S2RF: Semantically Stylized Radiance Fields

Moneish Kumar* Neeraj Panse* Dishani Lahiri*
Robotics Institute, Carnegie Mellon University
{moneishk, npanse, dishanil}@andrew.cmu.edu

Abstract

We present our method for transferring style from any arbitrary image(s) to object(s) within a 3D scene. Our primary objective is to offer more control in 3D scene stylization, facilitating the creation of customizable and stylized scene images from arbitrary viewpoints. To achieve this, we propose a novel approach that incorporates nearest neighborhood-based loss, allowing for flexible 3D scene reconstruction while effectively capturing intricate style details and ensuring multi-view consistency.

1. Introduction

For decades, recovering three-dimensional (3D) information from two-dimensional (2D) images has posed a persistent challenge in the field of computer vision. With the advent of cutting-edge differential rendering methods [17, 24], exciting new modalities have emerged, enabling the reconstruction of high-fidelity [21, 11, 7] and efficient [4, 9, 7] 3D scenes.

As advancements make 3D reconstructions more accessible, there is a growing demand for editing and manipulating these scenes. The ability to edit 3D scenes empowers creators to push the boundaries of imagination and precision. One such editing application is 3D style transfer, which aims to transfer artistic features from a single 2D image to a real-world 3D scene. Numerous remarkable works [6, 8, 10, 12, 20, 25] have successfully achieved this objective. However, these methods primarily concentrate on stylizing the **whole** scene by utilizing only a **single** style image.

The primary goal of this paper is to enhance the level of control while stylizing 3D scenes. With our method, highly customizable stylized scene images can be generated from arbitrary novel viewpoints. It not only facilitates the stylization of individual object(s) but also ensures that rendered images maintain spatial consistency. Figure 1 provides a summary of our stylization approach. To the best of our knowledge, this is the first approach that offers a single framework for semantic and instance-level style transfer for objects in a 3D scene.

Similar to previous methods [25, 12, 2, 26, 8, 6], addressing style transfer in 3D, we adopted an optimization-based approach. These methods aim to minimize (i) content loss, evaluating the difference between rendered stylized images and the original captured images, and (ii) style loss, quantifying the variance between rendered images and the style image. However, unlike these methods, we uniquely apply content and style loss exclusively to the relevant objects in the image, granting us superior control over the generated 3D scene. This approach allows for precise and targeted stylization, empowering us to achieve more tailored and refined results.

We present our results across a diverse range of 3D scenarios, showcasing how our approach serves as a stepping stone toward achieving more controllable 3D scene generation.

2. Related work

There is a plethora of work in scene style transfer for NeRFs [25, 12, 10, 19, 23, 15, 8]. In the majority of style transfer pipelines, a two-staged training framework is employed. The first stage entails training a photo-realistic 3D scene, while the second stage involves fine-tuning or modifying the 3D scene representation using the style image. Some of methods represent the 3D scene in the form of meshes [8, 15, 23], some in the form of point-clouds [10, 19] while other using an implicit radiance field [25, 12, 18].

These methods also differ in the way they fine-tune or modify the 3D scene representation. Some works utilize a separate hyper network [6, 12] while others alter the implicit representations themselves [25]. In the realm of 3D scene stylization, addressing spatial consistency emerges as one of the main challenges to be resolved. For example, StyleMesh [8] adopts a joint stylization approach, utilizing all input images to stylize the 3D scene and optimizing an explicit texture for accurate scene reconstruction. While Aristic Radiance field [25] employs an exclu-

*Equal contribution

Figure 1. Introducing S2RF, a technique for achieving stylizable 3D reconstruction. Our method processes a set of images capturing a 3D scene and offers users the ability to style specific objects within that scene. By incorporating user-defined styles for these objects, our approach generates a stylized 3D reconstruction.

sive color transfer to ensure view consistency.

Controlling the style transfer and restricting it to user-specified objects is a challenging task and an active area of research. The current state-of-the-art methods perform instance-based style transfer but the quality of the renderings are not as visually pleasing and contain artifacts. An interesting work Sine [2], requires one image from a scene edited by the user and can generate a 3D view of the scene with the edited objects. In this case, geometric priors are also used that constrain and maintain the geometric components of the objects in the scene. Albeit flexible, this method requires the user to edit one image and also the edits are semantically constrained, unlike our method where any style can latch on faithfully to an object of choice.

