

## A. OVERVIEW

In this appendix, we include more techniques, evaluation details, limitations and future work discussion. For better visualization, we encourage readers to watch the video results in the supplementary material folder.

## B. TECHNIQUE DETAILS

**Framework General Settings:** Identical to the baseline ARF [5], in ARF-Plus, we configure the stylization optimization to run for 10 epochs, with a learning rate that exponentially decreases from  $1e-1$  to  $1e-2$ . The scalar weight of style loss is set as  $\alpha = 1$ , and the weight of total variation loss is set as  $\gamma = 1$ . Unless stated otherwise, we assign a scalar weight of 0.001 to the content loss for forward-facing scenes and 0.005 for 360-degree scenes, which aligns with the baseline ARF.

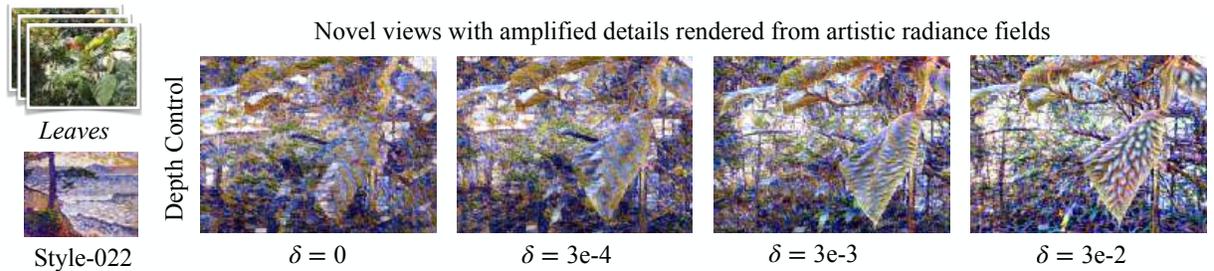
**Color Preservation Control Settings:** In our color preservation control, the style loss is calculated based only on the luminance channel. Due to the presence of only a single luminance channel for computing style loss, the value of style loss becomes diminished. However, for content loss, the calculation is still performed using all three RGB channels. As a result, to achieve balance, the content loss weight, denoted as  $\beta$  in Eq. 1, should be decreased accordingly. Specifically, we set  $\beta = 0.0001$  for forward-facing scenes and  $\beta = 0.0005$  for 360-degree scenes.

**Scale Control Settings:** All parameters for the scale control remain consistent with the baseline ARF.

**Spatial Control Settings:** We utilize a pre-trained fully-convolutional network model with a ResNet-101 backbone [27] to obtain the semantic segmentation spatial mask. All other settings remain consistent with the baseline ARF.

**Depth Enhancement Control Settings:** The pre-trained MiDaS network [23] is used as depth estimation network  $\phi_1$ . In ARF-Plus, We set the depth weight  $\delta = 0.003$  (in Eq. 10). All other settings remain consistent with the baseline ARF.

**Sensitiveness & Comparisons & Validation & Control issue w.r.t. hyperparams:** As mentioned above, hyperparams  $\alpha$ ,  $\beta$ , and  $\gamma$  align with those of the baseline ARF. Compared to ARF, our color preservation and spatial control have not introduced any additional hyperparams. Regarding depth control, a single hyperparameter  $\delta$  is introduced to control depth perception loss weight. The magnitude of its value depends on how realistic the user wants the scene to appear in perceptive depth, as shown in Fig. 9. For scale control, the adjustment and sensitiveness of scale hyperparams  $w_l$  and  $\mathcal{T}_S^l$  depends on the differences (e.g. in size) between the depicted content in the style image and the 3D scene, as well as people’s aesthetic preferences. In Appendix C.3, Fig. 11, Fig. 12, and Fig. 13 present comparisons.



**Fig. 9:**  $\delta$  positively affects the perceived visual depth in stylized scenes, but excessive values can imbalance total loss in Eq. 10, resulting in less aesthetically pleasing results.

## C. EVALUATION

In Section C.1, we detail the evaluation methods used in the main paper (Section 4) and the supplementary parts (Section C.2 - C.5). In Section C.2, additional experimental outcomes with color control are encompassed, including both qualitative and quantitative analyses. In Section C.3, we showcase influential factor studies concerning the receptive field and style image size. Additionally, we demonstrate further research on combining receptive field weights and utilizing various style image sizes. In Section C.4, we present additional quantitative experimental outcomes and intermediate training results. Additionally, we elucidate the procedure for computing ArtFid within the specified regions. In Section C.5, supplementary qualitative and quantitative results of depth enhancement control are provided.

