# CoARF: Controllable 3D Artistic Style Transfer for Radiance Fields

Deheng Zhang[1]     Clara Fernandez-Labrador[2]     Christopher Schroers[2]

[1]ETH Zürich     [2]DisneyResearch|Studios

## Abstract

*Creating artistic 3D scenes can be time-consuming and requires specialized knowledge. To address this, recent works such as ARF [57], use a radiance field-based approach with style constraints to generate 3D scenes that resemble a style image provided by the user. However, these methods lack fine-grained control over the resulting scenes. In this paper, we introduce Controllable Artistic Radiance Fields (CoARF), a novel algorithm for controllable 3D scene stylization. CoARF enables style transfer for specified objects, compositional 3D style transfer and semantic-aware style transfer. We achieve controllability using segmentation masks with different label-dependent loss functions. We also propose a semantic-aware nearest neighbor matching algorithm to improve the style transfer quality. Our extensive experiments demonstrate that CoARF provides user-specified controllability of style transfer and superior style transfer quality with more precise feature matching.*

## 1. Introduction

As the film and game industry continue to evolve, there is an increasing demand for more accessible methods to create and manipulate 3D scenes. Radiance field-based methods, such as Neural Radiance Fields (NeRF) [35], gained widespread attention due to their ability to produce high-quality novel-view images without the requirement of 3D assets like meshes or point clouds. Instead, they use implicit representation to encode the 3D scene information. However, despite the impressive photorealism achieved by these methods, the style richness of the resulting 3D scenes is still limited. To address this, researchers have turned their attention to style transfer for 3D content, which has gained significant interest as a means of providing inspiration and tools to 3D artists. There are many different style transfer approaches based on NeRF. For instance, [5] use a separate HyperNetwork to control the input features based on arbitrary style images, while [14, 28, 31] integrate AdaIN [13] module to the NeRF pipeline to achieve style

transfer. These methods cannot be easily integrated with NeRF variants such as [8, 38] since the structure is tailored for style transfer. As a comparison, the Artistic Radiance Fields (ARF) [57] applies a nearest neighbor feature matching (NNFM) style constraint to fine-tune the radiance information, thus can be generalized to NeRF variants and achieves superior stylization quality compared with the previous methods. However, artists need to control the style transfer target in practical applications, and ARF lacks fine-grained controllability in the stylization process. Other recent works [31, 52] provide controllability for compositional style transfer and text-based style transfer, but the resulting quality is inferior and the semantic correspondence is not well captured during style transfer for these methods.

In this paper, we propose Controllable Artistic Radiance Fields (CoARF), which endows a baseline model [57] with three controllability modules for 3D style transfer as shown in Figure 1. Namely, CoARF enables object selection, compositional style transfer and semantic-aware style transfer using multi-view 2D mask-based optimization and masks defined by a user. Similar to the 2D case [46], a naïve solution to achieve controllability is to define different optimization loss functions for different 3D regions. However, it is non-trivial to obtain the precise 3D object in practice. Therefore, we use multi-view 2D mask-based optimization with different loss definitions to achieve different controllable style transfer tasks and prove that multi-view optimization can automatically correct the error gradient.

In more detail, the object selection module allows the user to select which parts of the 3D scene should be stylized while keeping the rest photorealistic. To do so, we optimize the user's selected area using NNFM loss, and use Mean Squared Error (MSE) to preserve the rest photorealistic. The compositional style transfer module allows transferring of different styles to different parts of the 3D scene by using specific NNFM loss for corresponding mask labels. Last, our semantic-aware style transfer module uses 2D mask labels to match semantic regions between the 3D scene and the provided style image. We also show that while VGG features successfully encode texture, structure and color, they fail to encode semantic information. There-
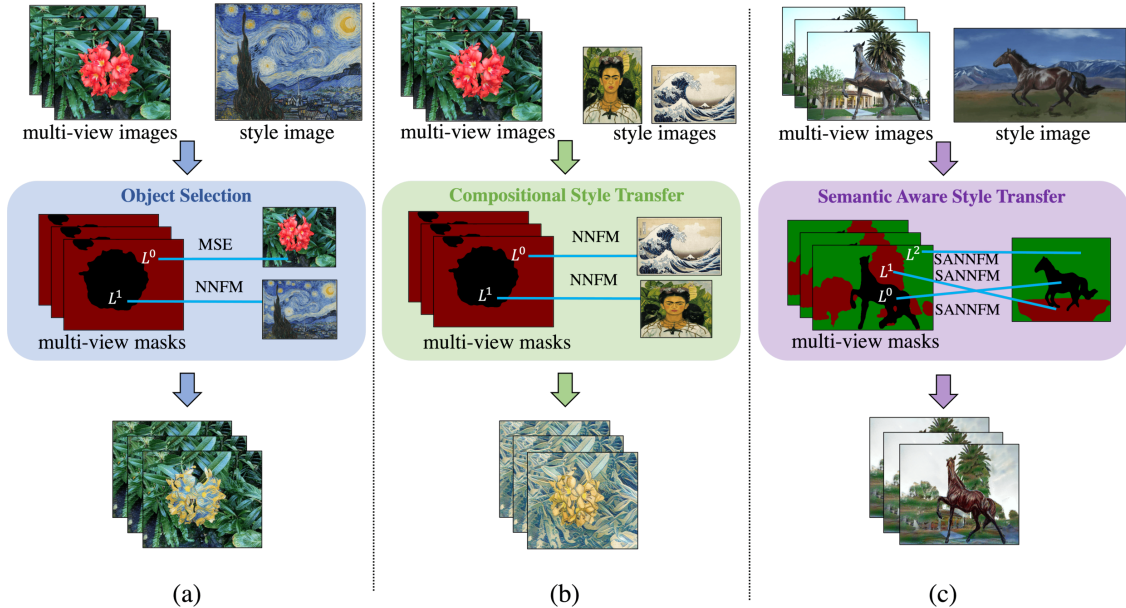
Figure 1. **CoARF Overview.** Given a set of multi-view ground truth images and style images, our controllable style transfer model allows the user to perform object selection (a), compositional style transfer (b), and semantic-aware style transfer (c) by using 2D mask-based optimization with different spatial-dependent loss definitions to achieve stylization.

