

## A APPENDIX

### A.1 ABLATION STUDY

We show two ablations of StylerDALL-E. Firstly, we ablate the usage of captions in formulating the prompt-based reward during the style-specific fine-tuning stage (Sec. 4.2). In more detail, instead of using the CLIP similarity between the stylized image  $I^s$  and the prompt  $t_q$  (which combines the style description  $t_s$  and the image caption  $t_a$ ) as the reward, we discard  $t_a$  and we compute the CLIP similarity between the stylized image  $I^s$  and the style description  $t_s$  as the reward. Secondly, we ablate the importance of down-sampling as introduced in Sec. 4. Specifically, we directly input the discrete tokens of the full-resolution image to the NAT model while conducting the same self-supervised pre-training and style-specific fine-tuning (i.e.,  $I' = I$ ).

In Fig. 6 we use the style “watercolor painting” and we show the comparison between the full StylerDALL-E method and the two ablation methods. By comparing the results of the full model and the “w/o captions”, we see that the results of the full model are slightly better. In the results of StylerDALL-E, the details are preserved better, and the colors are closer to the light and muted colors used in watercolor painting. Moreover, the results are overall harmonious as there are few abrupt brushstrokes. Nevertheless, “StylerDALL-E w/o captions” also present a satisfying style transfer quality, as the results keep a good balance between the stylization and content maintainness. This indicates our method can also work when no caption is provided, thus being less annotation-dependent. Finally, the results of “StylerDALL-E w/o scaling” show the importance of the scaling procedure in StylerDALL-E: when the NAT model is input with the discrete tokens of the full-resolution image, the style cannot be incorporated effectively through the Reinforcement Learning fine-tuning stage.



Figure 6: Ablation study on StylerDALL-E.

### A.2 ADDITIONAL RESULTS

In Fig. 7 we illustrate the additional comparing results between StylerDALL-E and CLIPStyler-Optimization (i.e., the mainly proposed method in the paper). As shown, CLIPStyler-Optimization suffers from two issues. Firstly, there are many inharmonious artifacts that appear in the stylized images. For example, there are many plant-like artifacts in stylized results of “Monet” and multiple suns in the “Monet Sun Impression” results. Secondly, the texts related to the language instructions are written in the stylized images unexpectedly. For instance, as in the top example of the “fauvism” train, the written text “fauvism” is on the front of the bus. In the middle example of “Monet”, there are also written texts shown on the train body and the building.

By contrast, StylerDALL-E does not have the above two issues. Furthermore, our results achieve well-characterized stylization results consistent with language instructions, and different styles are expressed with varied and distinctive brushstrokes related to the specific style.

In addition, we give more stylized results produced by StylerDALL-E in Fig. 8 Fig. 9 and Fig. 10. In particular, we also show the intermediate results  $\hat{I}$  (as in Fig. 2), which are generated with the out-

