Ref-NPR: Reference-Based Non-Photorealistic Radiance Fields for Controllable Scene Stylization: Supplementary Material

A. Supplementary Materials

We have prepared supplementary materials, including a document and a video, to provide a more comprehensive understanding of Ref-NPR. In the document, we discuss the technical details of our implementation in Sec. B, and provide visualizations in Sec. C to better illustrate the proposed modules in Ref-NPR. Moreover, we present additional examples and visualizations in Sec. D to demonstrate the performance and controllability of our method. Furthermore, we have prepared a video that showcases the results and comparisons of Ref-NPR. We also provide live demo examples on the project page.

Video link `https://youtu.be/jnsnrTwVSBw`.
Project page `https://ref-npr.github.io`.

B. Technical Details

Implementation details. Two worth-noting details may affect the visual quality of stylization results when implementing Ref-NPR.

- Before computing image-level loss terms ($\mathcal{L}_{\text{color}}$ and $\mathcal{L}_{\text{feat}}$), for LLFF [5] and T&T [4] dataset, we downsample both stylized and content views by 2x to speed up the calculation of patch-wise feature distance.

- Different from the implicit feature loss $\mathcal{L}_{\text{feat}}$, in order to get a high-level semantic color mapping for the color-matching loss $\mathcal{L}_{\text{color}}$, we evaluate distances between features extracted by the last stage (i.e., stage 5) of VGG backbone [7]. Besides, when calculating $\mathcal{L}_{\text{color}}$, we exclude the position of interest (i, j) where the semantic feature is not close enough to any feature in the reference view, to avoid over-matching. Such a constraint of the feature distance for valid position (i, j) is formulated as

$$\min_{i',j'} dist(F_I^{(i,j)}, F_{I_R}^{(i',j')}) < 0.4\,. \tag{1}$$

Details of comparison. Our experiments on Texler [8] are conducted using their official implementation. As the reference view can be freely chosen, it is possible that continuous views with high-quality temporal coherence do not appear in the test sequence. Therefore, we only use the RGB image sequence as input and follow the default training settings by training each scene for 30,000 iterations. However, it is important to note that Texler's method is unsuitable for videos with large movements and rotations, and we train it on the template view. Despite applying the Gaussian mixture strategy with a dense sample rate, error accumulation still leads to artifacts in the output.

Regarding SNeRF [6], we re-implement it based on Plenoxels [2] and use Gatys [3] as the stylization method. We train the stylization step for 10 iterations and the entire scene stylization for 10 epochs for each training view.

Quantitative comparison. In Sec. 4.3, we propose a reference-based perceptual similarity metric to evaluate our method. The detailed LPIPS scores for each scene are reported in Tab. B.1. It is worth noting that the scene-wise LPIPS scores exhibit significant variations. We speculate that these fluctuations may be due to the substantial differences in camera poses between the reference view and all other test views. Additionally, Texler [8] achieves slightly better reference-related LPIPS scores. However, it fails to produce satisfactory results when the camera pose diverges significantly from the reference camera φ_R, as demonstrated in Fig. D.7 and the supplementary video.

Fig. B.1 (a) depicts the procedure of the designed LPIPS evaluation in the paper. As only the reference image is given to evaluate the visual quality, we utilize LPIPS referring to CCPL [9] as a metric for frame-wise stylization consistency. The closest ten frames represent a frame-wise consistency of stylization results with the given style reference. Fig. B.1 (b) depicts our experiments investigating the robustness of stylization methods. For a stylized NeRF ω_{NP}, we render a set of views as the style reference and use them to get a set of stylized NeRFs. Given the same camera path, we compute the PSNR of rendering results between them and ω_{NP}.

C. Method Visualizations

Reference ray registration. Fig. C.2 gives two concrete examples of how ray registration provides supervision in reference-dependent areas. Rays related to the stylized reference S_R are projected to each training view to provide pseudo-ray supervision.