fore, we proposed a novel Semantic-aware Nearest Neighbor Feature Matching (SANNFM) algorithm which leverages a weighted sum of VGG feature distance and LSeg feature distance for stylization to improve general style transfer quality.

We summarize our contributions as follows: 1) We propose a multi-view 2D mask-based optimization framework to solve the general 3D controllable style transfer problem. 2) By leveraging the core framework, we can achieve object selection, compositional style transfer, and semantic-aware style transfer. 3) Our semantic-aware 3D style transfer algorithm utilizes a semantic-based nearest neighbor matching technique, which achieves better style transfer quality. Through extensive experimentation, we show that CoARF provides fine-grained control over the style transfer process and yields superior results compared to state-of-the-art algorithms [31, 57].

## 2. Related Work

**2D Style Transfer.** 2D style transfer can be achieved through various methods such as non-parametric algorithms [24] or texture synthesis [43]. However, since [11] introduced the use of a convolutional neural network, there has been significant improvement in performance. This approach utilizes features extracted from the VGG-19 [47] network to calculate a Gram matrix-based style constraint, and optimizes the output from a noise image. Follow-up methods [22, 27, 29, 30, 46] use different content and style

constraints to iteratively optimize the noise image. Recently, nearest neighbor-based algorithms [23, 27, 30] have been proposed to match content and style features and minimize the cosine distance between them. ARF [57] applies a similar nearest neighbor-based loss as [23], despite the loss function being only applied to the final scale instead of augmenting the style images. In our work, we improve upon ARF by modifying the nearest neighbor-finding algorithm using semantic information, which results in better style transfer outcomes. There are also several efforts [17, 19] to explore the controllability of 2D style transfer from the level of stylization and stroke perspectives. Previous works [12, 15, 46, 58] introduced user-defined masks or automatically generated soft masks for semantic-aware style transfer. Our work is similar to [46], but we propose a novel semantic-aware method to improve feature matching.

**Radiance Fields.** NeRF [35] proposed a radiance field-based representation of the 3D scene. The method is trained using multi-view 2D images for which the camera pose is provided and uses MLPs to predict the density and radiance of a given sample on the ray. Then the volumetric rendering equation is applied to render the image for different views. To further enhance the quality and diversity of rendered scenes, several tasks have been explored, including NeRF rendering performance improvements [2, 48, 54], NeRF speedup [3, 8, 38], dynamic scene rendering with radiance fields [39–41, 44], 3D scene editing and control [18, 21, 25, 51], view consistent segmentation

[7, 33, 50, 59], and novel view synthesis for real-world scenarios [32, 36, 53]. In this work, we use Plenoxels [8] as our backbone model for radiance prediction, which propose a sparse voxel model that matches NeRF's rendering performance while being much faster. Our algorithm can also be generalized to other differentiable radiance-field-based algorithms as demonstrated in ARF [57].

**NeRF-based Style Transfer.** There are various ways to enable stylized radiance prediction, including adding a tailored MLP [4, 5, 14] for stylization or applying AdaIN transformation to the input features of the radiance MLP [28]. However, these methods suffer from unstable results during training due to the use of a tailored stylization module. Moreover, they cannot be easily generalized to NeRF variants [8, 38], since the original scene encoded in NeRF is not edited. In contrast, ARF [57] applies an NNFM loss to the rendered image of the pre-trained radiance fields and fine-tunes the radiance information, resulting in exceptional stylization results compared to other methods. However, ARF lacks fine-grained controllability in the stylization process. An alternative approach, StyleRF [31], uses volumetric rendering to integrate features in the scene and deferred style transformation to enable controllability for compositional style transfer. Nevertheless, the stylization quality is worse than ARF, because volumetric rendering for neural features is not physically-based. Our algorithm introduces fine-grained controllability to ARF using a 2D mask-based optimization and achieves better style transfer quality compared with existing methods.

