

---

# Supplementary Materials for Styl3R: Instant 3D Stylized Reconstruction for Arbitrary Scenes and Styles

---

Anonymous Author(s)

Affiliation

Address

email

1 In the supplementary, we provide the following:

- 2 • more implementation details, including architecture and training hyperparameters for our  
3 model in Sec. [A](#);
- 4 • more technical details on training 2D and 3D baselines in Sec. [A](#);
- 5 • more visual results of our model and comparisons with baselines in Sec. [B](#).

6 *We highly recommend watching the **accompanying video** for a more comprehensive understanding of*  
7 *our method’s stylization quality and temporal consistency. The video includes:*

- 8 • qualitative comparisons with baselines on the RE10K [\[12\]](#) and Tanks and Temples [\[3\]](#)  
9 datasets;
- 10 • more out-of-domain stylization results on the Tanks and Temples and NeRF LLFF [\[8\]](#)  
11 datasets;
- 12 • style interpolation examples showcasing transitions between three different styles.

## 13 **A More Implementation Details**

14 **Training.** In terms of optimization, we employ AdamW optimizer. For Novel View Synthesis  
15 (NVS) pretraining, we train the stylization decoder, color head and structure head with initial learning  
16 rate of  $2 \times 10^{-4}$ , and fine-tune the other parameters with  $2 \times 10^{-5}$ . Then for stylization fine-tuning,  
17 we continue optimizing the color head and stylization decoder with initial learning rate of  $2 \times 10^{-4}$   
18 and fine-tune only the style encoder with  $2 \times 10^{-5}$ , and keep all the other parameters in the structure  
19 branch fixed.

20 **Architecture.** To expedite the inference of network, we use the flash attention implementation from  
21 xFormers [\[4\]](#) in all of our encoders and decoders. As in [\[10\]](#), we feed the tokens from the 1-st, 7-th,  
22 10-th and 13-rd block into DPT [\[9\]](#) for upsampling.

23 **Baselines Training.** For the 2D methods (StyTr2 [\[1\]](#), AdaIN [\[2\]](#), AdaAttN [\[7\]](#)), we directly utilize  
24 their publicly released pretrained checkpoints, available at [StyTr2 code](#), [AdaIN code](#), and [AdaAttN](#)  
25 [code](#), respectively. As these methods do not support 3D reconstruction from 2D images, we apply  
26 them directly to stylize the ground-truth 2D novel views, bypassing any reconstruction process.

27 In contrast, the 3D methods (ARF [\[11\]](#), StyleRF [\[5\]](#), StyleGaussian [\[6\]](#)) are unable to reconstruct  
28 geometry from sparse, unposed inputs. Therefore, we train them using all available scene images (on  
29 average, more than 100 per scene) along with their corresponding camera poses. This setup results in  
30 an unfair comparison with our method, which operates on sparse and unposed inputs.

## B More Visual Results

We present additional qualitative results in Fig. 1, Fig. 2, Fig. 3, Fig. 4, and Fig. 5, which highlight the superior performance of our method compared to prior state-of-the-art style transfer approaches. While existing methods rely on densely posed images, our approach enables instant 3D stylization across arbitrary scenes and styles without such constraints.

**More Comparisons with Baselines.** To better showcase the superiority of our method, we visualize more comparison results with 3D baseline methods on different scenes and styles, as shown in Fig. 1.

**More Visual Results of Our Method.** To validate our method is compatible with arbitrary scenes and styles, we show stylization results with exhaustive combinations from randomly selected scenes and styles, as shown in Fig. 2, Fig. 3, Fig. 4 and Fig. 5.

## References

- [1] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022.
- [2] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.
- [3] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.
- [4] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xformers: A modular and hackable transformer modelling library. <https://github.com/facebookresearch/xformers>, 2022.
- [5] Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8338–8348, 2023.
- [6] Kunhao Liu, Fangneng Zhan, Muyu Xu, Christian Theobalt, Ling Shao, and Shijian Lu. Stylegaussian: Instant 3d style transfer with gaussian splatting. In *SIGGRAPH Asia 2024 Technical Communications*, pages 1–4, 2024.
- [7] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.
- [8] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019.
- [9] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.
- [10] Botao Ye, Sifei Liu, Haofei Xu, Li Xueting, Marc Pollefeys, Ming-Hsuan Yang, and Peng Songyou. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [11] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *European Conference on Computer Vision*, pages 717–733. Springer, 2022.
- [12] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.