3. Method

The overview of our method is shown in Figure 2. Given a set of calibrated images of a scene and a set of user-defined objects in the scene, we aim to create a realistic and geometrically consistent image from any arbitrary viewpoint in the scene in which only the user-defined objects are styled according to reference style images. Our framework consists of three phases: generation of radiance fields, detection of objects, and stylization of radiance fields.

3.1. Generation of radiance field

Our method uses radiance fields (RF) to represent the scene in 3D. Given a set of calibrated images of a scene, this radiance field is optimized using a rendering loss on the training rays. The method is agnostic to the way radiance fields are represented but for efficient, we use Plenoxel's [24] sparse voxel grid ($\mathcal{V}$) to represent the 3D scene. Each occupied voxel stores a scalar opacity σ and a vector of spherical harmonic coefficients for each color channel. The radiance field is defined using trilinear interpolation over sparse voxel grid.

$$L(x, w) = \phi(x, \mathcal{V}) \tag{1}$$

Where x is the queried point in the 3D space, w is the queried unit directional vector, $\mathcal{V}$ is the voxel grid and the function ϕ is trilinear interpolation.

It uses the differentiable volume rendering model used in NeRF [17]. The color of the ray is determined by integrating all points along the ray.

$$\hat{C}(r) = \sum_{i=1}^{N} T_i (1 - e^{-\sigma_i \delta_i}) c_i \tag{2}$$

$$T_i = e^{-\sum_{j=1}^{i-1} \sigma_j \delta_j} \tag{3}$$

Where T_i represents the amount of light transmitted along the ray r, δ_i is the opacity of sample i, c_i is the color of sample i.

Voxel grid's opacity and spherical harmonic coefficients are optimized using mean square error ($\mathcal{L}_{mse}$) over the rendered pixels along with total variation ($\mathcal{L}_{tv}$) [22], beta distribution ($\mathcal{L}_{\beta}$) regularizers and sparsity prior ($\mathcal{L}_{s}$)[13]. The overall loss function ($\mathcal{L}_{rf}$) for the radiance field optimization is as follows:

$$\mathcal{L}_{rf} = \mathcal{L}_{mse} + \lambda_{tv}\mathcal{L}_{tv} + \lambda_{\beta}\mathcal{L}_{\beta} + \lambda_{S}\mathcal{L}_{S} \tag{4}$$

$$\mathcal{L}_{mse} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} ||C(r) - \hat{C}(r)||_2^2 \tag{5}$$

$$\mathcal{L}_{tv} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \sum_{d \in \mathcal{D}} ||\Delta(v, d)||_2 \tag{6}$$

$$\mathcal{L}_{S} = \sum_{i} \sum_{k} log(1 + 2\sigma(r_i(t_k))^2) \tag{7}$$

$$\mathcal{L}_{\beta} = \sum_{r} (log(T_{FG}(r)) + log(1 - T_{FG}(r))) \tag{8}$$

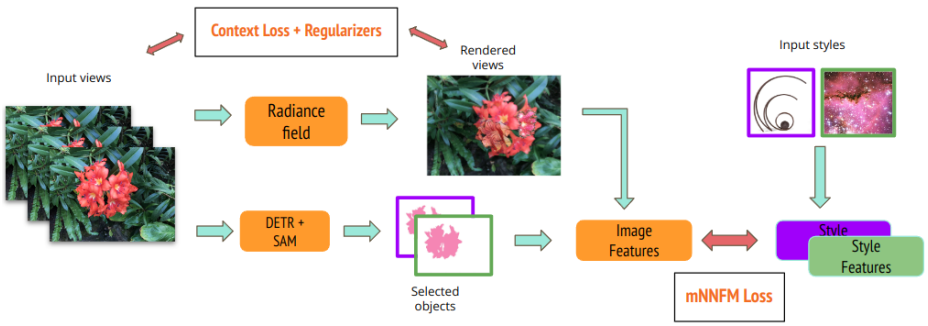

Figure 2. **Overview of our method**. We begin by reconstructing a photo-realistic radiance field and segmenting all objects from a set of scene images. Next, we apply stylization to this reconstruction by employing a masked Nearest Neighbor Feature Matching (mNNFM) style loss with the help of exemplar style images. Once the stylization process is complete, we can generate consistent free-viewpoint stylized renderings. For a more in-depth understanding of our results, we invite readers to watch the supplemental videos.