Template-based feature matching. Except for explicit supervision in R^3, the implicit supervision provided by TCM is essential to occluded regions. Fig. C.3 shows two examples of patch-wise replacement results. For guidance feature F_G, we select VGG features at stages 3 and 4. Since

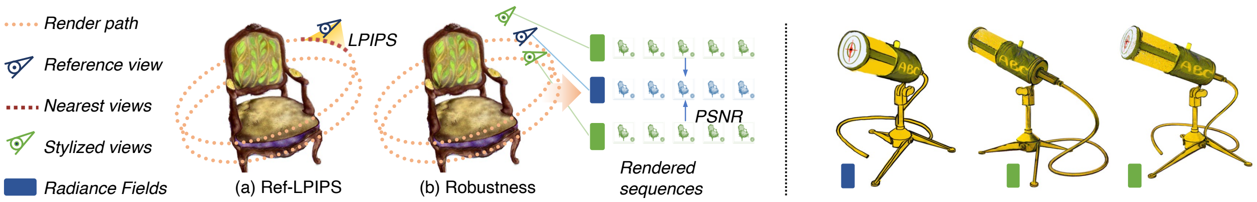

Figure B.1. Left: An illustration of the Ref-LPIPS and robustness test. Right: Tested views for robustness.

Ref-LPIPS ↓	Geo. Consist.	Chair	Ficus	Hotdog	Mic	Flower	Horn	Truck	Playground	Average
Texler [8]	✗	0.167	0.120	0.216	0.119	0.230	0.488	0.667	0.675	**0.335**
ARF [10]	✓	0.185	0.123	0.300	0.146	0.619	0.502	0.683	0.592	0.394
SNeRF [6]	✓	0.188	0.129	0.283	0.138	0.646	0.492	0.702	0.663	0.405
Ref-NPR	✓	0.164	0.122	0.273	0.126	0.289	0.471	0.669	0.596	**0.339**

Table B.1. Reference-related novel view LPIPS for each test scene.

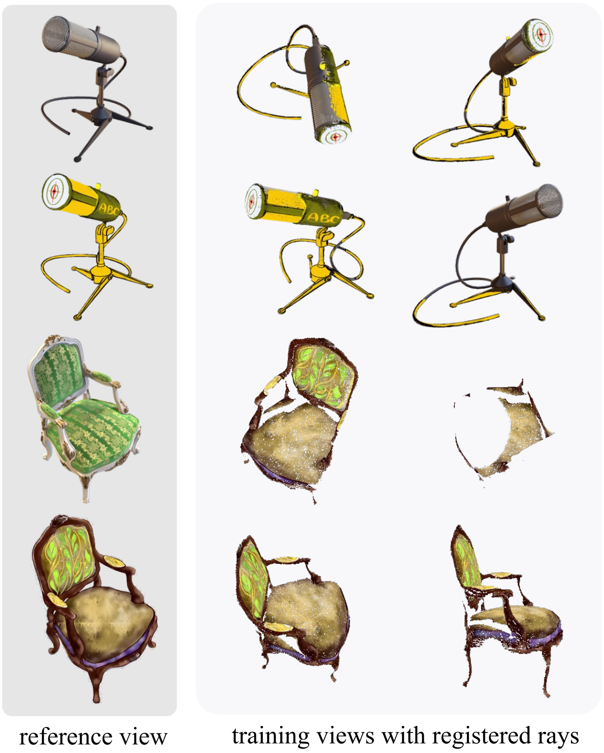

Figure C.2. Two examples to visualize registered rays in R^3. We paste pseudo-rays on content images in the first example for a better presentation.

the patch-wise semantic feature is a high-level representation for each patch, the receptive field is much larger than the corresponding image patch.

Conversely, directly using patch replacement results at the same stages for the color supervision $\mathcal{L}_{\text{color}}$ may result in a color mismatch problem, as highlighted in Fig. C.2.