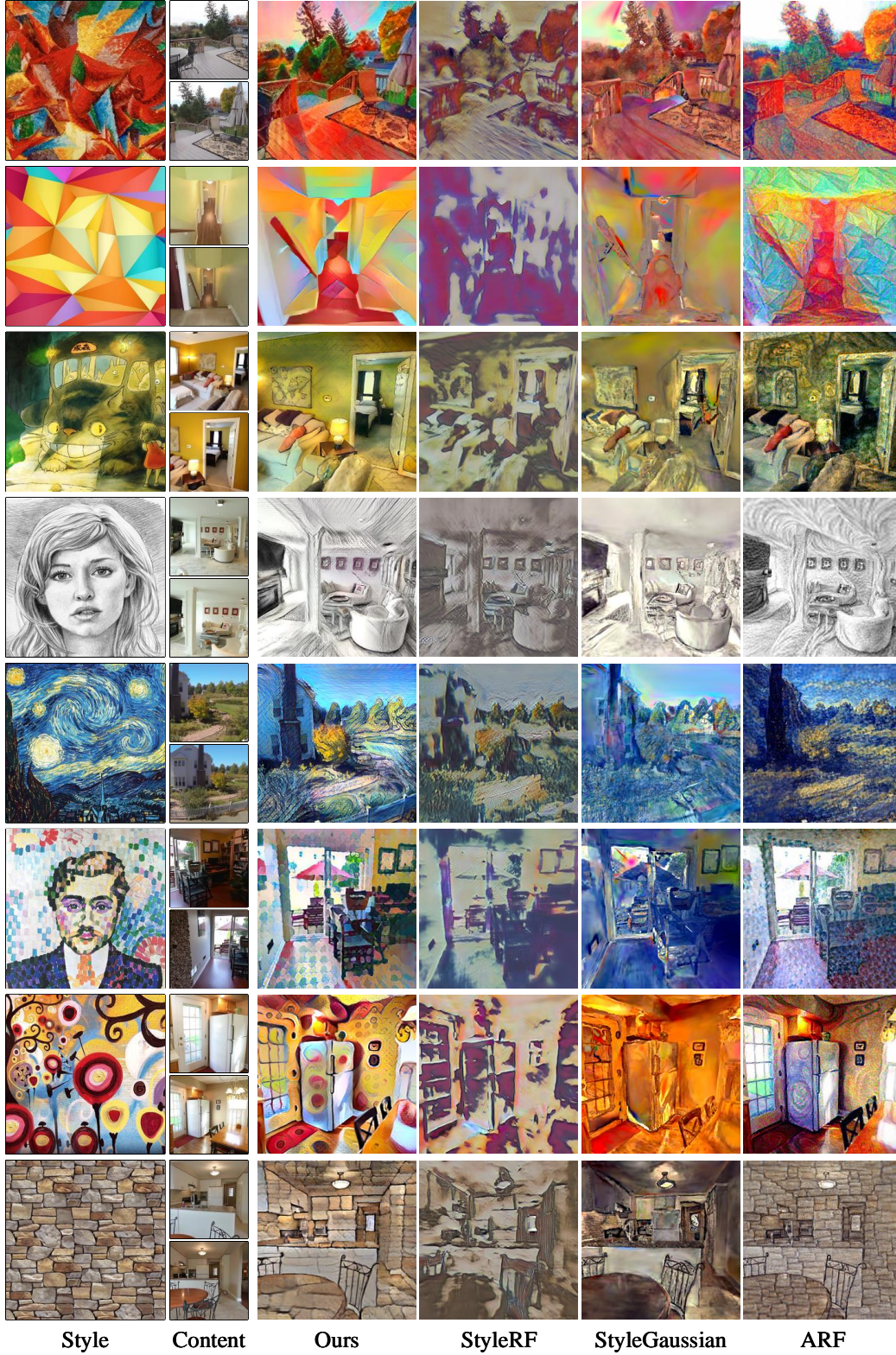


Figure 1: **Novel View Transfer Comparison on RE10K.** Our method faithfully preserves style and scene structure, even with limited image overlap. In contrast, StyleRF [5] and StyleGaussian [6] produce over-smoothed results with inaccurate color tones, while ARF [11] suffers from style overflow.



