S-DyRF: Reference-Based Stylized Radiance Fields for Dynamic Scenes

Supplementary Material

For a comprehensive evaluation and demonstration of our method, please refer to our supplementary video. This document includes the following contents:

1. Background.
2. Technical details.
3. Geometry of stylized radiance fields.
4. User study.
5. More results.
6. Limitation discussion.

A. Background

Neural radiance fields. Neural radiance fields [5] are an implicit representation that maps a spatial location $\mathbf{x} = (x, y, z)$ and viewing direction $\mathbf{d} = (\theta, \phi)$ to its corresponding emitted color $\mathbf{c}$ and volume density σ:

$$F_\Theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma),$$

where F_Θ is usually implemented with multilayer perceptrons (MLPs). The color $\hat{C}$ of a camera ray $\mathbf{r} = \mathbf{o} + t\mathbf{d}$ with near and far bounds t_n and t_f is given by:

$$\begin{aligned} \hat{C}(\mathbf{r}) &= \int_{t_n}^{t_f} T(t)\, \sigma(\mathbf{r}(t))\, \mathbf{c}(\mathbf{r}(t), \mathbf{d})\, dt, \\ \text{where } \; T(t) &= \exp\left(- \int_{t_n}^{t} \sigma(\mathbf{r}(s))\, ds \right). \end{aligned} \tag{1}$$

During the training phase, the neural radiance field is optimized by minimizing the difference between the predicted pixel colors $\hat{C}(\mathbf{r})$ and the ground truth pixel colors $C(\mathbf{r})$:

$$\mathcal{L} = \sum_{\mathbf{r} \in \mathcal{R}} \left\| \hat{C}(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2. \tag{2}$$

4D scene representation. A straightforward approach to representing dynamic 3D scenes is to condition the neural radiance fields on the timestamps [8, 9]. Nevertheless, empirical observations suggest that this design often struggles to effectively capture complex dynamic 3D scenes. Inspired by TensoRF [2], recent methods [1, 4, 6] propose to utilize 4D volume factorization to represent dynamic 3D scenes. In particular, if we denote XYZ, T, and F as the spatial resolution, temporal resolution, and feature size, HexPlane [1] represents a 4D feature volume $\mathbf{D} \in \mathbb{R}^{XYZTF}$ as:

$$\begin{aligned} \mathbf{D} = \sum_{r=1}^{R_1} \mathbf{M}_r^{XY} \circ \mathbf{M}_r^{XY} \circ \mathbf{v}_r^1 + \sum_{r=1}^{R_2} \mathbf{M}_r^{XZ} \circ \mathbf{M}_r^{YT} \circ \mathbf{v}_r^2 \\ + \sum_{r=1}^{R_3} \mathbf{M}_r^{YZ} \circ \mathbf{M}_r^{XT} \circ \mathbf{v}_r^3, \end{aligned} \tag{3}$$

where $\circ$ is outer product and each $\mathbf{M}_r^{AB}$ is a learned plane of features. For a 3D point at time t, its density and appearance feature can be quired from HexPlane using bilinear interpolations and vector-matrix product. The final color is then predicted through a tiny MLP that takes the appearance features and view direction as input. To train HexPlane, a combination of photometric loss and regularization terms, including Total Variational (TV) loss, are employed:

$$\mathcal{L} = \sum_{\mathbf{r} \in \mathcal{R}, t \in \mathcal{M}} \left\| \hat{C}(\mathbf{r}, t) - C(\mathbf{r}, t) \right\|_2^2 + \lambda_{reg} \mathcal{L}_{reg}, \tag{4}$$

where $\mathcal{R}$ is the set of all training rays and $\mathcal{M}$ is the time frame set.

B. Technical Details

To the best of our knowledge, we are the first to tackle the task of stylizing neural radiance fields for dynamic 3D scenes with a stylized 2D view as a reference. We compare our method with three representative stylization methods. In this section, we provide a detailed description of the adaptations we make to these methods.