## C.1. Evaluation Methods

We assess our approach’s effectiveness by conducting quantitative and qualitative comparisons with the baseline method ARF. We prioritize qualitative evaluation as the primary method in our study, while quantitative evaluation serves as a supplementary measure. This decision is motivated by two main factors. Firstly, both our baseline and most style transfer approaches primarily emphasize qualitative comparisons rather than quantitative metrics. Secondly, existing quantitative metrics have limitations in accurately capturing human visual perception.

### C.1.1. Qualitative Evaluation Method:

It is worth mentioning that there is always no single correct output in the field of neural style transfer. Currently, the most popular approach for evaluating neural style transfer models is a qualitative comparison of style and content views. To assess the qualitative performance, we present visual comparisons between ARF and our ARF-Plus (with different perceptual controls) for both forward-facing and 360-degree scenes. Because the results of our perceptual controls have a clear visual impact on the outputs and do not involve subjective aesthetic judgments, we do not conduct a user study to further evaluate the method.

### C.1.2. Quantitative Evaluation Method:

As for quantitative evaluation, some works [28, 29] employ classical perceptual metrics such as PSNR or SSIM [30], however, these metrics are generally not consistent with human perception [31]. Recently, Wright et al. [26] proposes a quantitative metric by combining two factors: 1. ContentDist: content preservation between generated images  $X_g$  and content images  $X_c$ , and 2. StyleFID: style feature distributions difference between generated images  $X_g$  and style images  $X_s$ . We utilize this metric for quantitative evaluation.

$$\begin{aligned} ArtFid(X_g, X_c, X_s) &= (1 + ContentDist) \cdot (1 + StyleFID) \\ ContentDist &= \frac{1}{N} \sum_{i=1}^N d(X_c^{(i)}, X_g^{(i)}) \\ StyleFID &= FID(X_s, X_g) \end{aligned} \quad (11)$$

We adopt this measurement metric from the 2D images domain and apply it to our research. To mitigate the impact of photo-realistic radiance fields reconstruction performance, we utilize generated photo-realistic views as  $X_c$  instead of using training photos. We choose a number of  $N$  viewpoints, render the photo-realistic radiance field, and generate multiple photo-realistic scene views  $\mathbf{X}_{real}$  before stylization, obtaining  $X_c = \{\mathbf{X}_{real}^{(1)}, \mathbf{X}_{real}^{(2)}, \dots, \mathbf{X}_{real}^{(N)}\}$ . After stylization, for each stylized radiance field, we render a series of images from the same viewpoints, which produces  $X_g = \{\hat{\mathbf{X}}^{(1)}, \hat{\mathbf{X}}^{(2)}, \dots, \hat{\mathbf{X}}^{(N)}\}$ . In order to generate  $X_s$ , we replicate the provided style image  $\mathbf{X}_{style}$  with  $N$  times, which gives us  $X_s = \{\mathbf{X}_{style}, \mathbf{X}_{style}, \dots, \mathbf{X}_{style}\}$ . For each forward-facing scene, we set the number of rendered novel views  $N$  to 120. For 360-degree scenes, where there are more angles to consider,  $N$  is increased to 200. The chosen novel view angles and the method employed for generating novel views remain consistent with the baseline ARF.

## C.2. Color Preservation Control

In order to enhance the substantiation of the essentiality and efficacy of our color preservation technique (employing a luminance-only approach), an additional preprocessing technique, namely color histogram matching, has been incorporated for comparative analysis. In the preprocessing color histogram matching step, the colors of the style image  $\mathbf{X}_{style}$  are adjusted to match the colors of the photo-realistic content view  $\mathbf{X}_{content}$  through an image to image color histogram matching. This creates a new style image  $\mathbf{X}'_{style}$  which is then used as a new style input, replacing the original  $\mathbf{X}_{style}$ . We utilize the linear method that Gatys et al. [9] used, which has been proven to produce effective outcomes in transferring image color. Unlike 2D image style transfer with just one content image, we have photos from various viewpoints as training data. When computing the style loss  $\mathcal{L}_{style}$  for each viewpoint, a new style image  $\mathbf{X}'_{style}$  is generated corresponding to the current training photo  $\mathbf{X}_{content}$ .