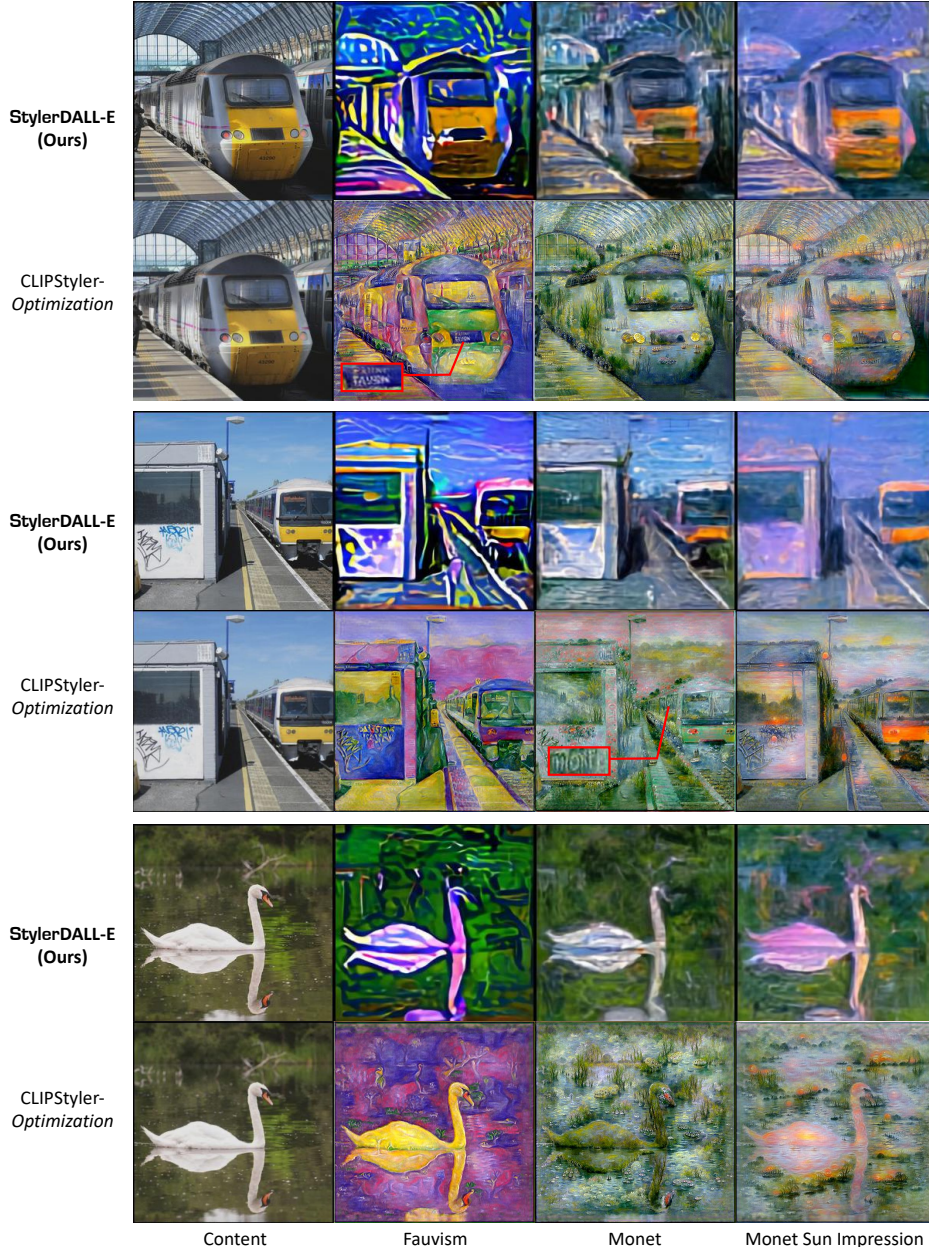


Figure 7: Comparisons between StylerDALL-E and CLIPStyler, styles are shown on the bottom.

put tokens using the model right after the self-supervised pre-training (and before the style-specific fine-tuning stage). Similar to what we have concluded, StylerDALL-E achieves distinctive, diverse, and harmonious stylized results on various styles and images. Besides, the differences between  $\hat{I}$ s and  $I^s$ s are significant. As shown,  $\hat{I}$ s are photo-realistic while  $I^s$ s present varied brush-strokes, edges, and colors with respect to each style instruction, indicating that StylerDALL-E has been effectively fine-tuned with our language-guided rewards in the Reinforcement Learning stage.