**Controllable NeRF.** The controllability of NeRF has been explored in various directions. For instance, [25] allows compositing objects from different scenes. [18] enables dynamic facial expression control using 2D annotations, while [10, 16, 42, 55, 56] edit the query for NeRF trained in static scenes, achieving user-defined deformation on the given scene. CLIPNeRF and SINE [1, 51] enable text-driven editing, whereas [21] distills the 2D semantic feature from LSeg [26] to train 3D semantic feature using volumetric rendering, enabling editing including colorization, translation, deletion, and text-driven editing. In our project, we use LSeg [26] to provide a 2D mask for optimization and calculate semantic features to assist semantic-aware style transfer. However, we use 2D masks directly for controllability, unlike [21], which calculates explicit 3D label distribution for controllability.

**Language-driven Recognition.** The field of language-driven recognition models has gained a lot of attention after the release of CLIP [45]. CLIP utilizes text-image encoders to encode image and text features into the same space via contrastive pre-training, which enables zero-shot class prediction. LSeg [26] builds on this work by using the pre-trained text encoder from CLIP and an image encoder based on ViT [6] to achieve pixel-wise semantic feature prediction. Other works [9, 34, 37, 49] have shown that the loss function defined by CLIP features has significant controllability for semantic information. In our work, we utilize a combination of semantic features from LSeg and features from VGG to guide nearest-neighbor searching and improve the robustness of style transfer.

## 3. Background

The general representation of the radiance fields is a mapping $f : \mathbb{R}^5 \mapsto \mathbb{R}^3$, which takes a 3D position $\mathbf{x}$ and direction $\mathbf{d}$ as input and outputs density $\sigma$ and radiance $\mathbf{c}$:

$$\sigma, \mathbf{c} = \text{RADIANCEFIELD}(\mathbf{x}, \mathbf{d}). \quad (1)$$

The radiance for a given ray can be calculated as the following weighted average by uniform sampling points $\mathbf{x}_i$ along the ray:

$$\mathbf{c}(\mathbf{r}) = \sum_{i=1}^{N} w_i \mathbf{c}_i, \quad w_i = T_i(1 - exp(-\sigma_i \delta_i)), \quad (2)$$

where $T_i = exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$. The radiance fields function is optimized by the $L_2$ distance between the ground truth and the rendered image. In particular, we use Plenoxels [8] as the radiance fields function, which represents the radiance as spherical harmonics functions in a voxel grid. Our CoARF is generalizable to the radiance fields representations.

In ARF [57], the radiance field is firstly pre-trained using the multi-view ground truth (content) images. Then, the density field is fixed, and ARF only optimizes the radiance prediction part of the radiance field. ARF fine-tunes the photorealistic radiance field by defining an averaged pixel-wise loss function on the rendered image:

$$L = \frac{1}{N} \sum_{x,y} (l_{\text{nnfm}}(\mathbf{F}_r(x,y), \mathbf{F}_s) + \lambda \cdot l_2(\mathbf{F}_r(x,y), \mathbf{F}_c(x,y)))$$
$$+ \lambda_{\text{tv}} \cdot l_{\text{tv}}, \quad (3)$$

where $N$ is the number of pixels, $\mathbf{F}_c$, $\mathbf{F}_s$, and $\mathbf{F}_r$ are features extracted from the multi-view content (ground truth), style, and rendered images using VGG [47] encoder. $\mathbf{F}(x,y)$ denotes the feature vector at pixel location $(x,y)$. $l_2$ represents pixel-wise MSE and $l_{tv}$ represents total variation loss. The pixel-wise NNFM style loss is defined as the nearest cosine distance in the style image:

$$l_{\text{nnfm}}(\mathbf{F}_r(x,y), \mathbf{F}_s) = \min_{x',y'} D(\mathbf{F}_r(x,y), \mathbf{F}_s(x',y'))$$
$$D(\mathbf{v}_1, \mathbf{v}_2) = 1 - \frac{\mathbf{v}_1^T \mathbf{v}_2}{\sqrt{\mathbf{v}_1^T \mathbf{v}_1 \mathbf{v}_2^T \mathbf{v}_2}}. \quad (4)$$

ARF also applies color transfer to the voxel grid when they use Plenoxels as the backbone and the multi-view ground truth image before and after the stylization, which aligns the distribution of the RGB values to the style image.

## 4. Methodology

We propose a novel controllable and generalizable style transfer model for radiance fields, which not only enables object selection and compositional style transfer but also produces a better style transfer quality for semantic sensitive style images. We first introduce the core 2D mask-based algorithm to provide controllability for the radiance fields in Section 4.1. After that, we demonstrate how to apply this algorithm with different optimization constraints to achieve different controllable tasks in Section 4.2. Last, we provide implementation details of the controllable style transfer algorithms in Section 4.3.

