ARF-Plus: Controlling Perceptual Factors in Artistic Radiance Fields for 3D Scene Stylization

Wenzhao Li[1], **Tianhao Wu**[1], **Fangcheng Zhong**[1] **Cengiz Oztireli**[1, 2]
[1]UniversityofCambridge
[2]Google Research
wl301@cam.ac.uk, tw554@cam.ac.uk, fz261@cam.ac.uk, aco41@cam.ac.uk

Abstract

The radiance fields style transfer is an emerging field that has recently gained popularity as a means of 3D scene stylization, thanks to the outstanding performance of neural radiance fields in 3D reconstruction and view synthesis. We highlight a research gap in radiance fields style transfer, the lack of sufficient perceptual controllability, motivated by the existing concept in the 2D image style transfer. In this paper, we present ARF-Plus, a 3D neural style transfer framework offering manageable control over perceptual factors, to systematically explore the perceptual controllability in 3D scene stylization. Four distinct types of controls - color preservation control, (style pattern) scale control, spatial (selective stylization area) control, and depth enhancement control - are proposed and integrated into this framework. Results from real-world datasets, both quantitative and qualitative, show that the four types of controls in our ARF-Plus framework successfully accomplish their corresponding perceptual controls when stylizing 3D scenes. These techniques work well for individual style inputs as well as for the simultaneous application of multiple styles within a scene. This unlocks a realm of limitless possibilities, allowing customized modifications of stylization effects and flexible merging of the strengths of different styles, ultimately enabling the creation of novel and eye-catching stylistic effects on 3D scenes.

1 Introduction

The neural style transfer technique (Gatys, Ecker, and Bethge 2016) unravels the mysteries of the artist. With the aid of this technology, a photograph may look to have been produced by a well-known artist. Video style transfer (Ruder, Dosovitskiy, and Brox 2018) extends the original neural style transfer from the 2D image domain to the temporal domain as well. 3D scene style transfer further extends it to the spatial-temporal domain, offering enhanced artistic transformations and immersive aesthetic experiences.

Nonetheless, 3D scene style transfer is not an easy task due to the domain gap between 2D and 3D. Traditional 3D scene stylization techniques (Huang et al. 2021; Mu et al. 2022; Yin et al. 2021; Höllein, Johnson, and Nießner 2022) are often based on mesh or point-cloud representations. The limited reconstruction accuracy and discrete geometric representation in these models constrain the stylization performance, resulting in observable artifacts in challenging real-world circumstances (Zhang et al. 2022). The Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) represents the optics of the 3D scene without explicit geometry which is inherently discontinuous. With its superior performance in 3D reconstruction and view synthesis, many researchers have attempted to perform 3D stylization based on radiance fields (Chiang et al. 2022; Huang et al. 2021, 2022; Zhang et al. 2022). However, they have focused on optimizing a unified style transfer across the whole scene, neglecting the control of perceptual factors that influence visual perception.

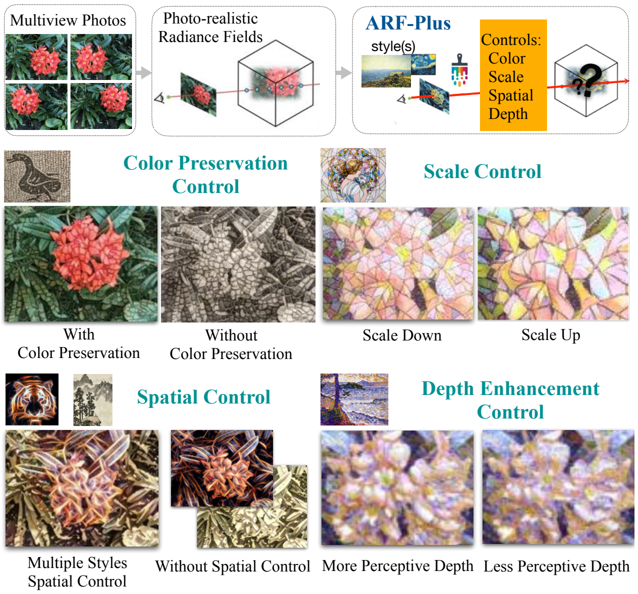

Figure 1: An overview of ARF-Plus framework. It facilitates the control of various perceptual factors and permits the use of multiple style images as input, resulting in perceptual flexibility in 3D scene stylization.

Color preservation control (Gatys et al. 2016), stroke size control (Jing et al. 2018; Reimann et al. 2022), spatial control (Castillo et al. 2017; Yu, Wu, and Wang 2019; Gatys et al. 2017), and other types of perceptual controls have been widely applied in the image domain style transfer. The research gap in the radiance fields style transfer hinders the diversity of choices and potential applications in 3D scene style transfer. In certain application scenarios, users may desire to selectively apply an abstract art style only to the scene's background, while keeping the main objects (e.g., people) in a realistic appearance. Alternatively, users may wish to add unique aesthetic effects to specific objects to create cartoon-like characters. Moreover, users might prefer to retain the original colors of the scene while applying transferred patterns and textures from a given style. Furthermore, users may have a preference to adjust the size of patterns or brushstrokes in the stylized outcomes for aesthetic adjustments. All these diverse 3D radiance fields style transfer requirements and numerous promising application scenarios are yet to be fully realized.

Hence, exploring perceptual control in 3D stylization on radiance fields and developing algorithms that offer perceptual-feature-aware controls hold promising potential for meeting application-level demands. We introduce ARF-Plus, a framework that enables the manipulation of four perceptual factors in 3D stylization, namely, color preservation, stroke scale, selective spatial area, and depth-aware perception. Our main contributions are listed as follows:

- We introduce ARF-Plus, a framework to *simultaneously* manipulate the influence of various perceptual factors on 3D stylization. To the best of our knowledge, we are the *first* to explore the control of various perceptual factors in the stylization of 3D models based on radiance fields.
- We demonstrate the effectiveness of ARF-Plus in controlling the aforementioned perceptual factors on 3D stylization through a series of qualitative and quantitative evaluations.