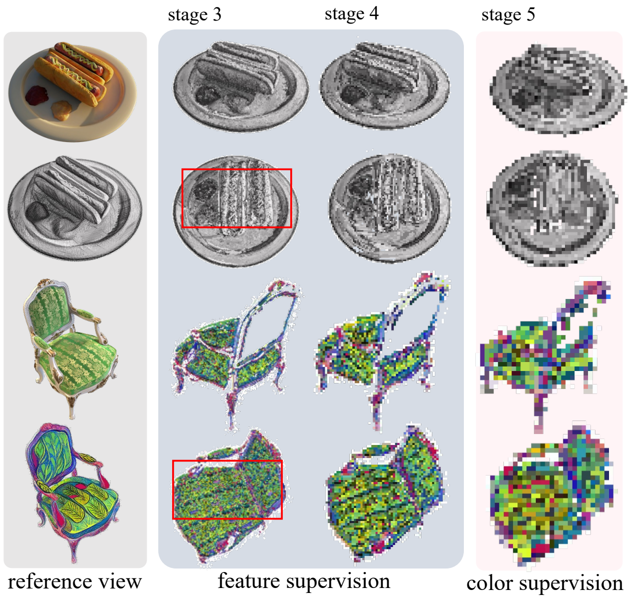

Figure C.3. Two examples of patch-wise replacement on VGG feature at the last three stages to visualize the semantic correspondence. Color mismatch problems in shallow semantic features are highlighted.

This problem is mainly caused by the receptive field difference between the feature patch and the image patch. Hence, as mentioned in Sec. B, we evaluate feature distances at the last VGG stage for color-matching supervision.

Loss balancing ablation. In addition to the ablation studies on the microphone example provided in Sec. 4.4, we conduct another ablation on the scene flower to discuss the effectiveness of color-matching loss $\mathcal{L}_{\text{color}}$ and the smooth content update strategy, which is described in Sec. 4.1.

For the same content view in Fig. C.4 (a), the color

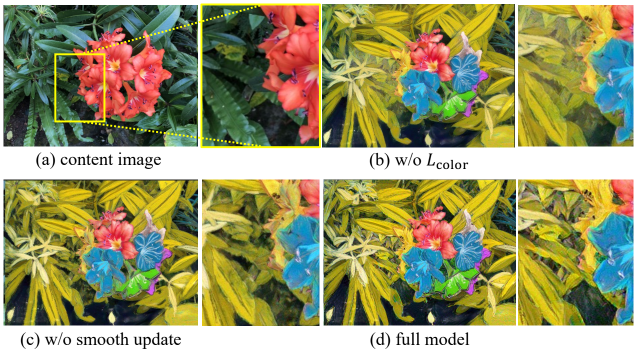

Figure C.4. Ablation on the color-matching loss and smooth update strategy. The occluded region is zoomed in.

mismatch problem would exist in occluded regions when we remove the color-matching loss $\mathcal{L}_{\text{color}}$, as shown in Fig. C.4 (b). In Fig. C.4 (c), we find that the stylized view without applying the smooth update strategy leads to occluded regions being under-stylized, which implies that the quality of semantic correspondence in the original content domain needs to be enhanced by TCM. A full model in Fig. C.4 (d) clearly shows a satisfying stylization result in terms of both color and style.

Discussion on TCM matching. We also validate the how effectiveness of the patch-wise matching scheme in TCM. Unlike epipolar correspondence, the deep semantic feature is calculated in 2D patch-wisely. The correspondence is only computed once and costs around 2 seconds for a set with 100 images. As shown in Fig. C.5 (a), a direct match with the stylized view often fails to get desired correspondence due to the domain gap in the semantic feature space. Conversely, in Fig. C.5 (b), TCM matches features within the same content domain. Hence the semantic correspondence is preserved at each level of semantic features.