### 4.1. 2D Mask-Based Optimization

Given a rendered image from one camera view $\mathbf{I}_r$, we can apply any 2D segmentation algorithm such as LSeg [26] to produce a segmentation mask $M_r$ with $M$ classes, such that each pixel $M_r(x, y) \in \{0, 1, ..., M\}$. We define the controllable loss function as:

$$L = (\frac{1}{N} \sum_{x,y} \sum_{m} \mathbb{1}[M_r(x, y) = m] L^m(x, y)) + \lambda_{\text{tv}} \cdot l_{\text{tv}},$$
(5)

where $L^m$ is a pixel-wise loss function defined for corresponding label $m$, and $\mathbb{1}[\text{condition}]$ is an indicator function with output one if the condition holds and zero if the condition does not hold. In particular, we assign a specific loss function for each mask label. As shown in Figure 3(a), this loss function is problematic for one-view optimization because 2D mask-based loss optimizes a 3D column, instead of accurately optimizing the 3D object. As a result, point A would be optimized using the incorrect foreground loss function. However, as shown in Figure 3(b), with multi-view optimization, this error can be automatically corrected. According to the volume rendering equation, the final gradient for the radiance of point A $\nabla_{\mathbf{c}_A} L$ is averaged among the multiple views:

$$\nabla_{\mathbf{c}_A} L = \sum_v w_A^v \nabla L^{m_v},$$
(6)

where $w_A^v$ represents the contribution of point A during volume rendering in Equation 2 for view $v$, $m_v$ is the point A masking label from view $v$, and $\nabla L^{m_v}$ represents the gradient of the corresponding loss w.r.t the rendered pixel value. During optimization, A has the correct background label when visible and the wrong foreground label when occluded. According to the physical property of the volume rendering equation, since point A is visible from views 1 and 2, the weight becomes larger dominating the final gradient and optimizing the radiance with the correct background loss. The detailed proof can be found in the supplementary materials. Based on this optimization framework, we pro-

pose to combine different loss functions $L^m(x, y)$ based on the type of control we want to achieve.

### 4.2. Controllable Optimization for Different Tasks

**Object Selection.** For object selection in 3D style transfer, we provide a 0-1 mask for each camera view to label objects that should be stylized (1) and objects that should be kept photorealistic (0). The loss function is the same as ARF for pixels with label 1, and we use MSE for pixels with label 0:

$$L^m(x, y) = \begin{cases} l_2 & m = 0 \\ l_{\text{nnfm}} + \lambda \cdot l_2 & m = 1 \end{cases}.$$
(7)

Note that the preservation loss $L^0$ is a crucial constraint to correct the gradient as mentioned in Section 4.1. We show the effectiveness of $L^0$ in Figure 7.

**Compositional Style Transfer.** Compositional style transfer assigns different style images for different mask regions. The loss function for label $m$ in compositional style transfer is defined as:

$$L^m(x, y) = l_{\text{nnfm}}(\mathbf{F}_r(x, y), \mathbf{F}_s^m) + \lambda \cdot l_2(\mathbf{F}_r(x, y), \mathbf{F}_c(x, y)),$$
(8)

where the feature $F_s^m$ represents the feature map extracted from the style image with label $m$.

**Semantic-aware Style Transfer.** The original ARF uses VGG features to match the nearest neighbor in the style image. However, VGG features encode more textural, structural, and color information compared with semantic information. Therefore, we use LSeg [26], which is trained together with a text encoder to extract semantic information for the nearest neighbor matching. Our **S**emantic **A**ware **N**earest **N**eighbor **F**eature **M**atching (SANNFM) function performs nearest neighbor matching between content and style features in both VGG ($\mathbf{F}_r^{VGG}, \mathbf{F}_s^{VGG}$) and LSeg ($\mathbf{F}_r^{LSeg}, \mathbf{F}_s^{LSeg}$) spaces for each specific pixel $x, y$ with label $m$:

$$\text{SANNFM}(x, y, m) = \text{argmin}_{x', y' \in S} D_{\text{sannfm}},$$
(9)

where $S = \{x', y' | M_s(x', y') = m\}$, and the distance function $D_{\text{sannfm}}$ is defined as a weighted average of the VGG cosine distance and LSeg cosine distance:

$$D_{\text{sannfm}} = \alpha \cdot D(\mathbf{F}_r^{VGG}(x, y), \mathbf{F}_s^{VGG}(x', y')) + (1 - \alpha) \cdot D(\mathbf{F}_r^{LSeg}(x, y), \mathbf{F}_s^{LSeg}(x', y')),$$
(10)

where $\alpha \in [0, 1]$ is a hyperparameter to control the weight of VGG and LSeg features and $D(.)$ is the cosine distance defined in Equation 4. The SANNFM style loss for the pixel $x, y$ is defined as:

$$l_{\text{sannfm}}(m, \mathbf{F}_r^{VGG}(x, y), \mathbf{F}_s^{VGG}, \mathbf{F}_r^{LSeg}(x, y), \mathbf{F}_s^{LSeg})$$
$$= \frac{1}{N} \sum_{x,y} D(\mathbf{F}_r^{VGG}(x, y), \mathbf{F}_s^{VGG}(\text{SANFFM}(x, y, m))).$$
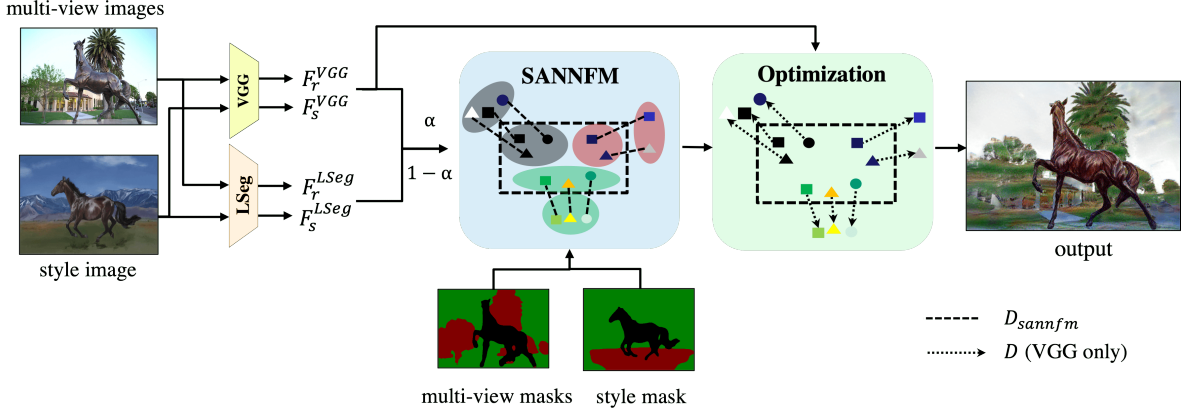(11)

Figure 2. **Pipeline of Semantic-aware Style Transfer.** The multi-view images and style image are used to extract VGG features $\mathbf{F}_r^{VGG}, \mathbf{F}_s^{VGG}$ and LSeg features $\mathbf{F}_r^{LSeg}, \mathbf{F}_s^{LSeg}$. Then the cosine distance is calculated and blended using the hyperparameter $\alpha$. As shown in the SANNFM module in the figure, we use the ellipses to represent different semantic labels, use the different shapes (square, triangle, circle) to represent semantic information of the pixels, use the color of the shape to represent the color and textural information of the pixels. The mixed distance is used to match the nearest neighbor in the style image with the same label for each pixel in the rendered image. Finally, the optimization uses VGG cosine distance only.
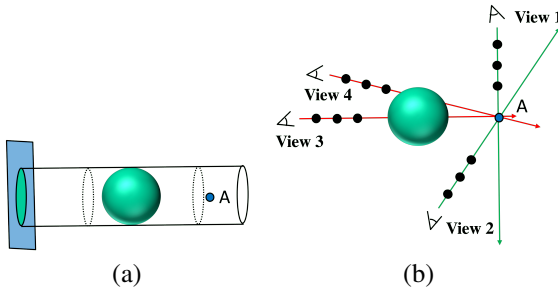


Figure 3. **Multi-view correction.** (a) For one view optimization, point A should be optimized using background loss, but optimized using foreground object loss. (b) Multi-view ray passing through point A. The first two views have the correct label for point A, represented as green rays, last two views have an incorrect label for point A, represented as red rays. The final gradient is dominated by correct loss gradients in the first two views.

Intuitively, the pipeline is shown in Figure 2, we search the nearest neighbor for a content image pixel only in the corresponding mask label of the style image. And the nearest neighbor metric is defined as a combination of VGG cosine distance and LSeg cosine distance. Finally, we only optimize the VGG cosine distance between the two features, which means LSeg feature is only used to match the features. The loss function for label $m$ in semantic-aware style transfer is defined as:

$$L^m(i,j) = l_{\text{sannfm}}(m, \mathbf{F}_r^{VGG}(x,y), \mathbf{F}_s^{VGG}, \mathbf{F}_r^{LSeg}(x,y), \mathbf{F}_s^{LSeg})$$
$$+ \lambda \cdot l_2(\mathbf{F}_r^{VGG}(x,y), \mathbf{F}_c^{VGG}(x,y)). \tag{12}$$

### 4.3. Implementation Details

For the VGG features extraction, we follow the same approach as ARF, using the conv3 block of VGG-16. For the LSeg features, we use the LSeg encoder similar to [21] to extract the feature for each pixel, and we use bilinear interpolation to match the resolution of VGG features. We set content loss weight $\lambda = 0.001$ in forward-facing captures, and $\lambda = 0.005$ in 360° captures, and the smoothness weight $\lambda_{\text{tv}} = 1$. During colorization, different from ARF, we only apply color transfer to the ground truth images as preprocessing and do not change the voxel grid, since we observe that the color transfer on the whole feature grid has an inaccurate rendering result since it is not semantically aware. For object selection, we only apply color transfer to the selected object in the ground truth image. And for compositional style transfer and semantic-aware style transfer, we apply color transfer to different ground truth regions using the corresponding semantic parts. All experiments are conducted with a single NVIDIA RTX 3090 Ti GPU.

## 5. Experiments

We conduct extensive experiments to evaluate object selection in Section 5.1, compositional style transfer in Section 5.2, and semantic-aware style transfer in Section 5.3. We also validate our 2D mask-based optimization and SANNFM module in the ablation studies in Section 5.4.

