 / 0.961
CLIP Text-Image Directional Similarity ↑	0.087	0.090 / 0.091	0.085 / 0.092	0.093 / 0.084	0.091 / 0.089	0.089 / 0.090

combining $\tilde{r}$ and w:

$$r^{Text} = w \cdot (1 + \tilde{r}). \tag{23}$$

Here, $\tilde{r}$ is extracted from the style-related target text description, while w represents the normal style element of the content image. This design ensures that when $\tilde{r}$ equals zero, the text-based style guidance has no effect, and only the content image's normal style is retained.

To further validate this design choice, we performed additional experiments using a simpler alternative that directly multiplies the parameters:

$$r^{Text} = w \cdot \tilde{r}. \tag{24}$$

As shown in Table 8, our original formulation consistently achieves better stylization quality, demonstrating the effectiveness of adding the constant 1.

3.3. Loss Terms

Our model is supervised by the loss function defined in Equation 22 of the main paper. Each loss term plays a distinct role. The importance of L_{ct}, L_{wt}, and L_{mm} has been demonstrated through ablation studies in Figure 9 and Table 3 of the main paper.

Here, we further investigate the remaining losses through ablation experiments, with results shown in Figure 11.

- L_{con} ensures that two images with the same content but different styles produce consistent style-less representations after the Style Removal Module, enabling effective removal of the original style. As shown in the second row of Figure 11, removing L_{con} degrades performance on content images that deviate from the "normal" style.
- L_{rec} constrains the two style-less images to reconstruct their original images through mutual style guidance. This endows the Style Injection Module with the ability to accurately inject styles. As shown in Figure 11, removing L_{rec} leads to meaningless outputs. This is because, as discussed in Section 3.4.1 of the main paper, without L_{rec} the network tends to learn a trivial solution where the Style Removal Module produces identical outputs regardless of the input to minimize L_{con}.
- L_{nor} allows the MLP l in Figure 4 (main paper) to extract the normal style element. Removing L_{nor} prevents MLP l from receiving gradients, leaving it at its initial weights and resulting in meaningless outputs.
- L_{CLIP} equips the model with text-guided style transfer capability. Without this loss, the model generates outputs that fail to align with the style text, as shown in Figure 11.

Table 10. Ablation on hidden layer dimension k. Best results are highlighted in **bold**.

k	Grayscale SSIM ↑	CLIP Image Similarity ↑	CLIP Text-Image Directional Similarity ↑
8	0.917	0.960	0.076
32	0.909	0.961	**0.088**
16 (Ours)	**0.919**	**0.962**	0.087

Figure 12. Additional text-guided qualitative results of our method with various text descriptions.

Table 11. Evaluation of dataset robustness. Best results are highlighted in **bold**.

Training Dataset	Grayscale SSIM ↑	CLIP Image Similarity ↑	CLIP Text-Image Directional Similarity ↑
MS-COCO	**0.925**	**0.963**	0.086
MIT-Adobe 5K	0.919	0.962	**0.087**

3.4. Loss Hyperparameters

We set the loss weights as described in Section 4.1 of the main paper. We follow previous work [6] to set λ_{con} and λ_{rec}, and set the others based on a reasonable scale. To examine robustness, we conduct an ablation study by varying each loss weight within a reasonable range while keeping others fixed. As shown in Table 9, our method is not sensitive to specific hyperparameter values and consistently maintains stable performance across all metrics.

3.5. Hidden Layer Dimension

The hidden layer dimension k of both rDRA-MLP and iDRA-MLP is set to 16 in our main experiments. To evaluate the influence of this setting, we conduct ablations with $k = 8$ and $k = 32$. As shown in Table 10, setting $k = 8$ leads to inferior performance due to insufficient representation capacity, while $k = 32$ yields saturated results without significant improvement but introduces unnecessary computational cost. Therefore, we adopt $k = 16$ as a balanced setting that achieves strong performance while maintaining efficiency.

3.6. Text Description

For simplicity, during training and inference, we generate style-related text prompts by directly appending the word "photo" to each color name (e.g., transforming "warm" into "warm photo"). To verify the robustness of our method across different prompts, we evaluate with a diverse set of prompt templates, as shown in Figure 12. While our training process utilizes a fixed template and concise descriptions, the model exhibits remarkable generalization capabilities. It successfully handles a wide spectrum of textual inputs, including those long and complex descriptions.

4. Additional Dataset Evaluation

As described in the main paper, we train the text-guided branch using the MIT-Adobe 5K dataset [2], which is split into 4,500 training and 500 testing images. To demonstrate that our approach is not dependent on a specific dataset, we additionally train the model on MS-COCO [10] while keeping the same testing set for evaluation.

The results are presented in Table 11 and Figure 13. As can be seen, our method still achieves excellent results when trained on a different dataset, confirming its robustness.