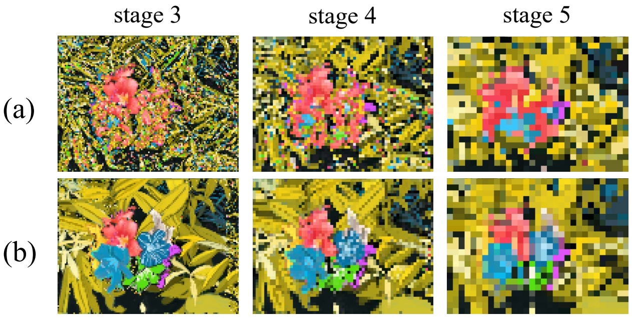

Figure C.5. Patch-wise replacement results on features from the last three stages of VGG backbone. (a) Matching with the style reference directly. (b) Matching with the content reference (TCM).

D. More Results

Comparsion with INS. We test INS [1] on examples with the same reference cases in Fig. 5 and Fig. D.10. Results are shown in Fig. D.6. Due to its simple supervision design, INS cannot generate satisfying results which contain local correspondence.

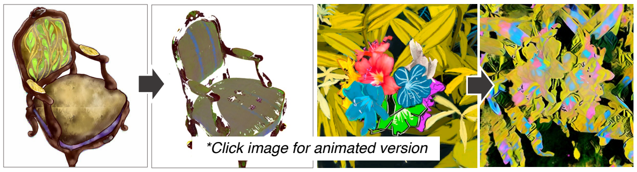

Figure D.6. Examples with INS.

More comparisons. Fig. D.7 offers two additional examples to compare our method with [6, 8, 10]. As discussed in Sec. 4.2, Texler can generate novel-view stylized results with a proper color distribution, but consistent results with the reference stylized view can be only obtained under the condition that the test camera pose is around the reference. More specifically, it fails to generate reasonable style in the occluded regions and has some flickering or ghosting artifacts in a continuous sequence. Two scene stylization methods [6, 10] are unable to find a desired style mapping to the entire scene. Neither in the reference-related regions nor the occluded regions. By contrast, results generated by Ref-NPR keep both semantic correspondence and geometric consistency with the reference view.

Flexibility & controllability. In Sec. 5, we show the ability of Ref-NPR to adapt with an arbitrary image as reference. Fig. D.10 gives two examples to demonstrate the flexibility of Ref-NPR, where the stylized reference view is generated by selecting one stylized view from ARF for each scene. In Fig. D.10 (a), we manually edit the selected view and take it as the style reference. Ref-NPR faithfully reproduces the textures in the edited regions. Meanwhile, as shown in Fig. D.10 (b), our method can reproduce the original novel-view stylizations by ARF through feeding in a stylized view as reference, which requires high-quality semantic correspondence.

Except for the local editing and scene stylization reproducing, the controllability of Ref-NPR can also be represented by adapting scene stylization to various styles. Fig. D.9 shows two examples of applying multiple styles to the same scene. Ref-NPR is capable of producing a faithful stylization result for each style owing to the modeling of cross-view semantic correspondence. Additionally, as shown in Fig. D.8, powered by controllable diffusion models [11], Ref-NPR is capable of text-driven controllable scene stylization as well.