Where $C(r)$ is the color of the ground truth ray, $\hat{C}(r)$ is the estimated color of the ray, $||\Delta(v,d)||_2$ is the squared distance between the dth values in the voxels. $\sigma(r_i(t_k))^2)$ is the opacity of the sample k along the ray i. $log(T_{FG}(r))$ denotes the accumulated foreground transmittance of ray r. $\lambda_{tv}, \lambda_\beta$ and λ_S are weights of the respective loss components.

3.2. Detection of objects

The second phase of our framework aims to provide a selection of objects in the scene to the user to which style can be transferred. Given a set of images of a scene, the output is a set of all object and mask pairs $\mathcal{O} = \{(o_i, m_i)\}$ (where $i \in [0, N]$ and N is the number of objects) in the scene.

We use a transformer-based object detector, DEtection TRansformer (DETR) [3] to detect objects. Segmentation masks are obtained using the Segment Anything Model (SAM) [1] for all images in the scene. Given an input image, DETR produces a comprehensive list of object boxes, each associated with a category tag and corresponding bounding box coordinates. SAM takes as input an image along with object bounding boxes and outputs segmentation masks corresponding to each of the input object queries. The segmentation masks generated using box prompts are much better than those generated using other prompts [5, 14] , hence we use DETR prior to SAM.

At this point, we have a list of objects (with segmentation masks) in the scene and the corresponding style images that need to be transferred. To ensure the reliability of the detected objects, we only retain those that appear in at least 80% of the frames throughout the scene images.

3.3. Stylization of radiance fields

Given a photo-realistic radiance field that is reconstructed using the method in section 3.1 and a set of objects and masks that are obtained utilizing the approach in section 3.2, our framework finetunes the photo-realistic radiance field, in which the objects are stylized according to their respective 2D style image. We achieve this by applying the Nearest Neighbor Feature Matching (NNFM) loss [25] to each object individually.

The NNFM loss aims to minimize the cosine distance of each feature in the feature map of the rendered image to its nearest neighbor feature in the style images' feature map. The rendered image from the radiance field is denoted by I_r and the style image is denoted by I_s. The VGG feature maps extracted from both these images are F_r and F_s respectively. The NNFM loss is given by:

$$\mathcal{L}_{NNFM} = \frac{1}{N} \sum_{i,j} \min_{k,l} \delta \left(F_r(i,j), F_s(k,l) \right) \tag{9}$$

where $F_*(i,j)$ denotes the feature vector at pixel location (i,j) for the feature map F_* and the function $\delta(v_1, v_2)$ computes the cosine distance between vectors v_1 and v_2.

We exclusively apply the NNFM loss (equation 9) to the pixels that correspond to each object separately. This selective application is achieved by employing the mask obtained in section 3.2. The mask allows to effectively confine the style transfer to the specific objects of interest. The masked NNFM loss (mNNFM) is as follows:

$$\rho = \sum_{i,j} \min_{k,l} m_o(i,j) D(F_r(i,j), F_s^o(k,l)) \tag{10}$$

$$\mathcal{L}_{mNNFM} = \frac{1}{N} \sum_{o=1}^{N} (\rho) \tag{11}$$

m_o represents the mask specific to object o, while F_s^o denotes the feature map extracted from the style image intended for transfer onto object o. ρ represents the masked-NNFM loss over a single object and the total loss is the average over all objects in the scene.

Figure 3. Shows examples of stylized radiance fields in two senarios. 1) Single object (chair) instance is stylized (Top row). 2) Multiple instances of the same object (flower) have been stylized (Bottom row). Images on the left show one of the input images for the scene along the object to be styled and style image (top-left). Images on the right show an image of the stylized image.

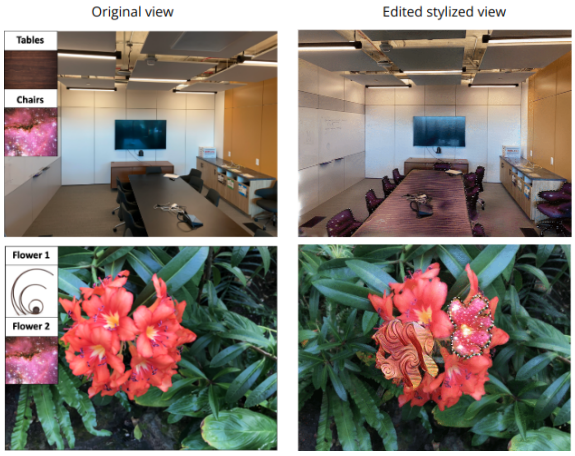

Figure 4. Shows examples of style transfer with 1) Multiple instances of multiple object(s) (Chairs and Table) are stylized (Top row). 2) Multiple instances of the same object (flower) are stylized (Bottom row). Images on the left show one of the input images for the scene along the object to be styled and style image (top-left). Images on the right show an image of the stylized image.