### C.2.1. Qualitative Evaluation Results

Fig. 10 demonstrates visual comparisons between methods applied to real-world forward-facing scenes - *Leaves*, *Flower*, *Trex*, and *Horns*. Fig. 10 (a) demonstrates the rendered views from photo-realistic fields. In the context of color preservation

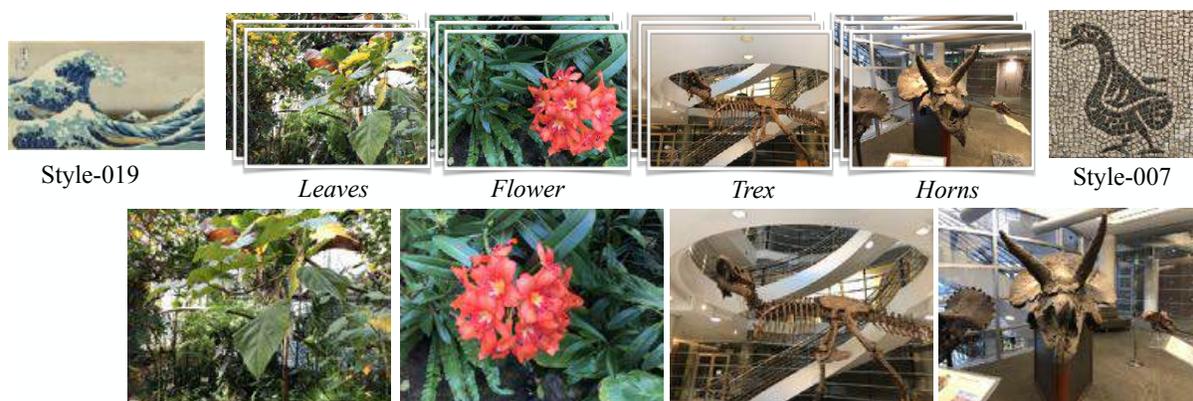
control, our goal is to ensure that the colors of the stylized scene remain consistent with those of the photo-realistic radiance fields. As depicted in Fig. 10 (b)-(c), our proposed color control method, effectively preserves the original scene’s color while successfully learning the desired style. In terms of details, our color preservation algorithm produces superior results compared to the preprocessing color histogram matching approach. For the scene *Flower*, red color leakage occurs on color histogram matching, resulting in the appearance of red dots on some of the leaves. Our color preservation method does not exhibit obvious color leakage. Moreover, the colors and brightness of the stylized outputs achieved through our proposed method closely resemble those of the photo-realistic view.