**Dataset.** We use a total of seven scenes including three forward-facing captures: *Flower, Fortress, Horns*, from [35], and four 360° captures: *Family, Horse, M60, Truck* from the *Tanks and Temples* dataset [20]. We use style images from the ARF style dataset, cartoon movies and other

paintings.

## 5.1. Object Selection

Our result for object selection is shown in Figure 4. We use ground truth mask from [7] and show the multi-view style transfer result for: *Flower, Fortress*, and *Truck* scene. As shown in the result, our method can correctly apply style transfer to the selected region, and leave other regions unchanged.

## 5.2. Compositional Style Transfer

Compositional style transfer results are shown in Figure 5. We compare our results with StyleRF [31], which uses volume rendering to generate a 2D feature map, and applies style transfer after the rendering. Thanks to the robustness of NNFM style constraint, our method can generate the composed style scenes that match the style images better in terms of brushstrokes and color, while StyleRF [31] generates over-photorealistic geometric structure, and fail to transfer the color of the style image. Also, we observe that our algorithm converges faster compared with StyleRF thanks to the efficient implementation of Plenoxels[8]. Since our method is optimization based, it can be further generalized to other radiance fields [2, 38] to either improve rendering quality or fasten the stylization convergence. But StyleRF cannot readily take advantage of other radiance fields.

## 5.3. Semantic-aware Style Transfer

We show comparisons between our semantic-aware style transfer and ARF [57] in Figure 6. We can observe that our results can match semantic information better compared with ARF. For example, in the horse scene, our results correctly match the color and texture of the horse (brown in the first style, dot in the second style), while our baseline ARF [57] can only matches the green color in the style image since the horse in ground truth image is green. This is because we use the semantic mask with the LSeg features to guide the nearest neighbor matching, while ARF [57] uses the VGG features only, which encodes more color and texture information.

**User Study.** We also conduct a user study to compare the style transfer quality between our semantic-aware style transfer and our baseline ARF [57]. Similar to Figure 6, we compare fourteen stylization results using the seven scenes of our dataset and two style images per scene that share similar semantics. For each example, we provide sample images from the photorealistic scene, the style image and the result multi-view videos from our method and ARF, and we let a large group of users select a preferred stylization or mark both results as similar. As a result, users prefer our result over ARF [57] 84% of the time, and 11.1% of the time think the results have similar quality. These results
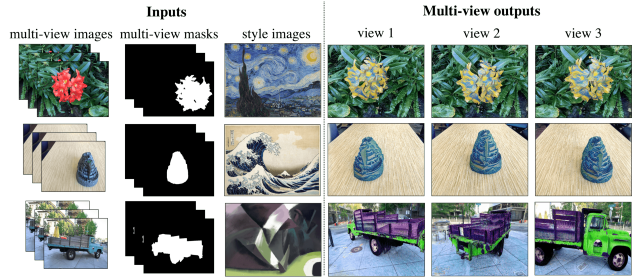


Figure 4. **Object selection result.** Our algorithm can generate the stylized object for the selected region and keep other objects photorealistic.
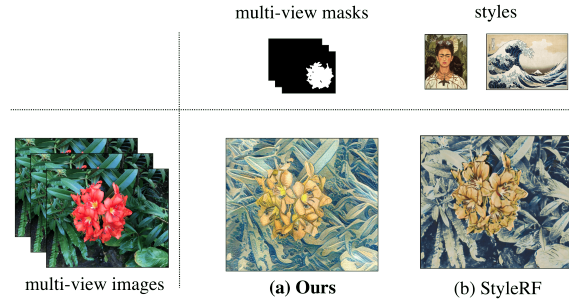


Figure 5. **Qualitative comparison of compositional style transfer.** Our result (a) can match the style image better in terms of the brushstrokes and color compared with StyleRF [31] (b).

show our method is preferable most of the time. We provide more detail in the supplementary materials.

## 5.4. Ablation Studies

**Multi-view Gradient Correction.** As mentioned in Section 4.1, the perservation constraint for object selection can correct the gradient for occluded objects. To prove the effectiveness of the constraint, we conduct an experiment to compare the stylized output with and without constraint. As shown in Figure 7, without the preserving constraint, many voxels are optimized incorrectly. As a result, there are some noisy pixels nearby the selected object. We call this effect style overflow. With the constraint, the incorrect gradients are corrected by the constraint gradient, thus the object selection is cleaner. This experiment shows that multi-view optimization can automatically correct the error gradient due to the column extended from the 2D mask.

**Mask Quality.** It is difficult to obtain ground truth masks in practice, therefore, we use two methods to extract masks. We can either use a text (such as *horse, flower*) or select a pixel belonging to an object directly in the image to get the mask using LSeg [26]. Thanks to our robust 2D mask-based optimization, we can use either ground truth masks or relatively more noisy masks generated by the aforementioned

Figure 6. **Comparison with the baseline method ARF[57] on real-world forward-facing and *Tanks and Temples* data.** Our method matches the semantics more faithfully compared with ARF.

two methods. As shown in Figure 8, although the pixel-based mask is noisy, our algorithm generates a reasonable result, which better matches the semantics in the style image significantly compared with ARF [57]. And the result using the text-based mask has comparable quality compared with the ground truth mask result.