Combining the masked NNFM loss with the loss mentioned in section 3.1, the overall loss that we optimize is:

$$\mathcal{L} = \mathcal{L}_{rf} + \mathcal{L}_{mNNFM} \tag{12}$$

The modified NNFM loss plays a crucial role in refining the radiance field generation process, ensuring that the applied style adheres precisely to the selected objects. This approach enhances the overall visual appeal and contextual consistency of the final output, making it more compelling and realistic.

4. Experiments

To assess the effectiveness of our method, we conduct qualitative evaluations, showcasing results from diverse real scenes where the objects are influenced by different style images. We demonstrate how our approach successfully applies various styles to objects within real-world contexts, providing visual evidence of its versatility and performance.

Datasets. We conduct our experiments on multiple real-world scenes which include: *Flower, Xmaschair, Room* from [16]. All these scenes are front-facing captures. The style images include a diverse set of images taken from [25].

Our qualitative results are presented in Fig. 3,4. We explore four different scenarios for style transfer:

Style transfer on a single instance of an object: In this scenario, we apply a style image to a single object within the 3D scene. Figure 3 (Top row) shows the result of applying the style image on the chair.

Style transfer on all instances of a single object: In this case, we transfer a single style to all instances of a single object. Figure 3 (Bottom row) shows the application of the given style on all the flower instances in the scene.

Style transfer on all instances of multiple objects object: In this example, we transfer different styles to separate objects in the scene. Figure 4 (Top Row) shows the application of different styles on the table and chairs in the 3D room scene.

Style transfer on multiple instances of a single object: In this case, we transfer different styles to separate instances of the same object. We apply different two different styles on two separate flowers in the same scene as shown in Figure 4 (Bottom row).

We encourage readers to watch the supplementary videos and the appendix A to better view the results.

5. Discussion

We propose a novel method for reconstructing stylized radiance fields from photorealistic radiance fields. The cornerstone of our method lies in the application of masked NNFM loss, enabling a more controllable style transfer. Our method effectively achieves style transfer on both semantic and instance level, successfully applying distinct style(s) to multiple object(s) within a single scene. While this serves as a compelling proof of concept, a more comprehensive evaluation is required to fully validate our approach. Future assessments should encompass a broader range of scenes, including 360-degree environments and scenes with an increased number of objects. Additionally, it is crucial to conduct a quantitative evaluation to thoroughly assess the effectiveness of our method.

References

[1] Segment anything. 2023.

[2] Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Sine: Semantic-driven image-based nerf editing with prior-guided editing field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20919–20929, 2023.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[4] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16569–16578, 2023.

[5] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*, 2023.

[6] Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, and Wei-Chen Chiu. Stylizing 3d scene via implicit representation and hypernetwork. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1475–1484, 2022.

[7] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. *arXiv preprint arXiv:2103.10380*, 2021.

[8] Lukas Höllein, Justin Johnson, and Matthias Nießner. Stylemesh: Style transfer for indoor 3d scene reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6198–6208, 2022.

[9] Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, and Jiaya Jia. Efficientnerf efficient neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12902–12911, 2022.

[10] Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, and Ming-Hsuan Yang. Learning to stylize novel views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13869–13878, 2021.

[11] Xudong Huang, Wei Li, Jie Hu, Hanting Chen, and Yunhe Wang. Refsr-nerf: Towards high fidelity and super resolution view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8244–8253, 2023.

[12] Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18342–18352, 2022.

[13] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019.

[14] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, page 102918, 2023.

[15] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022.

[16] 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, 2019.

[17] 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 *ECCV*, 2020.

[18] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20669–20679, 2023.

[19] Fangzhou Mu, Jian Wang, Yicheng Wu, and Yin Li. 3d photo stylization: Learning to generate stylized novel views from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16273–16282, 2022.

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

[21] Zhongshu Wang, Lingzhi Li, Zhen Shen, Li Shen, and Liefeng Bo. 4k-nerf: High fidelity neural radiance fields at ultra high resolutions, 2023.

[22] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020.