### C.2.2. Quantitative Evaluation Results

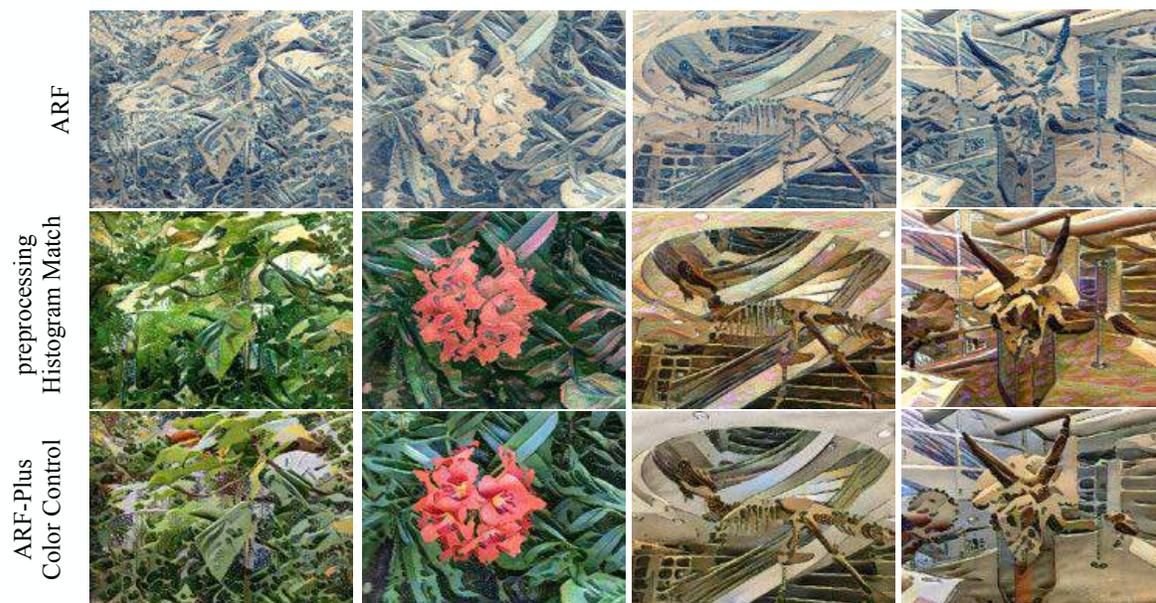
Table 2 displays the quantitative results on forward-facing data. Our color control methods have lower ContentDist scores than the baseline ARF, suggesting that the stylized scenes more closely resemble the photo-realistic scenes due to color preservation. Our color preservation method demonstrates superior performance on Style-019. However, for Style-007, the difference in effectiveness between our method and the preprocessing color histogram matching method is not significant. An intriguing observation is that when utilizing Style-007, the ContentDist values for color histogram matching on *Leaves* and *Trex* are smaller compared to the values obtained from our color preservation method. This may be attributed to the fact that color linear matching tends to yield satisfactory results when both the content and style images have relatively simple color compositions and distributions. The *Leaves* is mainly composed of green, while the *Trex* are primarily composed of dark brown. Similarly, the style image Style-007 has a relatively simple color composition with brown being the dominant color. As a result, the new style image generated by color histogram matching effectively integrates the colors of the scene. Another interesting finding is that our color preservation methods result in better StyleFID scores on Style-019, which means the transferred style is more similar to the given style image. We do not have a very confident explanation for this. One possible explanation is that the inconsistency between human judgements and machine evaluation metrics is inevitable. The StyleFID [26] can only roughly judge the style similarity, but its algorithm cannot accurately simulate the human brain’s understanding of style and content.

**Table 2:** Quantitative results of color preservation control: style transfer on forward-facing scenes. Results superior to the baseline ARF are highlighted in bold. The best results are in **bold**, and the second best results are in **blue**.

Style	Scene	ArtFID ↓			ContentDist ↓			StyleFID ↓		
		ARF	Preprocess Hist.match	ARF-Plus w/ Color Control	ARF	Preprocess Hist.match	ARF-Plus w/ Color Control	ARF	Preprocess Hist.match	ARF-Plus w/ Color Control
	Leaves	49.6562	47.6273	<b>38.0664</b>	0.5424	0.4416	<b>0.3817</b>	31.1951	32.0386	<b>26.5494</b>
	Flowers	49.8151	43.9636	<b>39.0186</b>	0.5893	0.5393	<b>0.3766</b>	30.3443	27.5606	<b>27.3451</b>
	Trex	47.4958	48.0271	<b>42.4836</b>	0.5644	0.5509	<b>0.4562</b>	29.3610	29.9681	<b>28.1745</b>
	Horns	45.5077	44.1148	<b>41.3908</b>	0.4889	0.4865	<b>0.4256</b>	29.5641	28.6767	<b>28.0330</b>
	Leaves	43.6201	51.6565	<b>41.7606</b>	0.4978	<b>0.4000</b>	0.4034	<b>28.1233</b>	35.8971	28.7565
	Flowers	<b>40.2618</b>	52.6178	46.8613	0.5552	0.4758	<b>0.3963</b>	<b>24.8884</b>	34.6547	32.5600
	Trex	<b>36.2511</b>	38.4576	46.2475	0.5386	<b>0.5053</b>	0.5212	<b>22.5606</b>	24.5483	29.4028
	Horns	<b>38.3892</b>	57.3361	51.5242	0.5088	0.5067	<b>0.5033</b>	<b>24.4442</b>	37.0529	33.2752