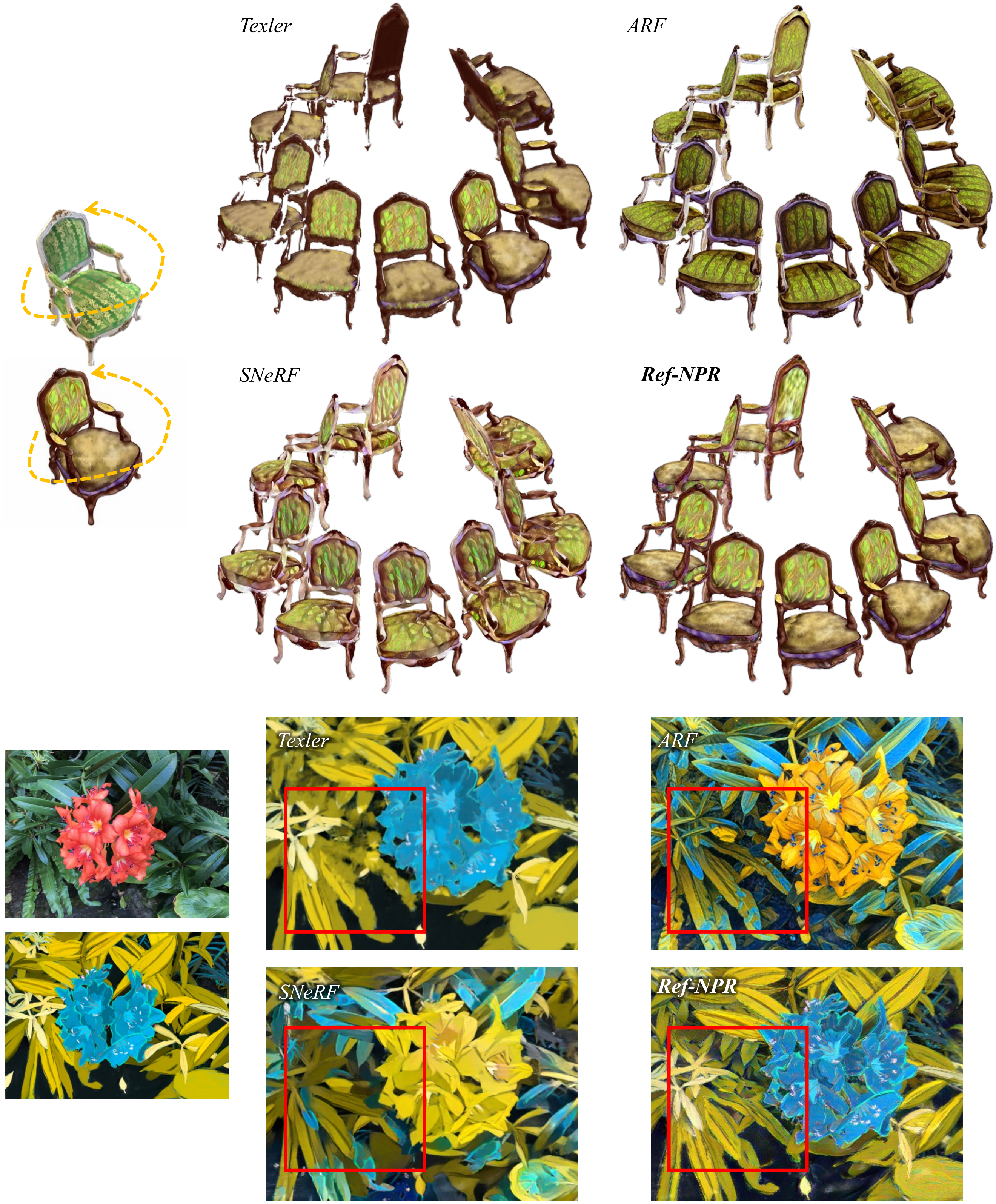

Figure D.7. Additional examples for qualitative comparisons.