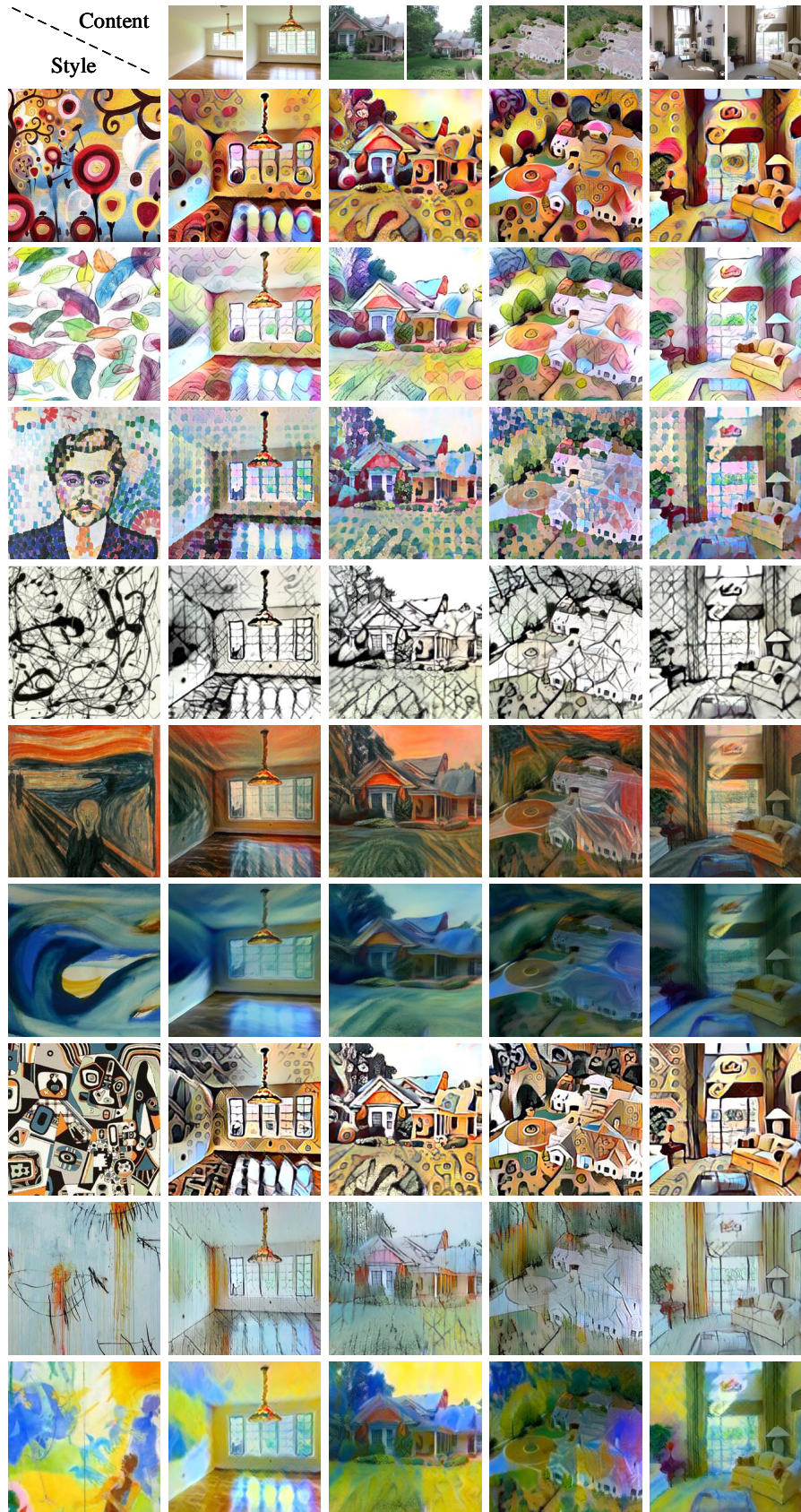