ARF*. ARF [10] is an arbitrary style transfer method for static 3D scenes. To ensure a fair comparison, we reimplement ARF and make adjustments to cater to dynamic scenes. This modified version is denoted as ARF*. Specifically, we replace Plenoxels [3] with a recently-proposed HexPlane [1] for its fast reconstruction and rendering speed of dynamic scenes. Notably, due to the inherent capability of the HexPlane representation to share information across different timesteps, the nearest neighbor feature matching loss can implicitly transfer style information not only across spatial dimensions but also across time dimensions to the entire scene.

Ref-NPR*. Ref-NPR [11] is a reference-based 3D stylization method with the ability to produce stylized novel views while preserving both geometric and semantic consistency given a stylized reference. Similarly, since it was initially designed for static 3D scenes, we also reimplement it and adapt it to dynamic scenes, which we denote as Ref-NPR*. Specifically, we substitute Plenoxels [3] with the more recently introduced HexPlane [1], chosen for its superior capabilities in reconstructing and rendering dynamic scenes. Following the insights from Ref-NPR [11], subsequent steps involve a reference-based ray registration process and a template-based feature matching scheme. These modules are employed to propagate style information across the scenes. The HexPlane representation, by inherently sharing information across timesteps, facilitates the implicit

transfer of style information to the entire scene during the aforementioned operations.

Video stylization method. We also include Texler et al. [7], a reference-based video stylization method that can propagate the stylized content from a stylized reference to the rest of the video sequence. We follow the official code[1] released on GitHub to synthesize stylized videos.

C. Geometry of Stylized Radiance Fields

As shown in Fig. C.1, we do not optimize density components, thereby ensuring that no alterations are made to the geometry.

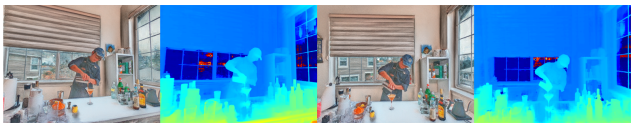

Figure C.1. **Stylized novel views and their corresponding depths.**

D. User Study

To evaluate the performance of our method, we organize a user study involving 56 participants with diverse backgrounds and expertise in the field. The study is conducted using an online website designed specifically for this purpose. A screenshot of the website interface is shown in Fig. D.2. Note that our user study is completely anonymous, and no personally identifiable data is collected from the participants. During the study, each participant is presented with a reference image, its stylized version, and two stylized videos: one generated by our method and the other from a randomly selected approach, with the order randomized. They are asked to click the better one that maintains semantic consistency with the provided stylized reference image, without noticeable flickering or artifacts, or none if it is hard to judge.

E. More Results

In this section, we present additional visual results in Fig. E.3, Fig. E.4, Fig. E.5, Fig. E.6, and Fig. E.7. One can observe that our method is capable of outputting visually pleasing stylized novel views and times while maintaining semantic consistency with the provided reference image across both spatial and temporal dimensions.

F. Limitation Discussion

Our method, S-DyRF, is a reference-based spatio-temporal stylization method for dynamic neural radiance fields, capable of outputting visually pleasing stylized novel views and times while maintaining semantic consistency with the provided reference image across both spatial and temporal dimensions. While our method can effectively transfer style information from the reference to the entire dynamic 3D scene, it faces challenges when flickering is present within the temporal pseudo-references, consequently resulting in subtle flickering in the rendered videos of our stylized radiance fields. Moreover, its performance may be compromised in the absence of meaningful semantic correspondence in the reference view.