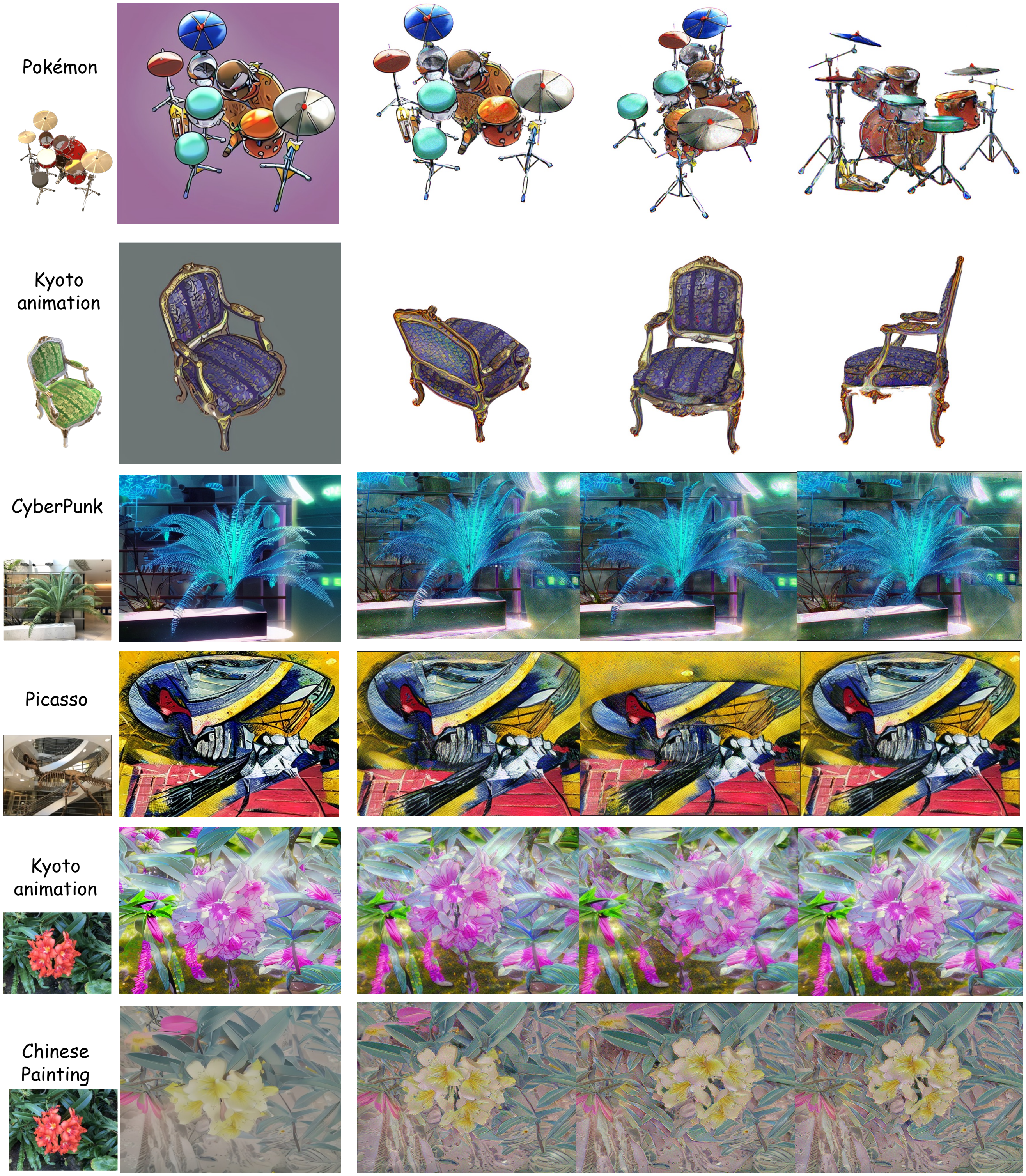

Figure D.8. Controllable scene stylization with ControlNet [11] and Ref-NPR. A text-driven stylization with an image diffusion model is used to generate reference (the second column), then Ref-NPR can propagate it to the whole scene.