(a) Novel views rendered from the reconstructed photo-realistic radiance fields



(b) Novel views rendered from artistic radiance fields in style-019



(c) Novel views rendered from artistic radiance fields in style-007

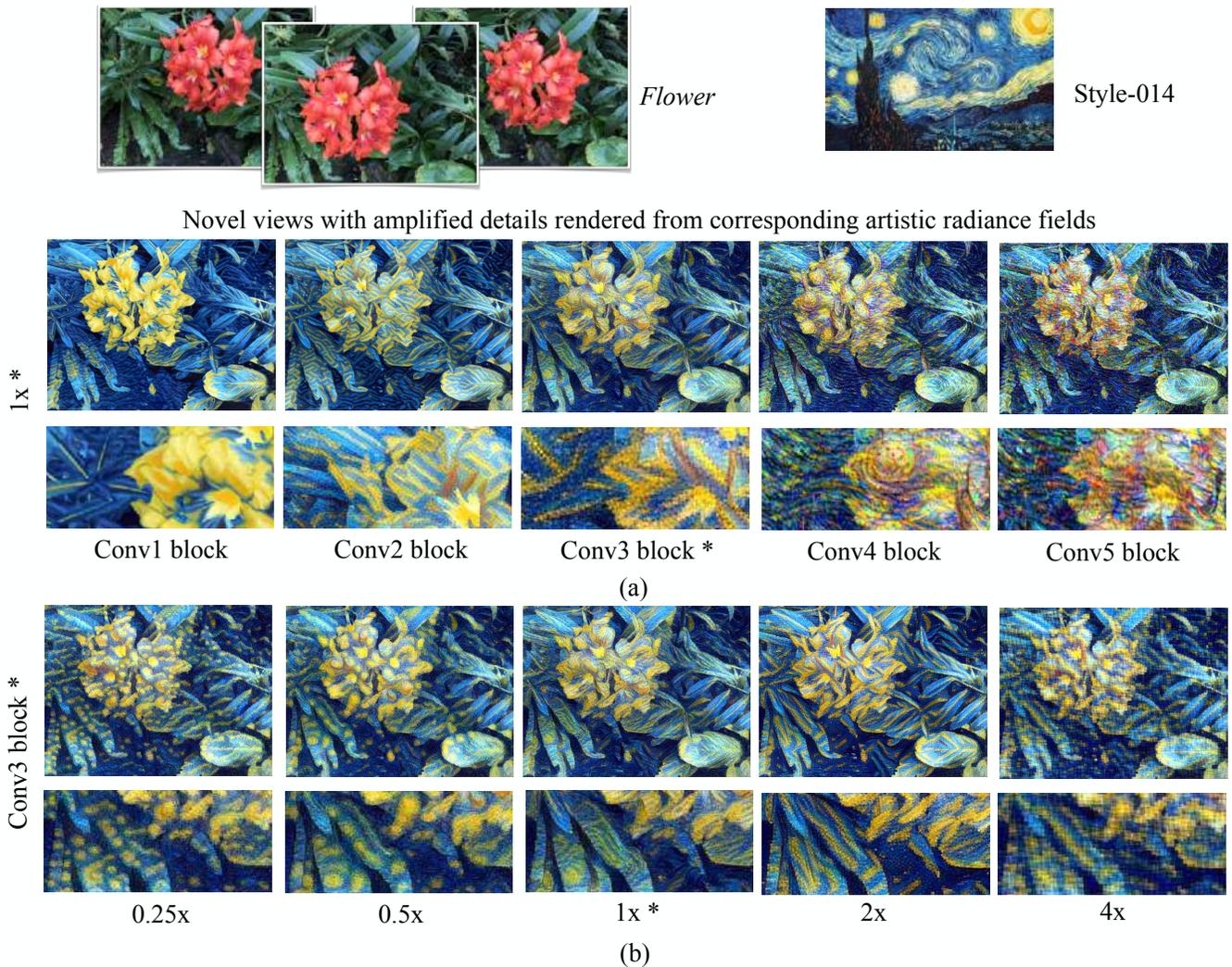
**Fig. 10:** Qualitative results of color preservation control: style transfer on forward-facing scenes *Leaves*, *Flower*, *Trex* and *Horns*. Our ARF-Plus with color preservation control successfully preserves the original colors while effectively transferring the style patterns. Compared to the preprocessing color histogram matching method which generates a new style image, our color preservation method yields better results.

### C.3. Scale Control

For scale control, we only apply qualitative evaluation. This is because ArtFID and its two components ContentDist and StyleFID, are unable to accurately measure the scaling of style patterns [26]. The scaling of style patterns can lead to variations in both the 3D scene’s style patterns and content appearance. As a result, both ContentDist and StyleFID, which are quantitative numerical metrics, undergo changes. However, these value changes are not directly proportional to the degree of style pattern size magnification or reduction. In other words, ArtFID cannot indicate whether the style pattern is being scaled down or scaled up.