[1] https://github.com/OndrejTexler/Few-Shot-Patch-Based-Training

References

[1] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–141, 2023. 1

[2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 333–350. Springer, 2022. 1

[3] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510, 2022. 1

[4] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12479–12488, 2023. 1

[5] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[6] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16632–16642, 2023. 1

[7] Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menclei Chai, Sergey Tulyakov, and Daniel Sỳkora. Interactive video stylization using few-shot patch-based training. *ACM Transactions on Graphics (TOG)*, 39(4):73–1, 2020. 2

[8] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994*, 2021. 1

[9] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9421–9431, 2021. 1

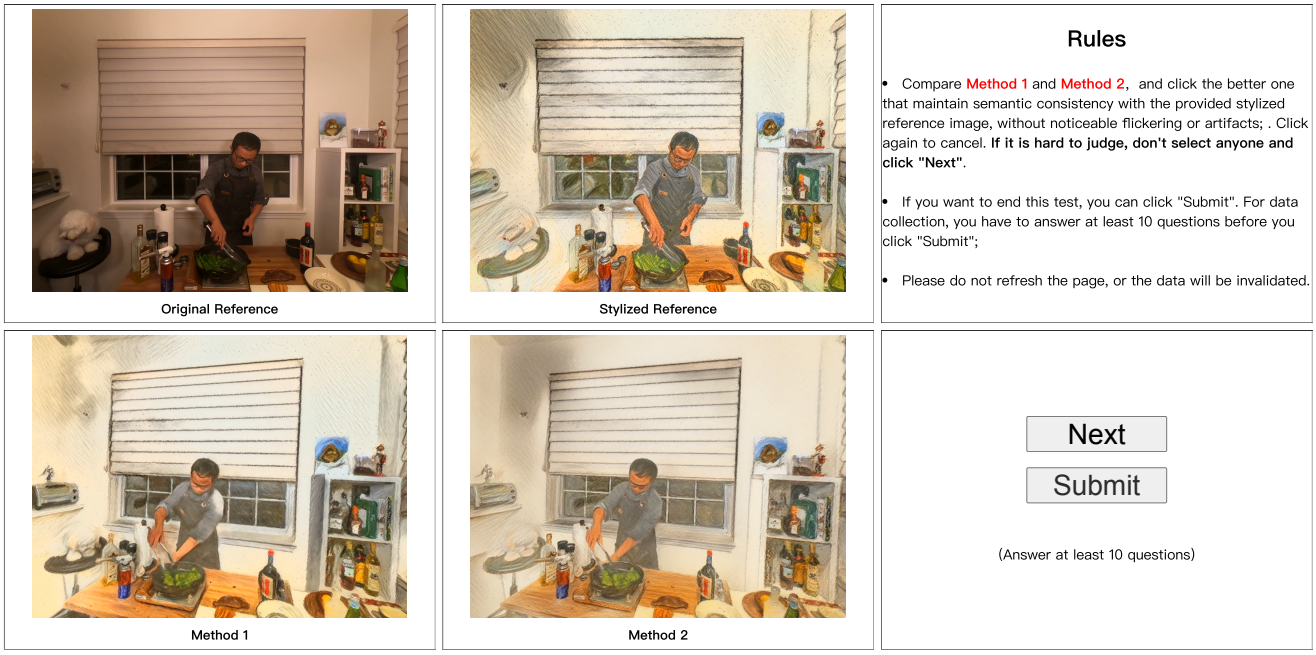

Figure D.2. **Interface of the user study website.**

[10] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 717–733. Springer, 2022. 1

[11] Yuechen Zhang, Zexin He, Jinbo Xing, Xufeng Yao, and Jiaya Jia. Ref-npr: Reference-based non-photorealistic radiance fields for controllable scene stylization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4242–4251, 2023. 1

Figure E.3. **Additional space-time view synthesis results from the Plenoptic Video dataset.**

Figure E.4. **Additional space-time view synthesis results from the Plenoptic Video dataset.**

Figure E.5. **Additional space-time view synthesis results from the Plenoptic Video dataset.**