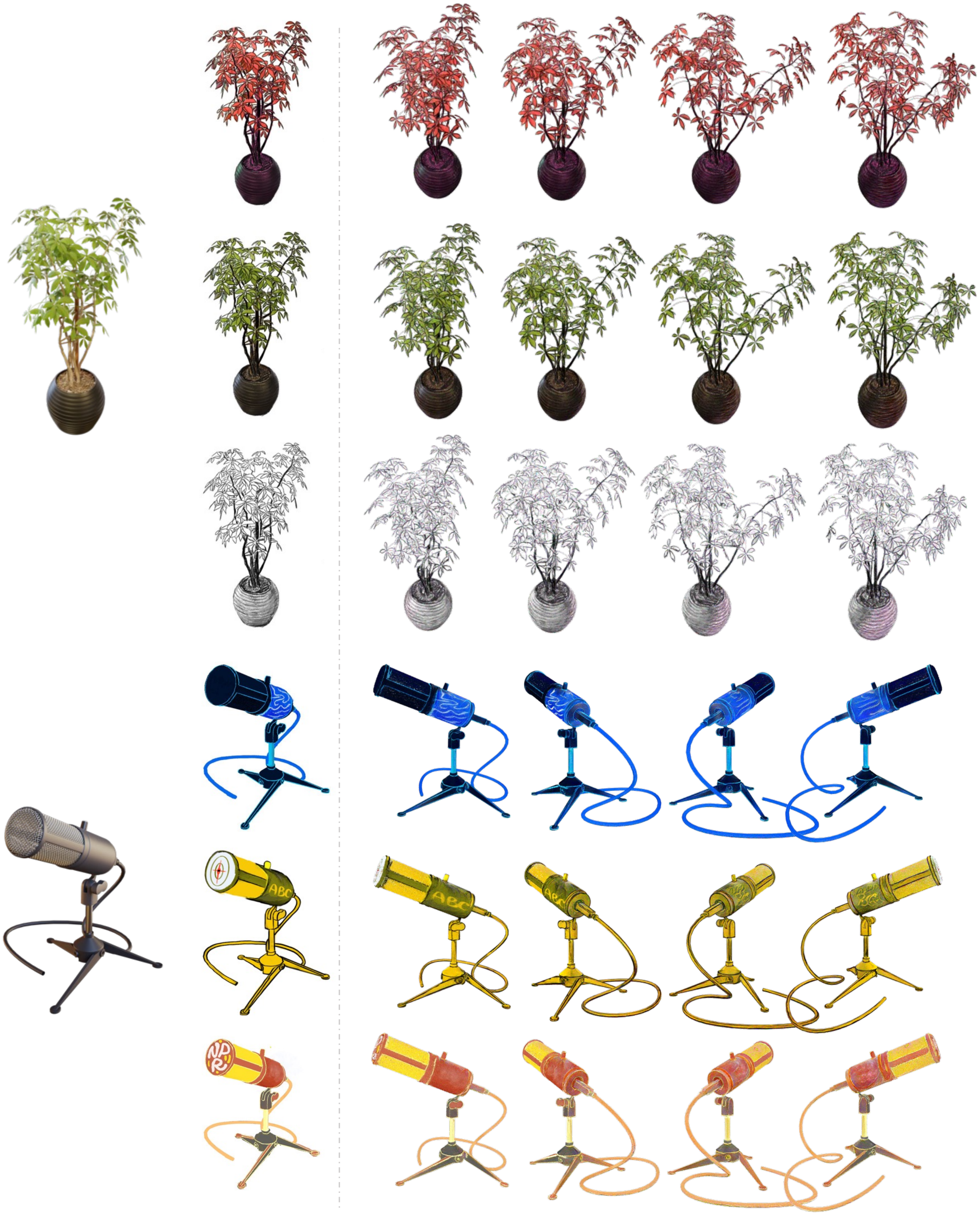

Figure D.9. Examples to show the controllability of Ref-NPR with hand drawing styles. Stylized novel-view rendering results are satisfactory with references in different styles.