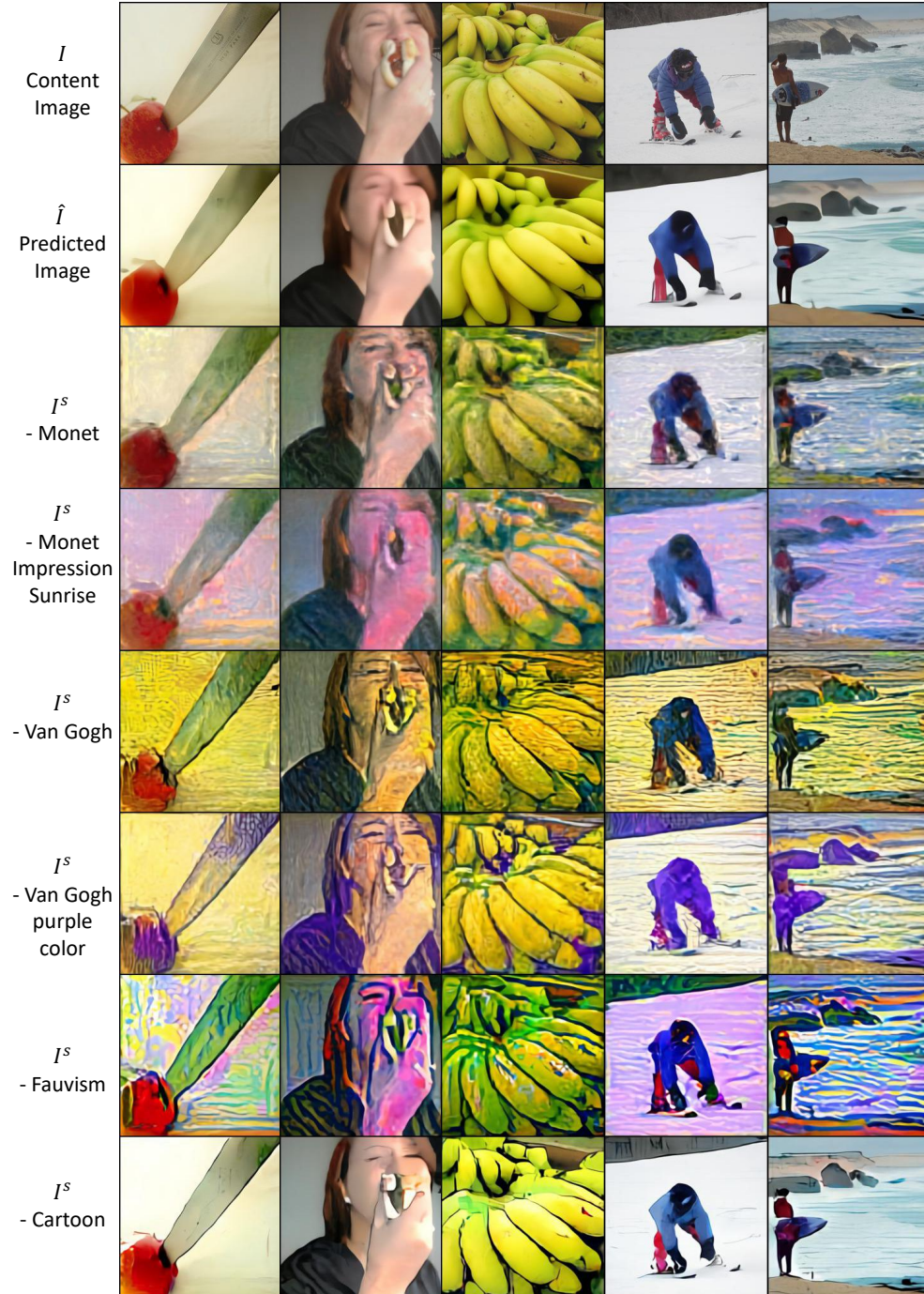


Figure 8: Additional stylized results of StylerDALL-E.