**SANNFM.** Original ARF [57] uses the cosine distance of the VGG features to match the nearest neighbor, while our method only matches the nearest neighbor with the same se-

mantic label, and uses a combination of VGG and LSeg cosine distance. To prove the effectiveness of these two modifications, we perform an ablation study. As shown in Figure 9, in column a), if we only use mask constraint but no Lseg distance to find the nearest neighbor, we have some artefacts on the horse body in the stylized result. In column b) if we remove the mask and there is no semantic similarity between the ground truth and style image (flower scene), the model behaves as standard methods, whereas if there
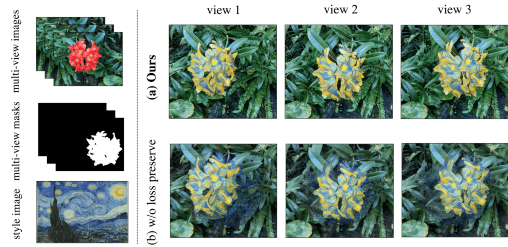
Figure 7. **Validation of the preservation constraint.** (a) shows our result, while (b) shows the result without the constraint. We can find that the preservation loss corrects the gradient, while style overflow happens when we remove this loss.
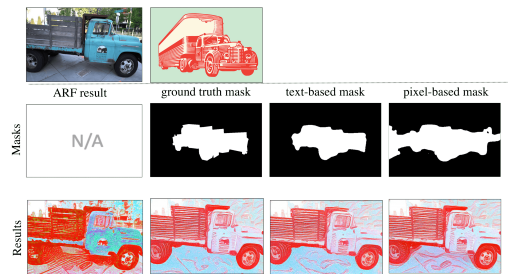


Figure 8. **Effect of mask quality on the style transfer.** We show how robust our method is with different 2D mask qualities. All of our methods generate more semantic-aware results compared with ARF [57]. And the quality of style transfer does not drop so much when the view consistency is not guaranteed. .
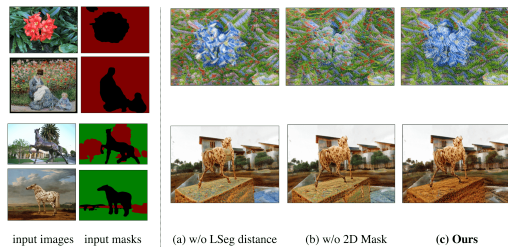


Figure 9. **Ablation studies for SANNFM.** The left column shows the ground truth images and masks, style images and masks. The right column shows (a) the result without LSeg distance (b) the result without mask constraint (c) our whole pipeline including LSeg distance and mask constraint.

is semantic similarity (horse scene), the model successfully transfers the style taking the semantics into account. This shows that our algorithm is robust when there is a correspondent object in the style image, but the 2D mask is necessary to provide more user control options when there is no correspondence. Column c) shows our results using both LSeg distance and 2D masks, which successfully transfer the style in both examples. In general, our method with mixed distance metric and masked constraint has a more robust controllability for semantic-aware style transfer.



Figure 10. **Different combination of VGG and LSeg features.** $\alpha = 0$ means we only use LSeg features, and $\alpha = 1$ means we only use VGG features. Note that all results are using the mask constraint.

**VGG & LSeg Features.** As shown in Figure 10, we show style transfer results using a different combination of VGG and LSeg distance. We can observe that if we only use the VGG features, the features matching has some artefacts on the horse. But if we only use the LSeg features, the result becomes noisy. Therefore, a combination of VGG and LSeg distance gives the best result by balancing the semantic and textural information.

# 6. Limitations

In this section, we discuss some limitations of the presented method and potential improvements. We notice that large scale differences between similar semantic objects in the scene and the style images could lead to undesired results. For example, if the scene object has a larger scale than the style object, the algorithm tends to copy multiple style objects to fill the scene object. One potential solution could be to use the corresponding segmentation masks to adjust the resolution of the style image. Besides, following ARF, we freeze the density field and only optimize the radiance field. However, we believe that optimizing the density could produce richer stylization results. Furthermore, our semantic-aware style transfer model is slower than ARF due to the computational requirement of the LSeg module.

# 7. Conclusion

In this paper, we propose different types of controllability including object selection, compositional style transfer, and semantic-aware style transfer. We achieve these tasks by multi-view 2D mask-based optimization with different constraints. This can be generalized to other optimization-based tasks such as CLIPNeRF [51] to achieve user-defined controllability. In addition, we propose SANNFM that combines the mask constraint with a mixed distance of VGG features and LSeg features. Our experiments show that our approach not only has better quality and robustness but also has better controllability given the user-defined semantics.