[23] Kangxue Yin, Jun Gao, Maria Shugrina, Sameh Khamis, and Sanja Fidler. 3dstylenet: Creating 3d shapes with geometric and texture style variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12456–12465, 2021.

[24] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021.

[25] 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.

[26] 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, pages 4242–4251, 2023.

A. Qualitative Results

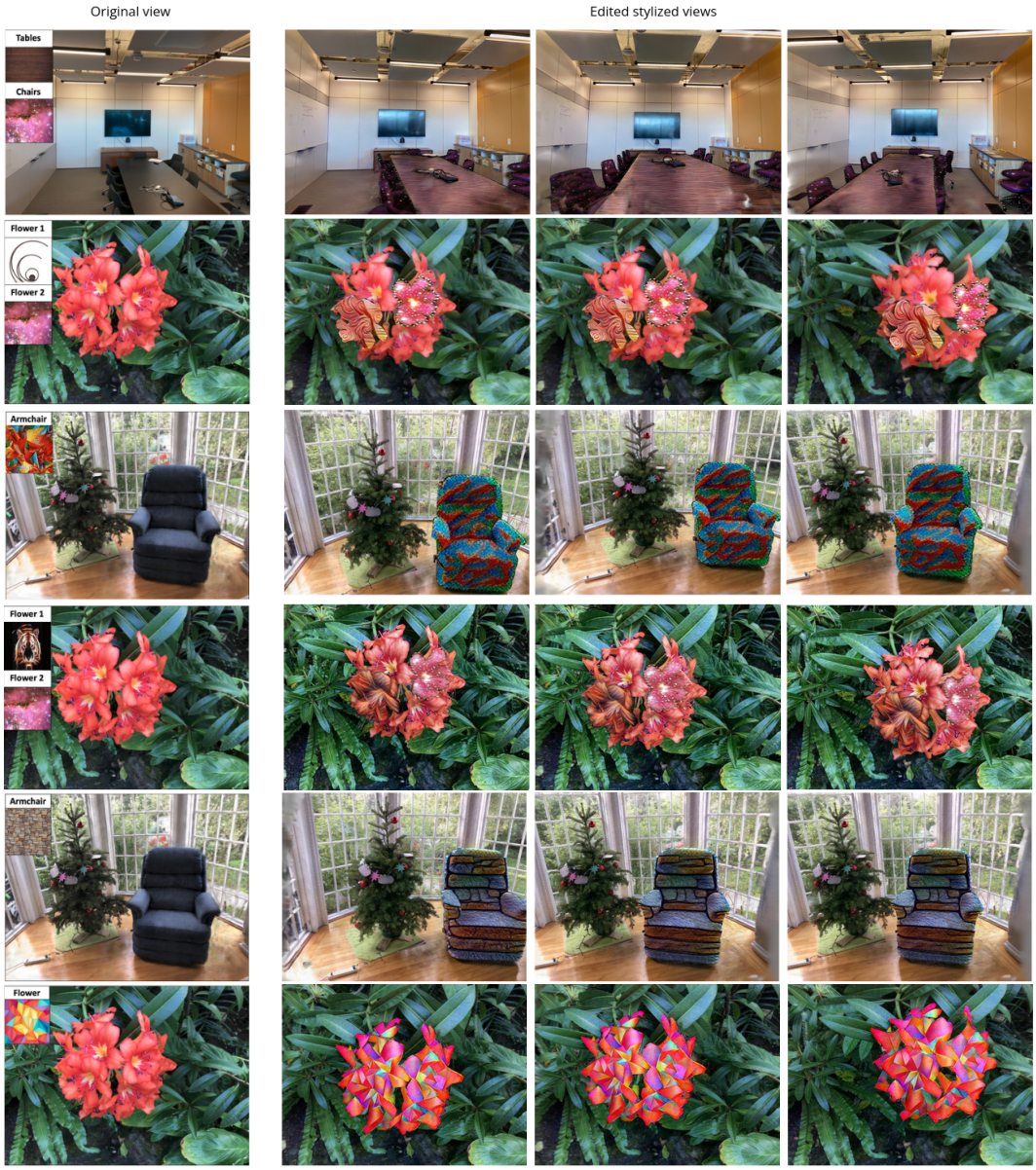

Figure 5. Shows qualitative results with examples of style transfers with multiple objects and style. Images in the first column show one of the input images of the scene along the object to be styled and the style image (top-left). Images on the right show three images from the stylized scene.