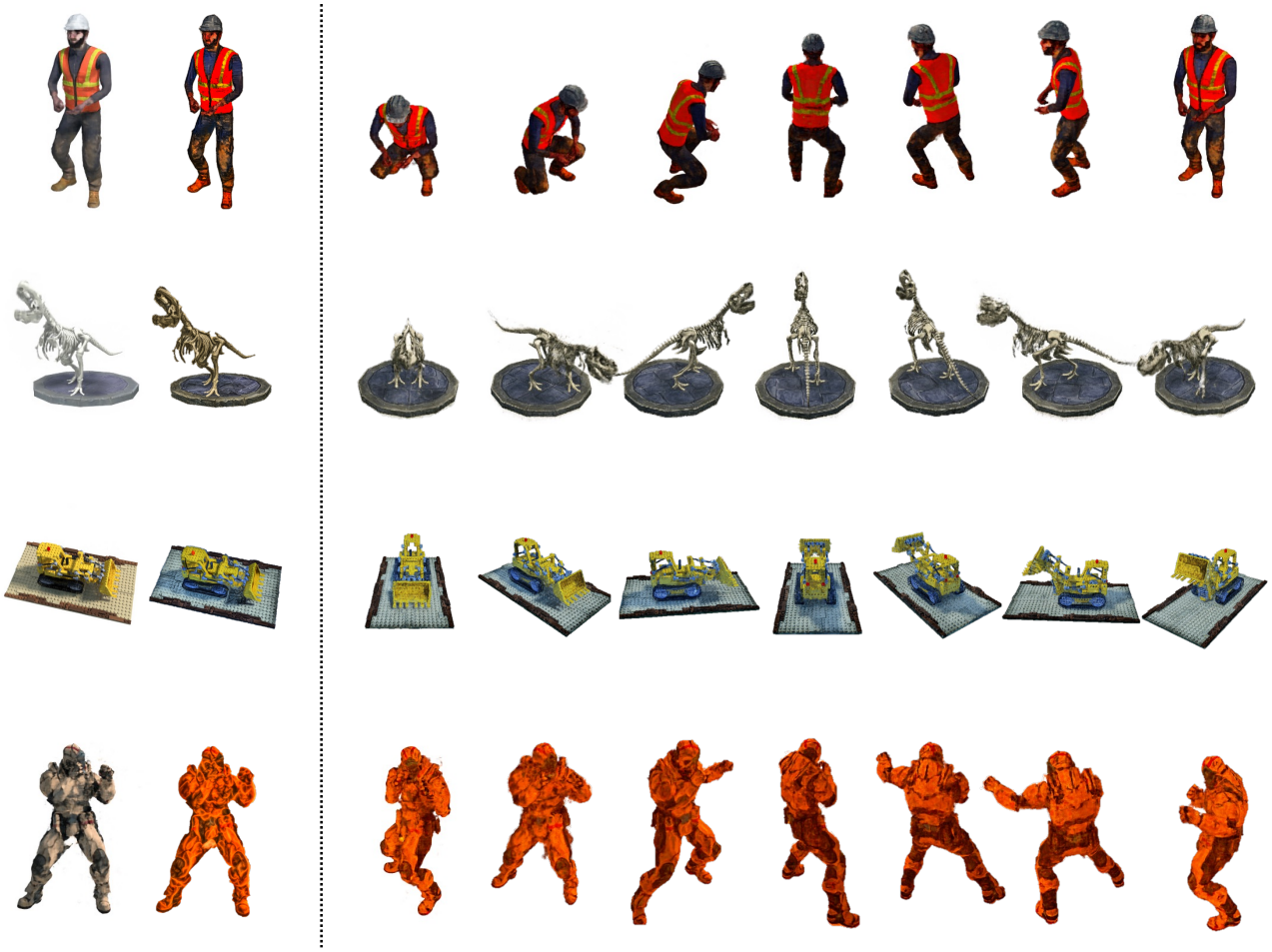

Figure E.6. **Additional space-time view synthesis results from the D-NeRF dataset.**