2 Background and Related Work

3D style transfer entails applying a reference style's visual characteristics to a target 3D scene while maintaining the 3D structure itself. Due to the exceptional 3D reconstruction capabilities of radiance fields, radiance fields style transfer has emerged as a promising approach to enhance the outcomes of 3D style transfer. Recent studies (Chiang et al. 2022; Fan et al. 2022; Huang et al. 2022; Nguyen-Phuoc, Liu, and Xiao 2022; Zhang et al. 2022; Chen et al. 2022b) have shown that style features can be transferred from a given style to an entire real 3D scene represented by radiance fields.

Zhang Kai et al. (Zhang et al. 2022) present the Artistic Radiance Fields (ARF) approach, which is highly effective at capturing style characteristics while preserving multi-view consistency. They present and implement a nearest neighbour-based feature matching (NNFM) loss, which can produce a consistent multi-view visual experience and achieve state-of-the-art visual quality. Additionally, their approach is independent of the radiance fields representation, making it applicable to a wider range of systems, including Plenoxels (Yu et al. 2021), NeRF (Mildenhall et al. 2021), TensoRF (Chen et al. 2022a), and others. We choose ARF as the basis of our architecture for two primary reasons. Firstly, ARF is currently one of the best-performing state-of-the-art methods. It captures style details while maintaining multi-view consistency and produces more convincing outcomes than other state-of-the-art techniques such as Huang et al. (Huang et al. 2022) and Chiang et al. (Chiang et al. 2022). Secondly, the ARF method is versatile and adaptable to various radiance fields representations. As a result, our ARF-Plus can benefit from ARF's exceptional performance and inherit ARF's broad applicability.

Based on our best knowledge, the concept of perceptual control has not been well utilized in the radiance fields style transfer. NeRF-Art (Wang et al. 2023) achieves text-driven neural radiance fields stylization, by addressing the mapping between a text prompt and a particular given style or aesthetic pattern. However, its matched stylized result with a given text prompt is still a uniformly stylized scene. Ref-NPR (Zhang et al. 2023) proposes a ray registration process, in order to maintain cross-view consistency and establish semantic correspondence in transferring style across the entire stylized scene. While Ref-NPR mentioned the concept of "controllable", its notion of controllability pertains to semantic correspondence and geometrically consistent stylizations. This differs from the user-level controllability discussed in this paper, where we aim to offer enhanced flexibility, enabling individuals to effectively customize controlled stylization. Nevertheless, similar to the demand for customized and diverse 2D image style transfer, there exists a substantial market for radiance field style transfer with personalized and varied characteristics. Therefore, our primary objective is to fill this research gap, and the subsequent sections will present our initial endeavours and discoveries.

3 Methodology

As illustrated in Figure 1, ARF-Plus involves the conversion of a photo-realistic radiance field, reconstructed from photographs of a real-world scene, into an artistic style radiance fields with a variety of perceptual factors controls. Initially, we reconstruct photo-realistic radiance fields from multiple photographs to represent a 3D scene. Next, with the given style image(s), we stylize the photo-realistic radiance fields in a fine-tuning process. The style transfer process of radiance fields is formulated as a problem of optimisation. Images are rendered from the radiance fields across different viewpoints in a differentiable manner. A variety of loss functions and gradient updating strategies are employed based on the selected perceptual controls. After the stylization process is completed, we achieve the stylized 3D scene radiance fields and can generate free-viewpoint and consistent stylized renderings from it.

3.1 Color Preservation Control

The general radiance fields style transfer algorithms convert the photo-realistic radiance fields, which depict a real 3D scene, to match the style of a specified style image. Once the stylization process is complete, the color distribution of the style reference is also transferred. However, in some cases, one may prefer to retain the colors of the photo-realistic 3D scene and only learn the painting patterns, while ensuring that stylization does not result in unnatural colors (e.g., the sky turning green or the grass becoming red).

Our method is inspired by the luminance-only algorithm proposed by Gatys et al. (Gatys et al. 2016). Their work is based on the utilization of the YIQ color space, which effectively separates luminance in Y, and color information in

I and Q. As shown in Equation 1, we extract the luminance channel in the rendered view $\hat{\mathbf{X}^{Y}}$ and the luminance channel of the style image $\mathbf{X}_{\text{style}}^{Y}$ for style loss $\mathcal{L}_{\text{style}}$ calculation, while keeping the RGB channels for the content loss $\mathcal{L}_{\text{content}}$ calculation.

$$\begin{aligned}\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{style}}(F_{\ell_s}(\hat{\mathbf{X}^{Y}}), F_{l_s}(\mathbf{X}_{\text{style}}^{Y})) \\ + \beta \cdot \mathcal{L}_{\text{content}}(F_{\ell_c}(\hat{\mathbf{X}}), F_{\ell_c}(\mathbf{X}_{\text{content}})) + \gamma \cdot \mathcal{L}_{\text{tv}}\end{aligned} \tag{1}$$

The style loss $\mathcal{L}_{\text{style}}$ enables the radiance fields to focus on the luminance variations (of pattern shapes) present in the style image, without explicitly learning its color. The content loss $\mathcal{L}_{\text{content}}$ quantifies the RGB color differences between the rendered view $\hat{\mathbf{X}}$ and the photo-realistic ground truth $\mathbf{X}_{\text{content}}$ at the level of content features. $\mathcal{L}_{\text{tv}}$ is the total variation regularizer. F_{ℓ_c} and F_{ℓ_s} stand for the representation of content features and style features, respectively. α, β and γ are scalar weights.