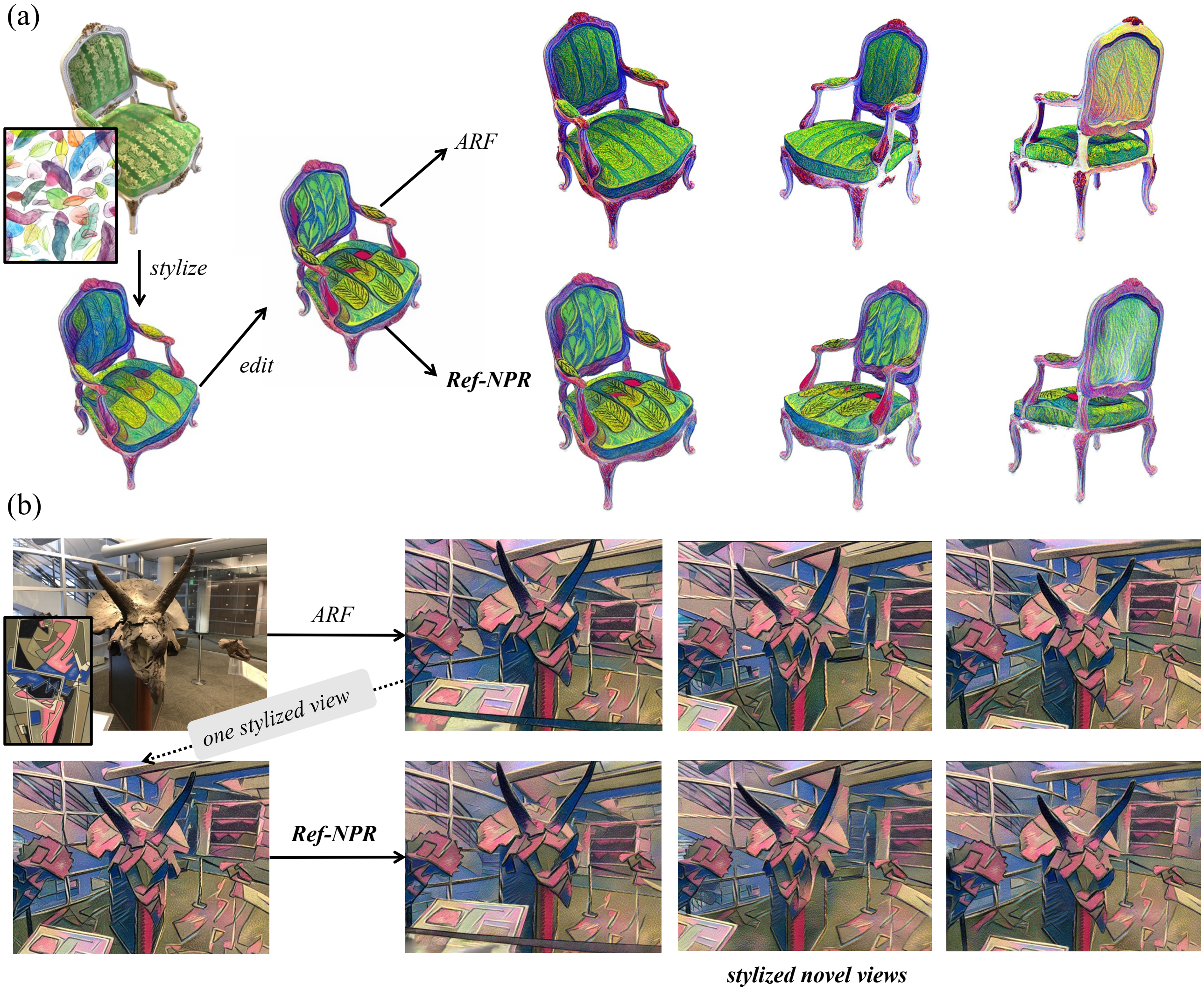

Figure D.10. Examples to show the flexibility of Ref-NPR: (a) reference editing based on a stylized view, and (b) reproducing novel-view stylization given one stylized view generated by ARF [10] as reference.

References

[1] Zhiwen Fan, Yifan Jiang, Peihao Wang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Unified implicit neural stylization. In *ECCV*, pages 636–654. Springer, 2022. 3

[2] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 1

[3] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 1

[4] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36(4), 2017. 1

[5] 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 Trans. Graph.*, 2019. 1

[6] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *arXiv preprint arXiv:2207.02363*, 2022. 1, 2, 3

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 1

[8] Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menglei Chai, Sergey Tulyakov, and Daniel Sýkora. Interactive video stylization using few-shot patch-based training. *ACM Trans. Graph.*, 39(4):73, 2020. 1, 2, 3

[9] Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. Ccpl: Contrastive coherence preserving loss for versatile style transfer. In *ECCV*, 2022. 1

[10] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 717–733. Springer, 2022. 2, 3, 7

[11] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3, 5