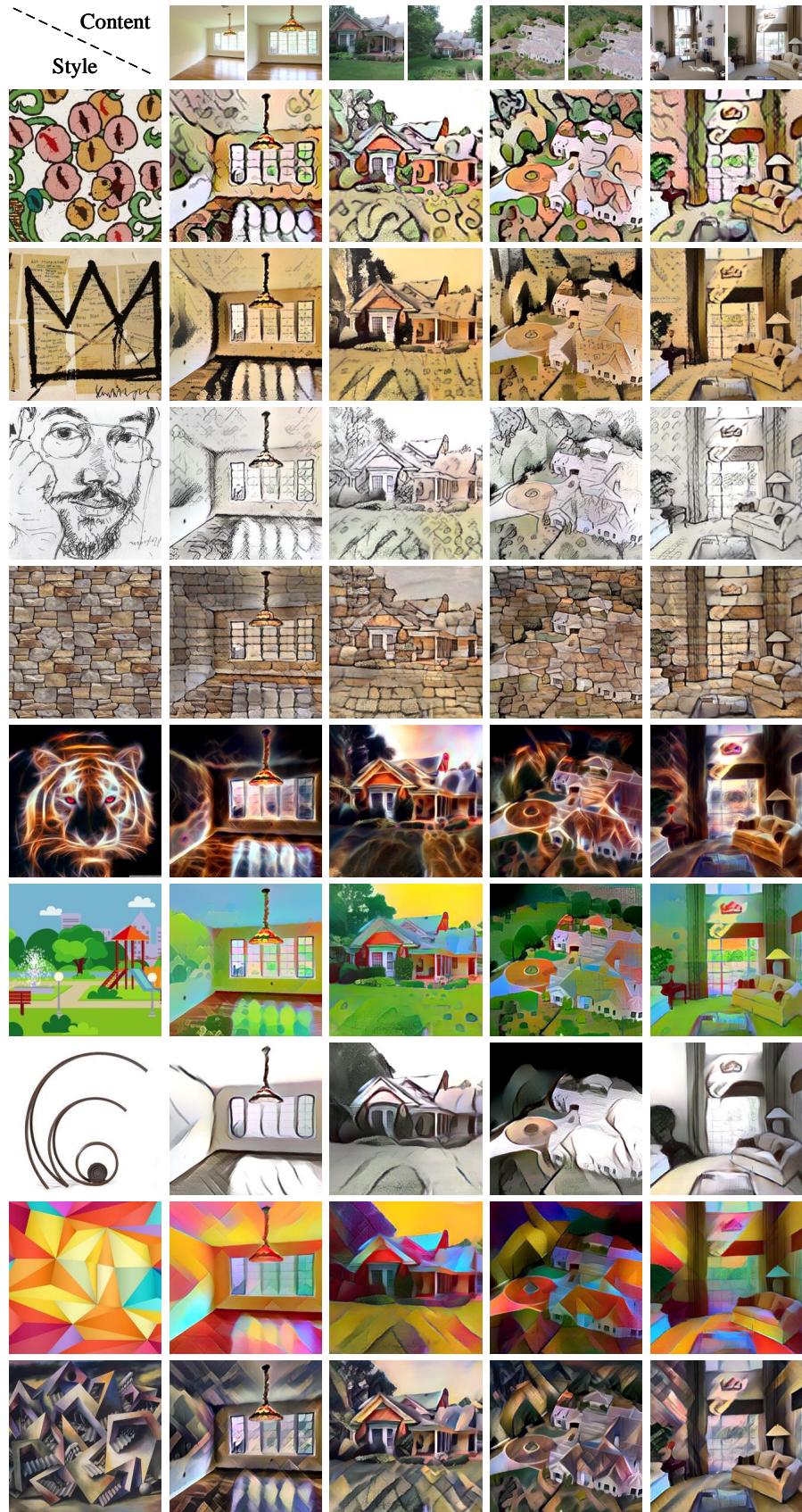