3.2 Scale Control

Stroke sizes and pattern shapes in drawing refer to two fundamental aspects of mark-making and line creation within an artistic composition. Stroke sizes pertain to the thickness or width of lines, brushstrokes, or marks used in a drawing, as shown in Figure 3 (a). Pattern shapes involve the specific forms, designs, or arrangements of strokes or marks that are repeated or used consistently within the artwork, as shown in Figure 3 (b). Artists have their own way of arranging strokes of different sizes and shape patterns, which reflects unique artistic "styles". For instance, Pablo Picasso favored simple shapes and a limited color palette when painting objects, people, and landscapes, whereas Claude Monet employed small, playful brush strokes to create his impressionist works. It is worth noting that, in some cases, the stroke and pattern sizes of the original style image may not be appropriate for the 3D scene after the style transfer has been performed, due to the differences (e.g. in size) between the depicted content in the style image and the 3D scene. Therefore, adjusting the pattern scale to match the 3D scene becomes an issue that requires attention.

Single Style Scale Control Inspired by the work of Jing Y. et al. (Jing et al. 2018) on stroke-controllable image style transfer, we take into account the two factors mentioned in their work that influence the scale of style patterns: the receptive field and the size of the style image.

$$\mathcal{L}_{\text{style}}^{\mathcal{T}_S} = \sum_{l \in \{l_1,\dots,l_5\}} w_l \cdot \mathcal{L}_{\text{style}}\left(\mathcal{F}_l\left(\mathcal{R}\left(\mathbf{X}_{\text{style}}, \mathcal{T}_S^l\right)\right), \mathcal{F}_l\left(\hat{\mathbf{X}}\right)\right) \tag{2}$$

As shown in Equation 2, $\mathcal{F}_l$ represents the feature map at l layer blocks, where l can be one of the five VGG-16 layer blocks $\{l_1, l_2, ..., l_5\}$. We adjust the range of the receptive field by altering the selection of l. $\mathcal{R}$ denotes the resize function for the style image. According to the specified scale $\mathcal{T}_S^l$ for each layer block, $\mathcal{R}$ resizes the style image for each layer block for feature extraction. The style losses from different layer blocks are subsequently combined together to form the total style loss $\mathcal{L}_{\text{style}}^{\mathcal{T}_S}$, with w_l controlling the corresponding weight. By having the flexibility to adjust both the size of the style image and the range of the receptive field, we are able to attain continuous control over the scale.

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{style}}^{\mathcal{T}_S} + \beta \cdot \mathcal{L}_{\text{content}}(F_{\ell_c}(\hat{\mathbf{X}}), F_{\ell_c}(\mathbf{X}_{\text{content}})) + \gamma \cdot \mathcal{L}_{\text{tv}} \tag{3}$$

Multiple settings might be purposefully employed to highlight either the delicate patterns or the harsh outlines within a single style image. Giving different weights to the losses of shallow and deep layers will achieve this. The limitless adjustment of scaling settings also allows for the creation of innovative aesthetic patterns. The total style loss $\mathcal{L}_{\text{style}}^{\mathcal{T}_S}$ is then added to the style transfer loss function to provide scale control in 3D scene stylization, as shown in Equation 3. There are three key aspects where our work differs from that of Jing Y. et al. (Jing et al. 2018): 1) In contrast to 2D image style transfer, we use the improved style transfer loss function for 3D radiance fields style transfer, leading to a different overall model architecture. 2) In order to ensure consistency across multi-view scenes, we utilize NNFM loss (Zhang et al. 2022) rather than the Gram matrix to calculate the style loss. 3) We optimized their continuous size control strategy. In addition to assigning different weights to the output feature maps, we allow alternative style image size $\mathcal{T}_S^l$ for each VGG layer in order to produce more adjustable results. In other words, it is possible to have different $\mathcal{T}_S^l$ for each layer in our method.

Multiple Style Scale Control There may arise a need to blend multiple styles. For instance, it may be desirable to incorporate the fine brushstrokes of one painting with the bigger shapes apparent in another artwork and blend them seamlessly into the 3D scene. It allows for fusing the benefit of two styles in this way. As shown in Equation 4 and 5, our approach can readily be extended to scenarios involving a mixture of multiple styles.

$$\mathcal{L}_{\text{style}}^{\mathcal{T}_M} = \sum_i \lambda_i \mathcal{L}_{\text{style}_i}^{\mathcal{T}_S} \tag{4}$$

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{style}}^{\mathcal{T}_M} + \beta \cdot \mathcal{L}_{\text{content}}(F_{\ell_c}(\hat{\mathbf{X}}), F_{\ell_c}(\mathbf{X}_{\text{content}})) + \gamma \cdot \mathcal{L}_{\text{tv}} \tag{5}$$

The scalar weight λ_i regulates the influence of each individual style image $\mathbf{X}_{\text{style}_i}$. The combination loss $\mathcal{L}_{\text{style}}^{\mathcal{T}_M}$ serves as the style loss term in the style transfer loss $\mathcal{L}_{\text{total}}$.

3.3 Spatial Control

General 3D style transfer approaches typically apply the style globally to the entire scene, producing a uniform stylization throughout the whole scene. However, there are circumstances where it is necessary to only stylize a specific area or object within a scene. For instance, one may desire to create an immersive virtual world experience by turning the background environment into a fantasy watercolor painting effect without affecting the people or main subjects of a scene. Another example is creating an augmented reality effect by turning the primary objects of a scene into cartoon-like, abstract figures while keeping the background intact.