5. Examples of Style-less Images

The first stage of our framework employs the Style Removal Module to generate a style-less version of the content image, which effectively disentangles the content and original

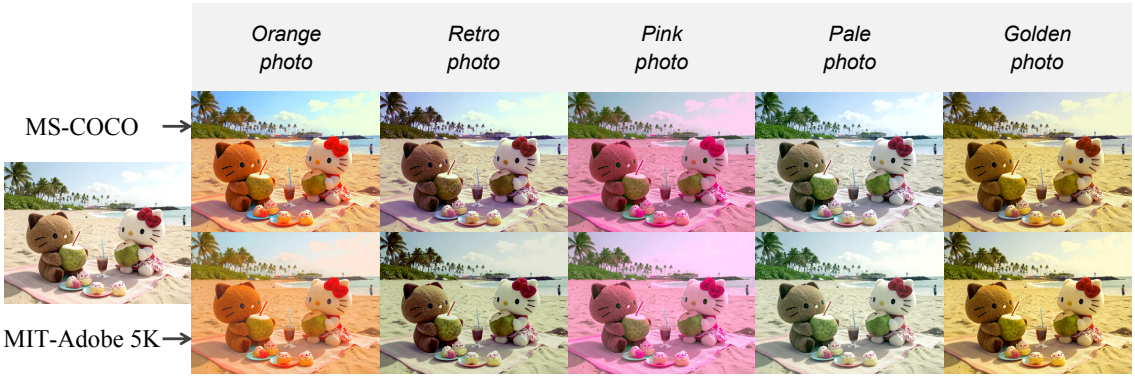

Figure 13. Qualitative results of training the text-guided branch on different datasets. Even when trained on MS-COCO instead of MIT-Adobe 5K, our method consistently produces promising photorealistic stylization results.

Figure 14. Examples of style-less images. The style-less image represents the "content", which is the result of removing all styles (including normal style) from a given image.

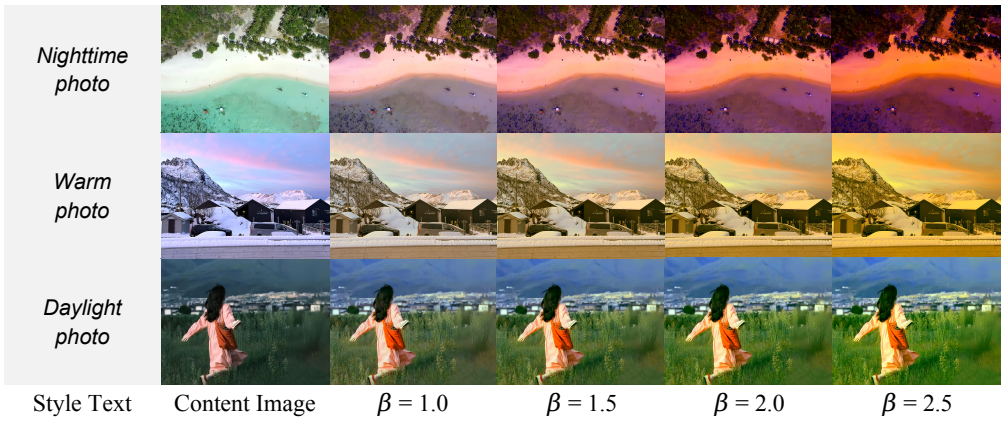

Figure 15. Flexible control of text-guided stylization intensity. By varying the scaling factor, the stylization strength can be adjusted from subtle to strong.

style. Since predefining this representation is challenging, we learn it via specific constraints during training, as detailed in Sections 3.3 and 3.4 of the main paper. To further illustrate the characteristics of the style-less images, we present multiple visual examples in Figure 14.

The results show that the style-less images successfully preserve the content while completely removing the original style of the content image, which aligns with our intended outcomes.

6. Flexible Text-Guided Stylization Intensity

As described in the main paper (Eq. 11), we obtain the text-guided mapping-adaptive parameters r^{Text} for iDRA-MLP by combining $\tilde{r}$ and w:

$$r^{Text} = w \cdot (1 + \tilde{r}). \tag{25}$$

Here, $\tilde{r}$ is extracted from the target prompt. We can further generalize this formulation by introducing a scaling factor β

Table 12. Quantitative comparison results of multimodal fusion-guided photorealistic style transfer. Best results are highlighted in **bold**.