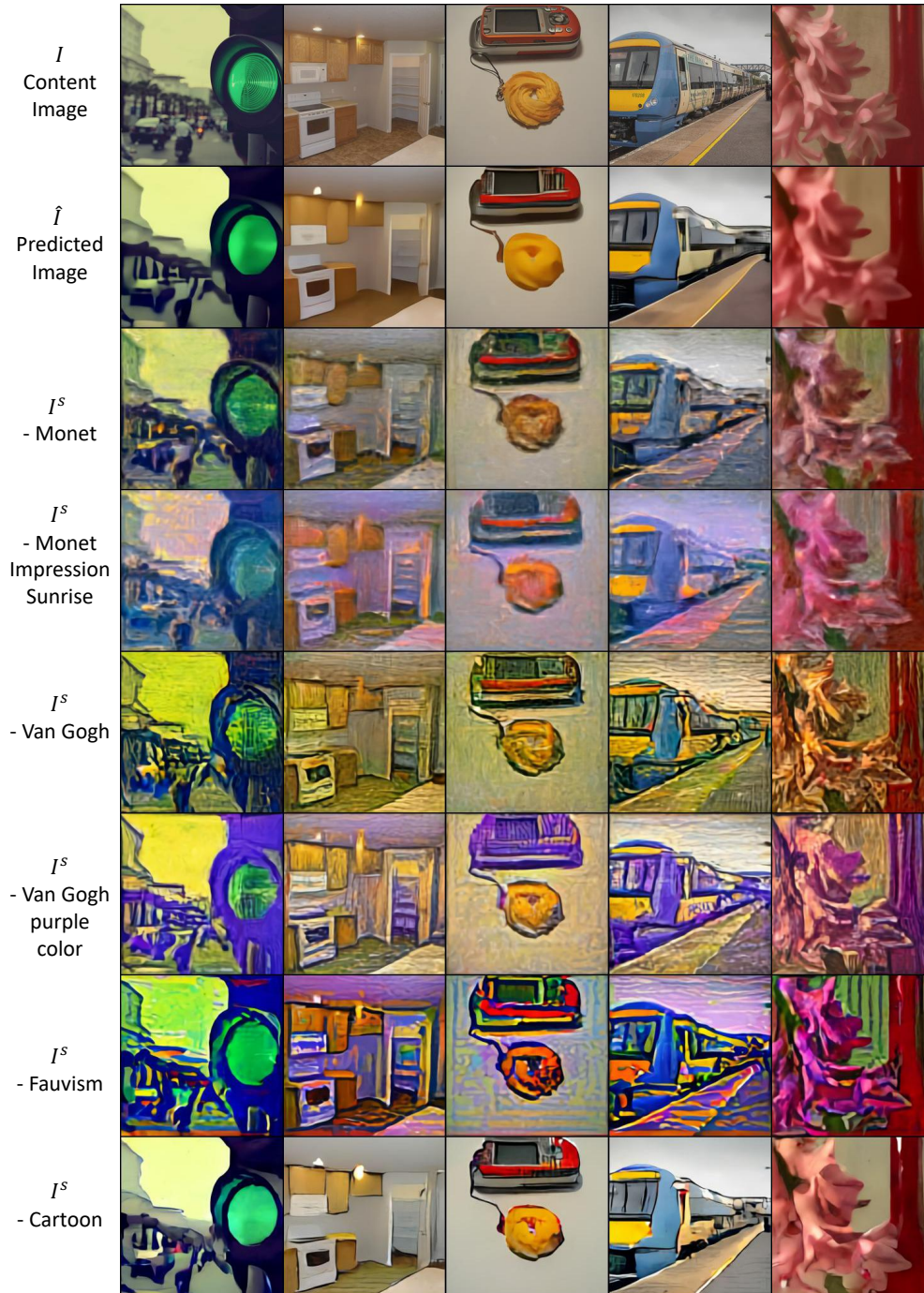


Figure 9: Additional stylized results of StylerDALL-E.