Single Style Spatial Control Our spatial control is integrated with the deferred back-propagation technique in ARF (Zhang et al. 2022). During the deferred back-propagation process, the loss and gradient of the predicted view are calculated with auto-differentiation disabled. Each 2D predicted view is associated with a cached-gradients map of the same size. The key aspect of our spatial control is applying a binary spatial mask $\mathbf{T}^r$ to the cached-gradients map $\hat{\mathbf{X}}_{\text{grad}}$, as shown in Equation 6, where $\circ$ denotes element-wise multiplication. Each element in $\mathbf{T}^r$ takes either a value of 0 or 1. The regions in the mask with a value of 0 correspond to the portions of the predicted view that are not subject to stylization, while the regions with a value of 1 reflect the areas of the view that need to be stylized. As a result, only the gradients corresponding to the chosen regions are used in the computation throughout the optimization step where the cached gradients are back-propagated.

$$\hat{\mathbf{X}}_{\text{grad}} = \mathbf{T}^r \circ \hat{\mathbf{X}}_{\text{grad}} \tag{6}$$

The source of the spatial mask is not limited and can be manually annotated. In this paper, we employ semantic segmentation maps $\mathbf{T}^r_{\text{semantic}}$ to determine the regions of specific semantic objects within the scene.

Multiple Style Spatial Control To accomplish the spatial control of multiple styles, we introduce a corresponding method called "Combined Cached-gradients Map", which still utilises the concept of the cached-gradients map in ARF (Zhang et al. 2022). Assuming the scene contains r areas that need individually corresponding to a unique style image $\mathbf{X}^r_{\text{style}}$. As shown in Equation 7 and 8, for each area $\mathbf{T}^r \circ \hat{\mathbf{X}}$ and its related style $\mathbf{X}^r_{\text{style}}$, we initially determine the corresponding cached gradients maps $\hat{\mathbf{X}}^r_{\text{grad}}$ according to the style transfer loss $\mathcal{L}^r_{\text{total}}$ with auto-differentiation disabled. $\hat{\mathbf{X}}^r_{\text{grad}}$ are then added to generate the final cached gradients maps $\hat{\mathbf{X}}_{\text{grad}}$ applied in deferred back-propagation. As a result, during the optimization process, each selected region is updated according to its target style.

$$\begin{aligned}\mathcal{L}^r_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{style}}(F_{\ell_s}(\mathbf{T}^r \circ \hat{\mathbf{X}}), F_{l_s}(\mathbf{X}^r_{\text{style}})) + \gamma \cdot \mathcal{L}_{\text{tv}} \\ + \beta \cdot \mathcal{L}_{\text{content}}(F_{\ell_c}(\mathbf{T}^r \circ \hat{\mathbf{X}}), F_{\ell_c}(\mathbf{T}^r \circ \mathbf{X}_{\text{content}}))\end{aligned} \tag{7}$$

$$\hat{\mathbf{X}}_{\text{grad}} = \textstyle\sum_r \mathbf{T}^r \circ \hat{\mathbf{X}}^r_{\text{grad}} \tag{8}$$

3.4 Depth Enhancement Control

While the style transfer technique yields impressive effects by evenly applying style patterns across the entire scene, there are instances where it inadvertently compromises the scene's depth perception. According to some research findings (Liu et al. 2017; Jing et al. 2019), stylized results are especially unsatisfactory for situations with a large range of depths. This happens as a result of the uncontrollable alteration of the content layout, which obscures the line separating the foreground from the background and the boundaries between various objects.

Our depth-aware control technique is based on the Liu et al (Liu et al. 2017)'s depth-aware approach for 2D image style transfer. They note that the depth map serves as an effective representation of the spatial distribution within an image, and propose to preserve the depth map of the content during stylization. It is important to note that, to achieve smooth gradient optimization, ARF (Zhang et al. 2022) fixes the density component in the radiance fields during stylization. Hence, the actual depth information within the radiance fields remains unchanged, while only the appearance of the 3D scene is altered. However, as the style transfer is done in a 2D per-view way without any knowledge of the scene depth or geometry, object boundaries can be easily blurred or lost during stylization. Hence, we rely on a pre-trained depth estimation network ϕ_1 (Ranftl et al. 2020) to approximate the perceived depth of the current stylized appearance, as the depth information remains fixed in the radiance field. This can assist ARF-Plus in promptly understanding the changes in stylization's appearance and whether they would result in harmful alterations to the visually perceived depth.

In order to preserve depth-aware information, we take the outputs of the pre-trained depth estimation network ϕ_1 and compute the (squared, normalized) Euclidean distances as the $\mathcal{L}^{\phi_1}_{\text{depth}}$. The photo-realistic view $\mathbf{X}_{\text{content}}$ and stylized view $\hat{\mathbf{X}}$ have the same size $H \times W \times C$.

$$\mathcal{L}^{\phi_1}_{\text{depth}} = \frac{\left\|\phi_1(\hat{\mathbf{X}}) - \phi_1(\mathbf{X}_{\text{content}})\right\|_2}{H \times W \times C} \tag{9}$$

As shown in Equation 10, the depth loss $\mathcal{L}^{\phi_1}_{\text{depth}}$ is incorporated into the style transfer loss $\mathcal{L}_{\text{total}}$, where δ is a scalar weight. $\mathcal{L}^{\phi_1}_{\text{depth}}$ serves as an additional aspect to guide and influence the stylization based on depth-aware information.

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{style}} + \beta \cdot \mathcal{L}_{\text{content}} + \delta \cdot \mathcal{L}^{\phi_1}_{\text{depth}} + \gamma \cdot \mathcal{L}_{\text{tv}} \tag{10}$$

4 Experiments and Evaluation

4.1 Experiments Design

Baseline To the best of our knowledge, we are the first to systematically address the issue of enhancing perceptual flexibility controls for radiance fields style transfer. As such, there are no similar works in the current state-of-the-art that can be compared to ours. Since our proposed ARF-Plus framework is based on the ARF (Zhang et al. 2022), we choose it as the baseline for our experimental evaluation.