Figure 2: Additional Results on RE10K from Our Method.







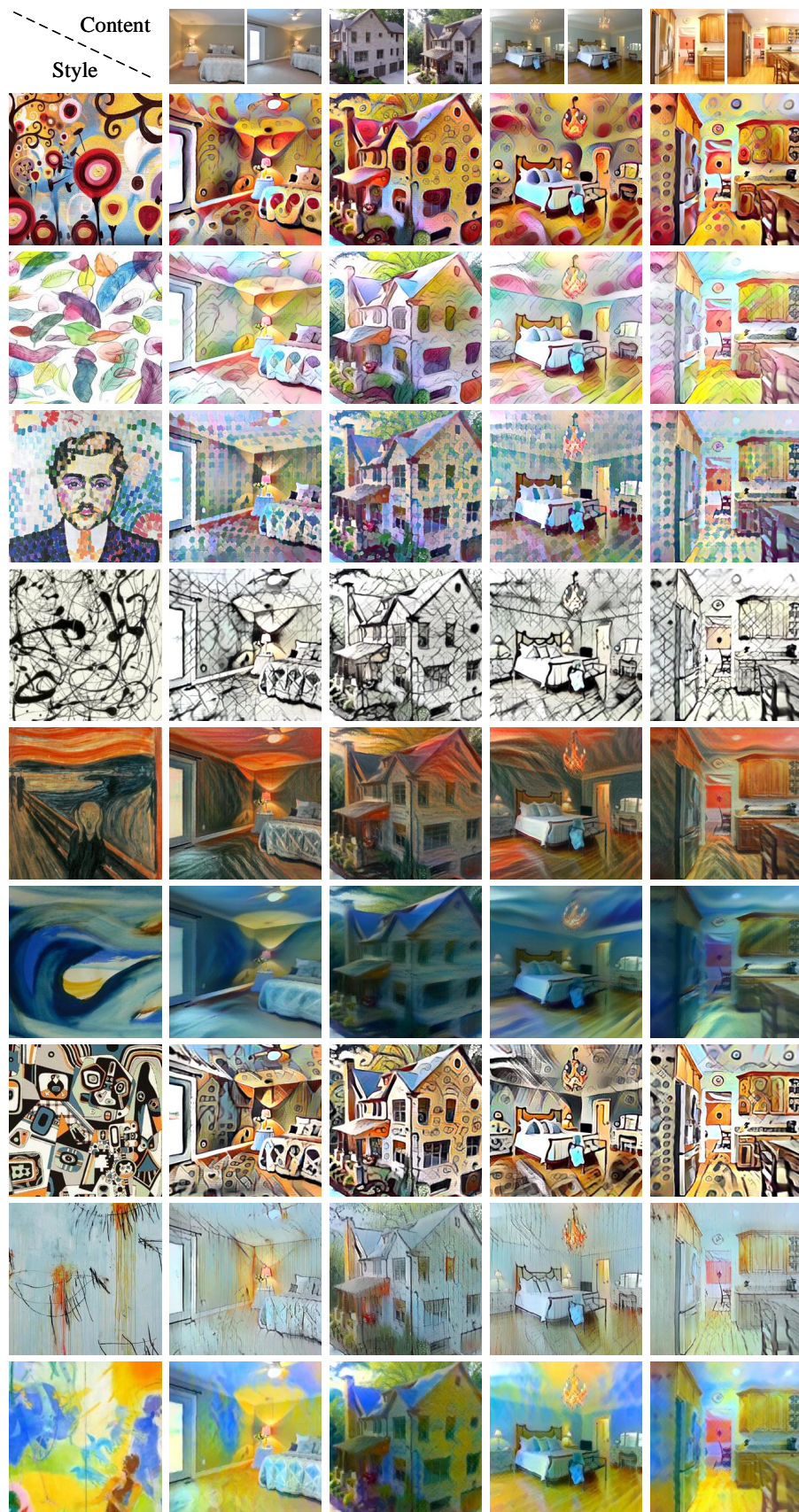


Figure 4: Additional Results on RE10K from Our Method.





Figure 5: Additional Results on RE10K from Our Method.