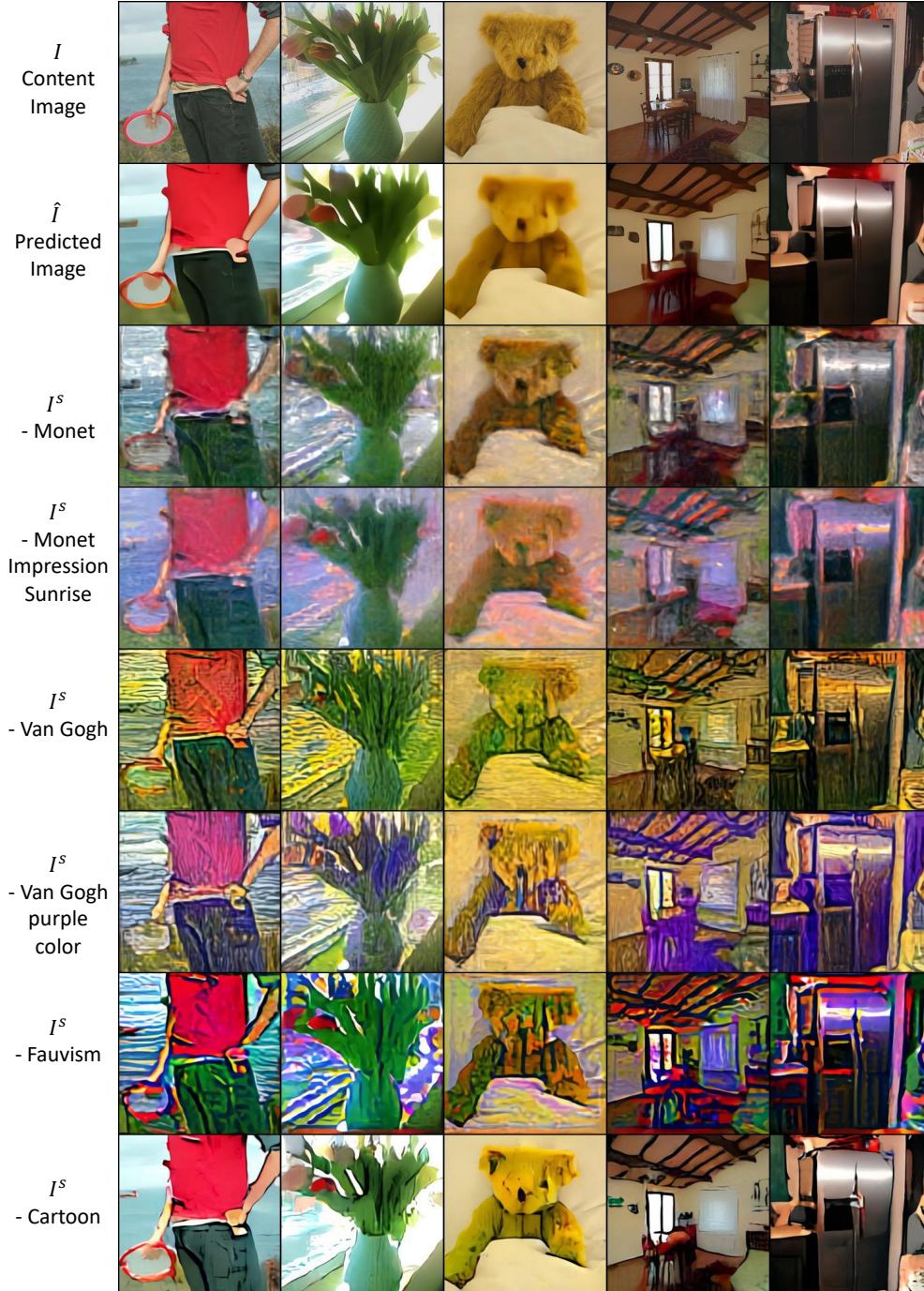


Figure 10: Additional stylized results of StylerDALL-E.

### A.3 IMPLEMENTATION DETAILS

The StylerDALL-E NAT model consists of a 4-layer encoder and an 8-layer decoder while the attention head number is 8 and the hidden dimension is 512. We use Pytorch (Paszke et al., 2019) to implement our method. To train our model, we use the train-set of COCO (Lin et al., 2014) dataset, which contains 82,783 images while each image has 5 captions. In the self-supervised pre-training stage, we only use the images in the COCO train-set. We train the NAT model for 25 epochs with a learning rate of  $1e-4$ . We use Adam (Kingma & Ba, 2014) optimizer. In the style-specific fine-tuning stage, we use both the images and the captions. In particular, we utilize all the caption annotations to enhance the model robustness, as usually human annotates different captions of a single image. Notably, the caption annotations are only used at fine-tuning stage. In other words, StylerDALL-E does not need to use image caption as input at inference time,. We only fine-tune the decoder of the NAT model, keep the encoder frozen, and train the model for 5 epochs. We use Adam optimizer with a learning rate of  $1e-6$ . We use the officially released dVAE of DALL-E<sup>1</sup> and the CLIP ViT-B/32 model<sup>2</sup>. For both training stages, the models are trained on single A6000 GPUs for 24 hours.

To compare with CLIPStyler, we use the official implementation<sup>3</sup>. For all reference image-based comparing methods, we use the officially released trained models<sup>4</sup>.

<sup>1</sup>DALL-E: <https://github.com/openai/dall-e>

<sup>2</sup>CLIP: <https://github.com/openai/CLIP>

<sup>3</sup>CLIPStyler: <https://github.com/cyclomon/CLIPstyler>

<sup>4</sup>AesUST: <https://github.com/EndyWon/AesUST> StyTr2: <https://github.com/diyiiyii/STyTR-2> AST: <https://github.com/CompVis/adaptive-style-transfer>