Datasets To guarantee comparability with the baseline and achieve more accurate differentiation, we utilise the same datasets as those employed in ARF (Zhang et al. 2022).
a) Artworks Dataset (Zhang et al. 2022): This dataset is made up of a variety of images which act as given styles. It is used to assess the effectiveness of our proposed ARF-Plus framework in learning and managing multiple styles while adhering to specific perceptual factor controls.
b) LLFF Dataset (Mildenhall et al. 2019): This dataset contains real-world scenes that are photographed from forward angles. It is utilized to create photo-realistic radiance fields, representing the 3D scenes prior to stylization.
c) Real 360 Degree Scenes Dataset (Knapitsch et al. 2017): This dataset includes real, unbounded, 360-degree large outdoor scenes with complex geometric layouts and camera trajectories. The scene data is also employed to train photo-realistic radiance fields before stylization.

Training Settings Identical to the baseline, in ARF-Plus, we configure the stylization optimization to run for 10 epochs, with a learning rate that exponentially decreases from 1e-1 to 1e-2. For single-style input, the training time of our ARF-Plus is comparable to that of baseline, as it does not entail additional training tasks during the stylization optimization process, which typically takes around 20 minutes on a single GPU.

Implementation Details Our ARF-Plus framework, as previously mentioned, comprises reconstructing photo-realistic radiance fields from numerous pictures and then stylizing the resulting reconstruction with perceptual factor choices. 1) Regarding the first step of reconstructing a photo-realistic radiance field, our work aligns with the baseline approach in primarily relying on Plenoxels (Yu et al. 2021) to achieve. Plenoxels is renowned for its fast reconstruction and rendering speed. All training settings of our method in this process remain consistent with the baseline. 2) Although ARF (Zhang et al. 2022)'s core algorithm achieves acceptable style transfer in both style pattern and color, its authors introduce a recoloring strategy in the stylization process to enhance visual quality. In ARF-Plus, we continue to employ the recoloring method for scale control and depth enhancement control in single-style cases. However, we choose to disable the recoloring process in scenarios where color preservation control, spatial control, and multiple input styles are involved. In those instances, it is inappropriate to follow the recoloring process in ARF which uses the colors of a given style image to recolor the entire scene. To ensure a fair comparison, when our ARF-Plus deactivates the recolor process, the baseline ARF also disables this step accordingly. Please refer to the Appendix for additional technical details.

4.2 Color Preservation Control

Figure 2 demonstrates visual comparisons between methods applied to real-world scenes. In the context of color preservation control, our goal is to ensure that the colors of the stylized scene remain consistent with those of the photo-realistic radiance fields. As illustrated, our proposed color preservation control method effectively preserves the original scene's color while successfully learning the desired style. Moreover, the colors and brightness distribution of the stylized outputs achieved through our method closely resemble those of the photo-realistic view. Please refer to the Appendix for additional qualitative and quantitative evaluations.

4.3 Scale Control

We demonstrate the effects of scale control on both single and multiple styles. Please refer to the Appendix for additional experiments regarding the study of influential factors of scale control.

Single Style Scale Control Figure 3 illustrates the successful scaling up and scaling down of style patterns achieved by our scale control method. In addition, this approach provides enhanced flexibility for resizing fine strokes or coarse patterns by selectively assigning different weights to various receptive fields and generating diverse resized style images for each receptive field. In Figure 3 (a), by configuring the range of receptive field, specifically selecting Conv1 and Conv2 for feature extraction (setting $w_{l_1} = 0.3, w_{l_2} = 0.7$ and the rest to 0 in Equation 2), style transfer places greater emphasis on capturing fine details from the style image. It can be observed that the fine strokes - the green and white lines - are successfully scaled up or down in the stylized scenes. In Figure 3 (b), by selecting Conv3 and Conv4 (setting $w_{l_3} = 0.7, w_{l_4} = 0.3$ and the rest to 0 to in Equation 2), style transfer focuses more on learning coarse patterns. Different layers are then assigned different values of the resize setting of the style image to achieve a continuously adjustable scale. The results of the stylized scene demonstrate that the coarse shape patterns are effectively scaled up or down.

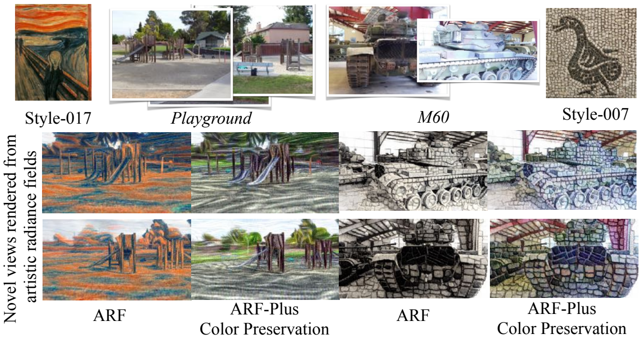
Style-017
Playground
M60
Style-007
Novel views rendered from artistic radiance fields
ARF
ARF-Plus Color Preservation
ARF
ARF-Plus Color Preservation

Figure 2: Qualitative results of color preservation. Please refer to supplementary materials for better visualization.