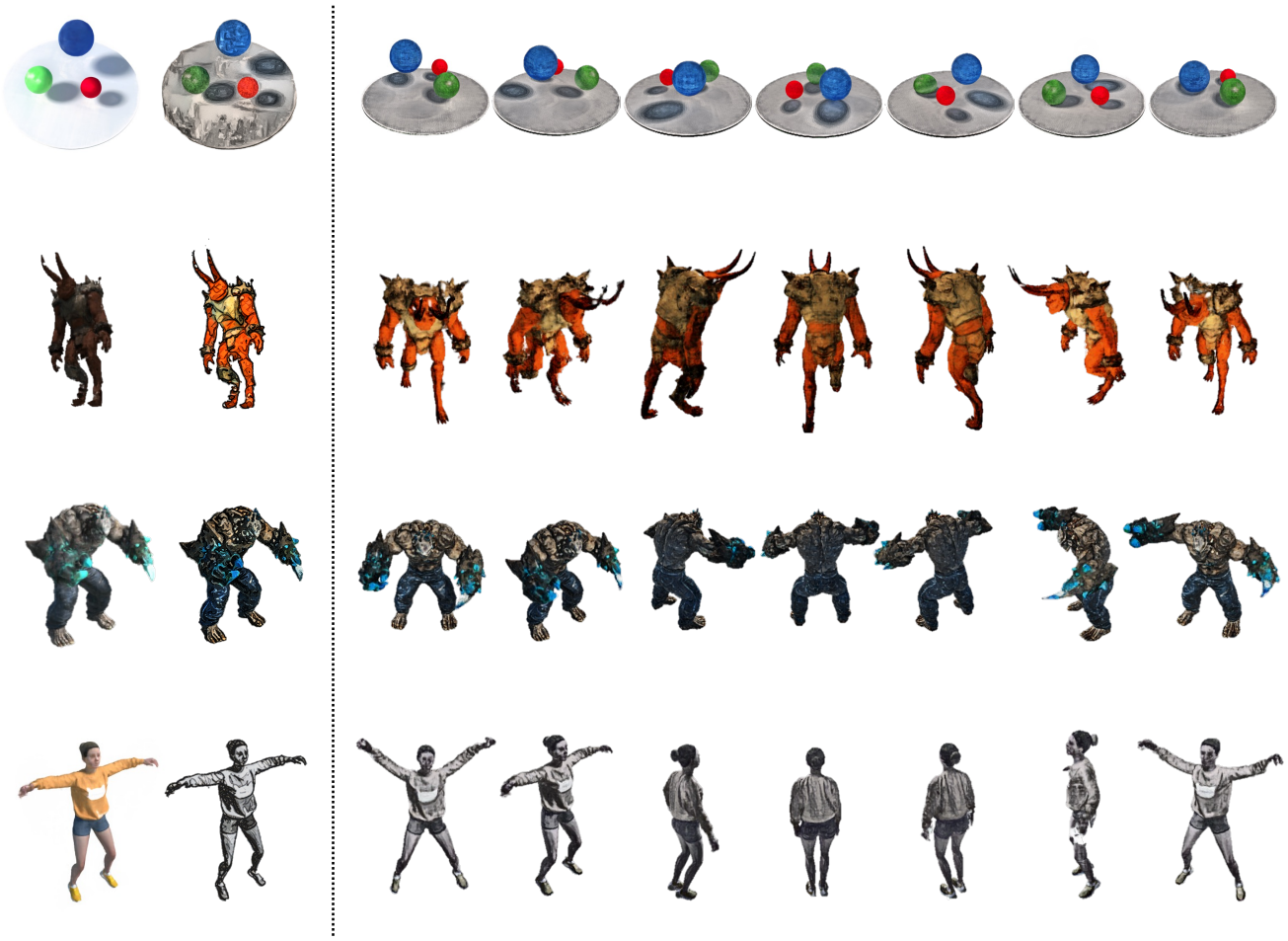

Figure E.7. **Additional space-time view synthesis results from the D-NeRF dataset.**