Method	Content Similarity ↑	Grayscale SSIM ↑	CLIP Image Similarity ↑
MMIST [14]	0.471	0.415	0.607
StyleBooth [4]	0.562	0.518	0.652
Ours	**0.758**	**0.926**	**0.964**

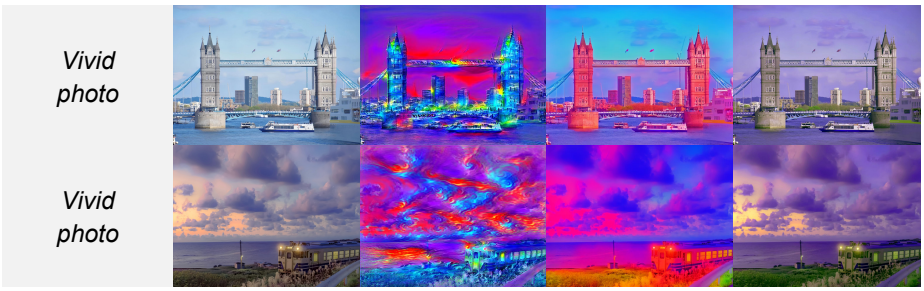

Figure 16. Illustration of CLIP-induced bias. Given the text prompt "Vivid photo", our method tends to produce images with a purple hue. Similar behavior is also observed in other CLIP-based methods such as CLIPstyler [7] and IP2P [1], suggesting that the effect originates from the inherent bias of the CLIP model.

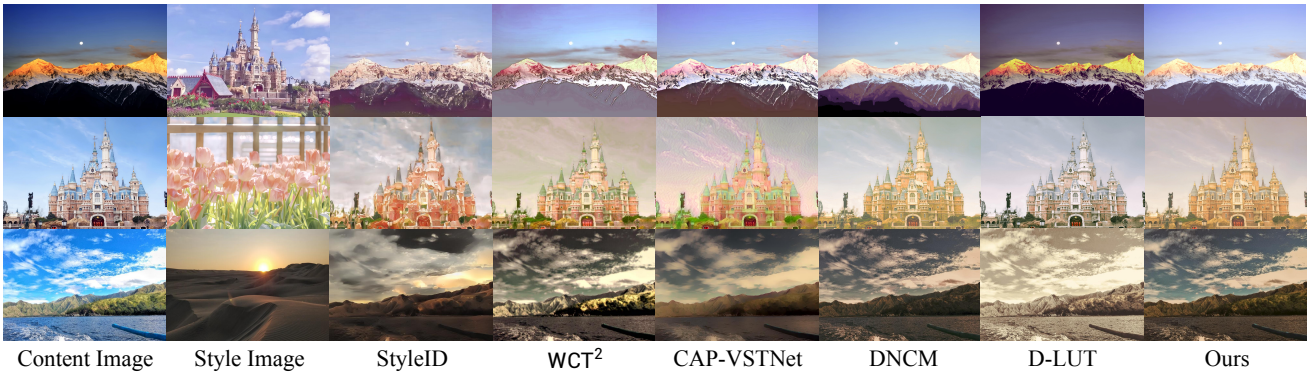

Figure 17. Additional qualitative comparison results of image-guided photorealistic style transfer.

before $\tilde{r}$, which in our current setting is $\beta = 1$. By adjusting β, we can flexibly control the intensity of text-guided stylization: larger values of β correspond to stronger stylization effects, while smaller values yield subtler results.

We conduct qualitative experiments to demonstrate this capability, as shown in Figure 15. The results clearly illustrate that gradually increasing β allows fine-grained control over the stylization strength. This flexibility enables users to customize the output according to their preferences or application requirements, without retraining the model.

7. Quantitative Comparisons of Multimodal Fusion-Guided Photorealistic Style Transfer

In this section, we present quantitative evaluations of multimodal fusion-guided photorealistic style transfer. Specifically, we randomly selected 50 triplets, each consisting of a content image, a style image, and a style text prompt. For each triplet, we applied our method with interpolation coefficients of 1.00, 0.75, 0.50, 0.25, and 0.00.

To demonstrate the ability of our approach to better preserve fine details of the content images, we measured three metrics: Content Similarity, Grayscale SSIM, and CLIP Image Similarity. We then averaged the results across all 250 outputs (50 triplets × 5 coefficients). As shown in Table 12, our method consistently achieves the highest scores across all metrics, confirming its superior performance in preserving content information compared to baseline methods.

8. Limitations and Failure Cases

First, our approach may inherit the inherent biases from the CLIP model. As illustrated in Figure 16, when given the text prompt "Vivid photo", our method consistently pro-

Figure 18. Additional qualitative comparison results of text-guided photorealistic style transfer.