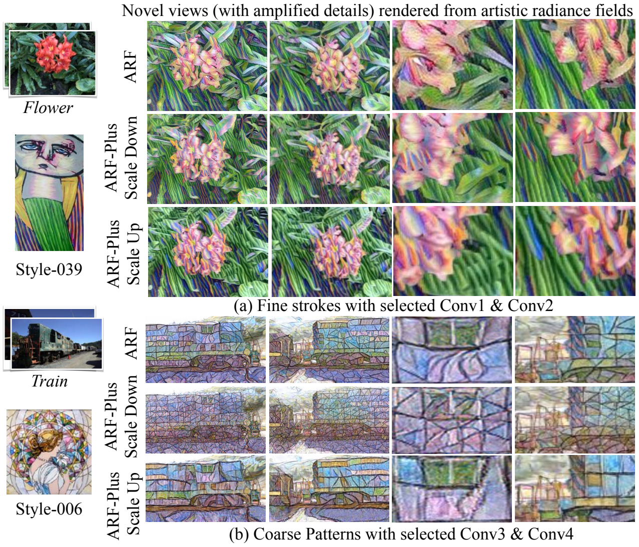
Novel views (with amplified details) rendered from artistic radiance fields
Flower
Style-039
ARF
ARF-Plus Scale Down
ARF-Plus Scale Up
(a) Fine strokes with selected Conv1 & Conv2
Train
Style-006
ARF
ARF-Plus Scale Down
ARF-Plus Scale Up
(b) Coarse Patterns with selected Conv3 & Conv4

Figure 3: Qualitative results of single style scale control. Our scale control effectively achieves style scaling. Please refer to supplementary materials for better visualization.

Multiple Style Scale Control After analyzing how well our scale control on individual styles, we move on to examine its effectiveness in handling multiple blended styles. We utilize color preservation to minimize the visual influence of different colors from multiple styles, enabling a more effective comparison of pattern scale variations. Figure 4 illustrates how our ARF-Plus scale control method, when applied to blended styles, effectively controls each individual style's scale. The noticeable changes in the size of leaf patterns are evident when scaling up or down the coarse pattern of Style-121. When the fine pattern scale changes for Style-122, there are observable effects on the leaves and petals.

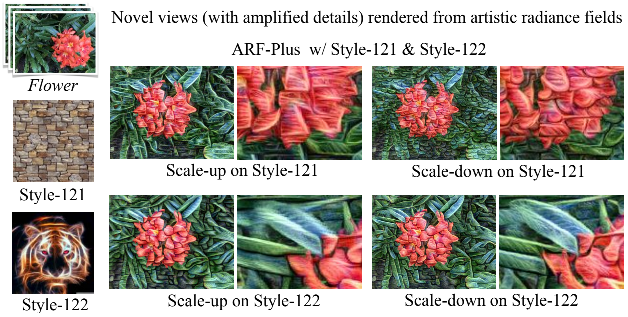

Figure 4: Qualitative results of multiple styles scale control. For better visualization, color is preserved by using our color preservation control. Our ARF-Plus with scale control can selectively scale up or scale down a particular style in stylization with blended multiple styles.

4.4 Spatial Control

Single Style Spatial Control Figure 5 presents the results of spatial control with semantic segmentation masks. Our method successfully stylizes the associated elements of the scene *Room*, namely the chairs, table, and TV. A noteworthy observation is that, despite the semantic segmentation masks being occasionally inaccurate or not fully covering the objects, our method manages to effectively paint nearly whole objects' areas in the stylized scene. One possible explanation is multiple training perspectives partially offset the shortcomings of semantic segmentation. Although certain views lead to poor segmentation results, alternative views' semantic maps provide broader and more accurate coverage. This enables the successful coloring of the overlapping regions across multiple views. Ultimately, the entire area of the semantic object is successfully stylized.

We set the number of rendered novel views N to 120 and render a series of images from the stylized radiance fields. Table 1 provides quantitative results of the stylized scenes. In comparison to ARF in the mask, our spatial control yields overall superior ArtFID scores (Wright and Ommer 2022), which can be attributed to a significant enhancement in styleFID. There are potentially two reasons behind the improvement in masked styleFID scores. Firstly, spatial control leads to a reduction in the regions requiring stylization within the scene. This results in fewer areas participating in gradient back-propagation during the same training epoch, allowing for the acquisition of more style information. Secondly, the appearance of semantic objects (such as TV) often exhibits similar textures. The uniform texture makes it easier for those areas to learn the style. It is also notable that in Scene *Room*, for ARF-Plus, the individual use of three separate semantic masks results in higher StyleFID values compared to the combined use of all three masks as a single mask (Room-All). Because ArtFID integrates both ContentDist and StyleFID, changes in StyleFID contribute to an elevated ArtFID score. This reveals that the size of the area controlled by spatial control has an impact on the effectiveness of style transfer. When the mask is too small, indicating a limited stylized area, it becomes difficult for that region to encompass all the style patterns depicted in the style image. Please refer to the Appendix for the explanation of calculating ArtFID on specific areas and additional qualitative evaluation.

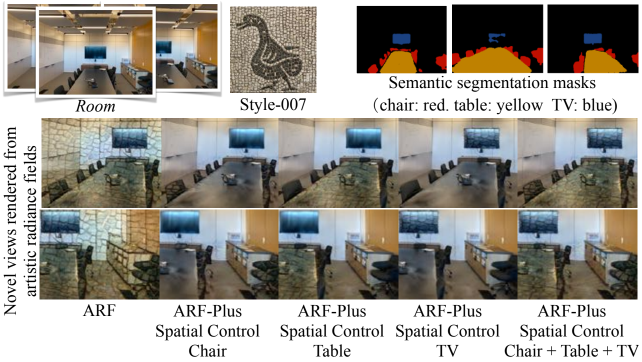
Room
Style-007
Semantic segmentation masks
(chair: red. table: yellow TV: blue)
Novel views rendered from artistic radiance fields
ARF
ARF-Plus Spatial Control Chair
ARF-Plus Spatial Control Table
ARF-Plus Spatial Control TV
ARF-Plus Spatial Control Chair + Table + TV

Figure 5: Qualitative results of spatial control with semantic segmentation masks. Our ARF-Plus with spatial control effectively stylizes specific semantic objects - chair, table, and TV - within the scene. Please refer to supplementary materials for better visualization.