# 8. Acknowledgment

# References

[1] 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 *The IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2023. 3

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2, 6

[3] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 333–350. Springer, 2022. 2

[4] Yaosen Chen, Qi Yuan, Zhiqiang Li, Yuegen Liu Wei Wang Chaoping Xie, Xuming Wen, and Qien Yu. Upstnerf: Universal photorealistic style transfer of neural radiance fields for 3d scene. *arXiv preprint arXiv:2208.07059*, 2022. 3

[5] 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. 1, 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. *arXiv preprint arXiv:2209.08776*, 2022. 3, 6

[8] 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*, pages 5501–5510, 2022. 1, 2, 3, 6

[9] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 3

[10] Stephan J Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. Voltemorph: Real-time, controllable and generalisable animation of volumetric representations. *arXiv preprint arXiv:2208.00949*, 2022. 3

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[12] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 2

[13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization, 2017. 1

[14] 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. 1, 3

[15] Zixuan Huang, Jinghuai Zhang, and Jing Liao. Style mixer: Semantic-aware multi-style transfer network. In *Computer Graphics Forum*, pages 469–480. Wiley Online Library, 2019. 2

[16] Clément Jambon, Bernhard Kerbl, Georgios Kopanas, Stavros Diolatzis, George Drettakis, and Thomas Leimkühler. Nerfshop: Interactive editing of neural radiance fields. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(1), 2023. 3

[17] Yongcheng Jing, Yang Liu, Yezhou Yang, Zunlei Feng, Yizhou Yu, Dacheng Tao, and Mingli Song. Stroke controllable fast style transfer with adaptive receptive fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 238–254, 2018. 2

[18] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzciński, and Andrea Tagliasacchi. Conerf: Controllable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18623–18632, 2022. 2, 3

[19] Sunwoo Kim, Youngjo Min, Younghun Jung, and Seungryong Kim. Controllable style transfer via test-time training of implicit neural representation. *arXiv preprint arXiv:2210.07762*, 2022. 2

[20] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 5

[21] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *arXiv preprint arXiv:2205.15585*, 2022. 2, 3, 5

[22] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. 2

[23] Nicholas Kolkin, Michal Kucera, Sylvain Paris, Daniel Sykora, Eli Shechtman, and Greg Shakhnarovich. Neural neighbor style transfer. *arXiv e-prints*, pages arXiv–2203, 2022. 2

[24] Jan Eric Kyprianidis, John Collomosse, Tinghuai Wang, and Tobias Isenberg. State of the" art": A taxonomy of artistic stylization techniques for images and video. *IEEE transactions on visualization and computer graphics*, 19(5):866–885, 2012. 2

[25] Verica Lazova, Vladimir Guzov, Kyle Olszewski, Sergey Tulyakov, and Gerard Pons-Moll. Control-nerf: Editable feature volumes for scene rendering and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4340–4350, 2023. 2, 3

[26] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 3, 4, 6

[27] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2479–2486, 2016. 2

[28] Shaoxu Li and Ye Pan. Instant neural radiance fields stylization. *arXiv preprint arXiv:2303.16884*, 2023. 1, 3

[29] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017. 2

[30] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017. 2

[31] Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric Xing. Stylerf: Zero-shot 3d style transfer of neural radiance fields. 2023. 1, 2, 3, 6

[32] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 3

[33] Yang Miao, Iro Armeni, Marc Pollefeys, and Daniel Barath. Volumetric semantically consistent 3d panoptic mapping, 2023. 3

[34] 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. 3

[35] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2, 5

[36] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022. 3

[37] Shailesh Mishra and Jonathan Granskog. Clip-based neural neighbor style transfer for 3d assets. *arXiv preprint arXiv:2208.04370*, 2022. 3

[38] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1, 2, 3, 6

[39] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[40] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[41] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2

[42] Yicong Peng, Yichao Yan, Shengqi Liu, Yuhao Cheng, Shanyan Guan, Bowen Pan, Guangtao Zhai, and Xiaokang Yang. Cagenerf: Cage-based neural radiance field for generalized 3d deformation and animation. *Advances in Neural Information Processing Systems*, 35:31402–31415, 2022. 3

[43] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40:49–70, 2000. 2

[44] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[46] Eric Risser, Pierre Wilmot, and Connelly Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017. 1, 2

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 3

[48] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 2

[49] Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 3

[50] Bing Wang, Lu Chen, and Bo Yang. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. *arXiv preprint arXiv:2208.07227*, 2022. 3

[51] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 2, 3, 8

[52] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural

radiance fields stylization. *arXiv preprint arXiv:2212.08070*, 2022. 1

[53] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 106–122. Springer, 2022. 3

[54] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 2

[55] Tianhan Xu and Tatsuya Harada. Deforming radiance fields with cages. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 159–175. Springer, 2022. 3

[56] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: geometry editing of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18353–18364, 2022. 3

[57] 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. 1, 2, 3, 6, 7, 8

[58] Hui-Huang Zhao, Paul L Rosin, Yu-Kun Lai, and Yao-Nan Wang. Automatic semantic style transfer using deep convolutional neural networks and soft masks. *The Visual Computer*, 36:1307–1324, 2020. 2

[59] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 3