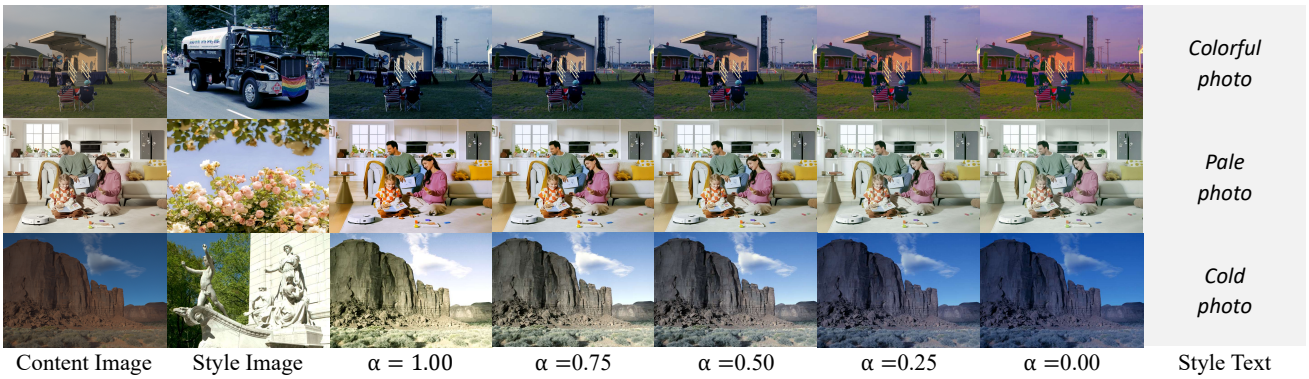

Figure 19. Additional multimodal fusion-guided photorealistic style transfer results of the proposed method.

duces images with a noticeable purple hue. It is worth noting that this phenomenon is not unique to our method but is also observed in other CLIP-based approaches such as CLIPstyler [7] and IP2P [1]. This observation suggests that the effect is likely due to biases within the CLIP model itself rather than to our design.

Second, our method relies on pixel-wise color mapping, which restricts it to global tone adjustments while preventing local refinements. Extending the framework to support local adjustments is a promising direction for future work.

9. Additional Qualitative Comparisons

In this section, we present additional visual examples of various methods. The comparison results of image-guided, text-guided, and multimodal fusion-guided photorealistic style transfer are demonstrated in Figure 17, Figure 18, and Figure 19, respectively.

As shown in Figure 17 and Figure 18, our method stands out by consistently transferring the target style based on image or text guidance, and effectively preserving the texture and structure of content images. Furthermore, Figure 19 showcases the effectiveness of our approach in handling multimodal fusion applications. The stylized images are generated through the interpolation of different reference style images and text descriptions. The results are not only semantically reasonable but also visually appealing, underscoring the versatility of our method in diverse scenarios.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023. 1, 2, 6, 7

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 97–104, 2011. 4

[3] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8795–8805, 2024. 1, 2

[4] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. *arXiv preprint arXiv:2404.12154*, 2024. 6

[5] Kibeom Hong, Seogkyu Jeon, Junsoo Lee, Namhyuk Ahn, Kunhee Kim, Pilhyeon Lee, Daesik Kim, Youngjung Uh, and Hyeran Byun. Aespa-net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22758–22767, 2023. 2

[6] Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson WH Lau. Neural preset for color style transfer. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14173–14182, 2023. 1, 2, 4

[7] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18062–18071, 2022. 1, 2, 6, 7

[8] Hyeongmin Lee, Kyoungkook Kang, Jungseul Ok, and Sunghyun Cho. Cliptone: Unsupervised learning for text-based image tone adjustment. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2942–2951, 2024. 1, 2

[9] Mujing Li, Guanjie Wang, Xingguang Zhang, Qifeng Liao, and Chenxi Xiao. D-lut: Photorealistic style transfer via diffusion process. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9206–9214, 2025. 1, 2

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 4

[11] Fan Chen Xiaoying Xing Longyin Wen Chen Chen Sijie Zhu Ming Li, Xin Gu. Superedit: Rectifying and facilitating supervision for instruction-based image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1, 2

[12] ITU-R BT.500 Recommendation. Methodologies for the subjective assessment of the quality of television images. 2019. 1

[13] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, pages 34–41, 2001. 1, 2

[14] Hanyu Wang, Pengxiang Wu, Kevin Dela Rosa, Chen Wang, and Abhinav Shrivastava. Multimodality-guided image style transfer using cross-modal gan inversion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4976–4985, 2024. 6

[15] Linfeng Wen, Chengying Gao, and Changqing Zou. Cap-vstnet: Content affinity preserved versatile style transfer. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18300–18309, 2023. 1, 2

[16] 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. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 327–342, 2020. 1, 2

[17] Zipeng Xu, Songlong Xing, Enver Sangineto, and Nicu Sebe. Spectralclip: preventing artifacts in text-guided style transfer from a spectral perspective. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5121–5130, 2024. 1, 2

[18] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22873–22882, 2023. 1, 2

[19] 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 (ICCV)*, pages 9036–9045, 2019. 1, 2