Multiple Style Spatial Control As shown in Figure 6, compared to the baseline's unified stylization, our proposed spatial control method effectively transfers various styles to different regions. The modification of the loss function (in Eq. 7) may result in minor variations in the details. More specifically, our method calculates differences solely between a specific region and its corresponding style and segmented photo-realistic view. In contrast, the baseline ARF (Zhang et al. 2022) computes the loss differences between the entire scene, style, and the overall photo-realistic view. Please refer to the Appendix for intermediate training results.

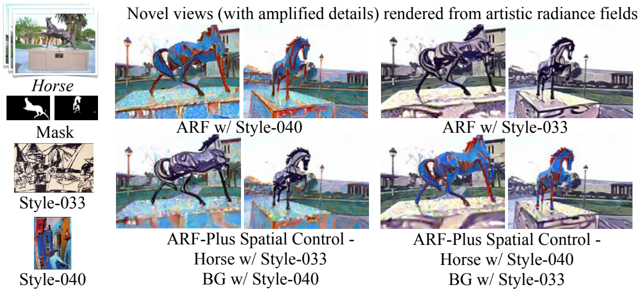
Novel views (with amplified details) rendered from artistic radiance fields
Horse
Mask
Style-033
Style-040
ARF w/ Style-040
ARF w/ Style-033
ARF-Plus Spatial Control - Horse w/ Style-033 BG w/ Style-040
ARF-Plus Spatial Control - Horse w/ Style-040 BG w/ Style-033

Figure 6: Qualitative results of multiple styles spatial control. Our multi-style spatial control successfully stylizes various locations with distinct styles. Please refer to supplementary materials for better visualization.

Table 1: Quantitative results of spatial control with semantic segmentation masks. Evaluation metric ArtFID (Wright and Ommer 2022) is a combination of ContentDist and StyleFID. Room-All represents the combined usage of three semantic masks: Chair, Table, and TV. Results better than the baseline in the mask are in bold.

		ArtFID ↓		ContentDist ↓		StyleFID ↓	
Style	Scene	ARF in Mask	ARF-Plus w/ Spatial Control in Mask	ARF in Mask	ARF-Plus w/ Spatial Control in Mask	ARF in Mask	ARF-Plus w/ Spatial Control in Mask
007	*Room*-Chair	41.4614	**39.3775**	0.0435	**0.0434**	38.7340	**36.7384**
007	*Room*-Table	43.2753	**40.8616**	**0.0473**	0.0489	40.3191	**37.9581**
007	*Room*-TV	41.2685	**38.4802**	0.0195	**0.0193**	39.4799	**36.7500**
007	*Room*-All	41.0287	**37.9573**	**0.0950**	0.0983	36.4707	**33.5591**

4.5 Depth-aware Control

Figure 7 presents comparisons between our depth-aware control method and the baseline. Qualitative results demonstrate that our method can successfully preserve perceptive depth across varying views. The foreground leaf has a more distinct outline and stands out from the other leaves in the background. For the *Horns* scene, our depth control method better preserves the nostril (hole) located next to the horn. Please refer to the Appendix for quantitative evaluations.

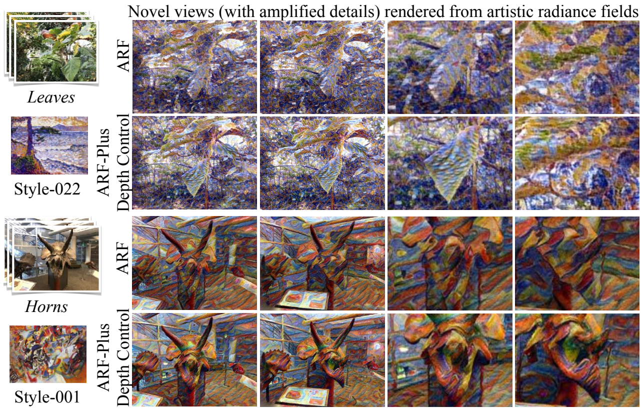

Figure 7: Qualitative results of depth-aware control. Our depth-aware control demonstrates superior performance in preserving perceptive depth across varying views - the forefront leaf in *Leaves*, and the nostril (hole) in *Horns*. Please refer to supplementary materials for better visualization.

4.6 Combination of Controls

In Figure 8, the 3D scene has successfully been infused with distinct shapes from two different styles. Through spatial control, the semantic object (horse) has been adorned with regular cuboid shapes transferred by Style-007, while the rest of the background showcases triangular patterns from Style-131. It avoids monotony by incorporating multiple shapes from different styles. This imbues the entire scene with an abstract sensation, particularly noticeable in elements like the leaves of trees. Furthermore, in the ARF baseline, the shape of the horse sculpture appeared disproportionately large, giving the impression of significant fractures in the sculpture. With our scale control, the patterns on the horse sculpture have been reduced to a more suitable size, enhancing visual aesthetics. Moreover, the color preservation feature allows us to freely select style images without excessive concern for whether their colors match our scene. In conclusion, compared to ARF, our control methods provide greater flexibility for 3D scene stylization, opening up endless possibilities for the imagination.

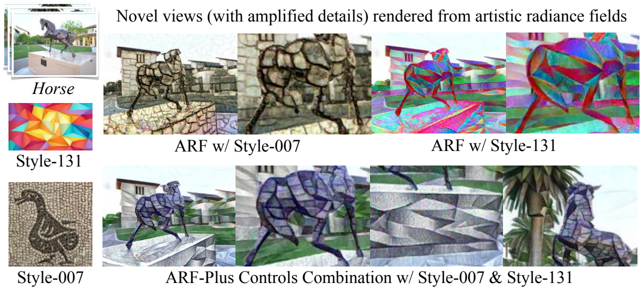
Novel views (with amplified details) rendered from artistic radiance fields
Horse
Style-131
Style-007
ARF w/ Style-007
ARF w/ Style-131
ARF-Plus Controls Combination w/ Style-007 & Style-131

Figure 8: Qualitative results of controls combination - Multiple Styles Spatial Control (with Style-007 on the horse & Style-131 on the background), Scale Control (with patterns scale-down on the horse), and Color Preservation Control (on the whole scene). Please refer to supplementary materials for better visualization.

5 Summary

In this paper, we identify a research gap - the lack of sufficient perceptual control - in the emerging radiance fields style transfer. Both quantitative and qualitative evaluations on real-world datasets demonstrate that ARF-Plus's four types of control methods effectively achieve perceptual control for 3D scene stylization. The combination use of our proposed controls through the ARF-Plus framework offers boundless freedom, enabling the production of novel and eye-catching stylized 3D scene effects. We wish that our research outcomes will encourage further studies in this novel and intriguing field.

The limitations of our work include: 1) Our spatial control method, while generally producing the desired stylization effects, sometimes demonstrates an issue of style leakage. This occurs because the accuracy of the semantic segmentation output differs among views, resulting in instances where the incomplete semantic mask fails to correctly represent object boundaries. 2) Since different styles have varying aesthetic characteristics, in order to achieve a visually appealing visual outcome, our scale control sometimes needs manual tuning of the hyperparameters. For example, determining the weight configuration for different convolution layers to emphasize finer details or coarser patterns. For a given 3D scene, it would be desirable if the system could automatically learn and tune the hyperparameters for each individual style in a more intelligent manner.

References

Castillo, C.; De, S.; Han, X.; Singh, B.; Yadav, A. K.; and Goldstein, T. 2017. Son of zorn's lemma: Targeted style transfer using instance-aware semantic segmentation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1348–1352. IEEE.

Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; and Su, H. 2022a. Tensorf: Tensorial radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, 333–350. Springer.

Chen, Y.; Yuan, Q.; Li, Z.; Xie, Y. L. W. W. C.; Wen, X.; and Yu, Q. 2022b. UPST-NeRF: Universal Photorealistic Style Transfer of Neural Radiance Fields for 3D Scene. *arXiv preprint arXiv:2208.07059*.

Chiang, P.-Z.; Tsai, M.-S.; Tseng, H.-Y.; Lai, W.-S.; and Chiu, W.-C. 2022. Stylizing 3d scene via implicit representation and hypernetwork. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1475–1484.

Fan, Z.; Jiang, Y.; Wang, P.; Gong, X.; Xu, D.; and Wang, Z. 2022. Unified implicit neural stylization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, 636–654. Springer.

Gatys, L. A.; Bethge, M.; Hertzmann, A.; and Shechtman, E. 2016. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Gatys, L. A.; Ecker, A. S.; Bethge, M.; Hertzmann, A.; and Shechtman, E. 2017. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3985–3993.

Höllein, L.; Johnson, J.; and Nießner, M. 2022. Stylemesh: Style transfer for indoor 3d scene reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6198–6208.

Huang, H.-P.; Tseng, H.-Y.; Saini, S.; Singh, M.; and Yang, M.-H. 2021. Learning to stylize novel views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13869–13878.

Huang, Y.-H.; He, Y.; Yuan, Y.-J.; Lai, Y.-K.; and Gao, L. 2022. 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*, 18342–18352.

Jing, Y.; Liu, Y.; Yang, Y.; Feng, Z.; Yu, Y.; Tao, D.; and Song, M. 2018. Stroke controllable fast style transfer with adaptive receptive fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 238–254.

Jing, Y.; Yang, Y.; Feng, Z.; Ye, J.; Yu, Y.; and Song, M. 2019. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11): 3365–3385.

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Liu, X.-C.; Cheng, M.-M.; Lai, Y.-K.; and Rosin, P. L. 2017. Depth-aware neural style transfer. In *Proceedings of the symposium on non-photorealistic animation and rendering*, 1–10.

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4): 1–14.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Mu, F.; Wang, J.; Wu, Y.; and Li, Y. 2022. 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*, 16273–16282.

Nguyen-Phuoc, T.; Liu, F.; and Xiao, L. 2022. Snerf: stylized neural implicit representations for 3d scenes. *arXiv preprint arXiv:2207.02363*.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3): 1623–1637.

Reimann, M.; Buchheim, B.; Semmo, A.; Döllner, J.; and Trapp, M. 2022. Controlling strokes in fast neural style transfer using content transforms. *The Visual Computer*, 1–15.

Ruder, M.; Dosovitskiy, A.; and Brox, T. 2018. Artistic style transfer for videos and spherical images. *International Journal of Computer Vision*, 126(11): 1199–1219.

Wang, C.; Jiang, R.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2023. Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics*.

Wright, M.; and Ommer, B. 2022. Artfid: Quantitative evaluation of neural style transfer. In *DAGM German Conference on Pattern Recognition*, 560–576. Springer.

Yin, K.; Gao, J.; Shugrina, M.; Khamis, S.; and Fidler, S. 2021. 3dstylenet: Creating 3d shapes with geometric and texture style variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12456–12465.

Yu, A.; Li, R.; Tancik, M.; Li, H.; Ng, R.; and Kanazawa, A. 2021. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5752–5761.

Yu, Z.; Wu, Y.; and Wang, T. 2019. A Method for Arbitrary Instance Style Transfer. *arXiv preprint arXiv:1912.06347*.

Zhang, K.; Kolkin, N.; Bi, S.; Luan, F.; Xu, Z.; Shechtman, E.; and Snavely, N. 2022. ARF: Artistic Radiance Fields. In *European Conference on Computer Vision*, 717–733. Springer.

Zhang, Y.; He, Z.; Xing, J.; Yao, X.; and Jia, J. 2023. 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*, 4242–4251.