#### C.3.1. Receptive Field and Style Image Size:

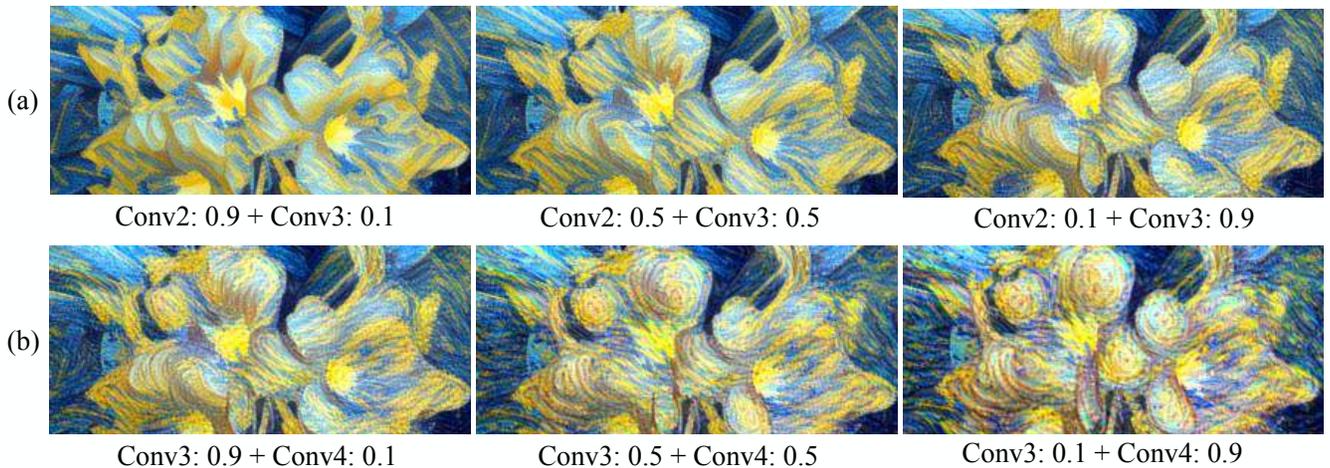
As our method involves the consideration of both the receptive field (selected convolution layers) and style image size, we add additional influential factor studies experiments to assess the impact of each factor on the style scaling independently. Fig. 11 shows the rendered views derived from 10 stylized (artistic) radiance fields generated with a variety of settings, including changes to the receptive field and alterations to the size of the style image. Each stylized (artistic) radiance fields corresponds to a stylized 3D scene.



**Fig. 11:** The impact of receptive field and style image size on style scaling. (a) Selecting different Conv layer blocks with the fixed style image size 1x. (b) Resizing style image with the fixed Conv layer block. Settings that are identical to the baseline ARF are marked with the asterisk (\*).

In Fig. 11 (a), we modify the receptive field by selecting different convolution layer blocks while keeping the original style image size. As we delve deeper into the VGG network, the receptive field undergoes a rapid expansion. In the ultimate layer, it encompasses a substantial portion of the input image. This phenomenon has implications for the stylized radiance fields, as it leads to an overall reduction in style pattern scale and incorporation of a greater amount of coarse styles. For instance, when choosing the Conv4 block, the stylized output exhibits an increased presence of circular petal-like forms, which are similar to the yellow moon style in Style-014. Conversely, with the Conv2 block, the emphasis in the output is more on fine details, reminiscent of brush strokes in Style-014. Moreover, the result of the Conv2 block presents a larger scale in terms of the style pattern. In Fig. 11 (b), we keep a fixed receptive field (similar to the one selected in the baseline, which is conv3) while altering the size of the style image. The results reveal that changes in the style image size also impact the resulting style pattern scale. When the style image is resized to 0.25x, the small circular spots on the leaves in the stylized radiance fields correspond to the moon pattern from Style-014. However, when the style image is resized to 0.5x, the circular spots on the leaves also increase in size accordingly. Based on the results shown in Fig. 11, both the receptive field and style image size have an impact on the pattern scale in the stylized 3D scenes. The receptive field serves as a control mechanism for learning fine or coarse patterns from the style image and also influences the size of the patterns. On the other hand, the style image size assists in adjusting the overall scale of the style.

Novel views (amplified in details) rendered from corresponding artistic radiance fields



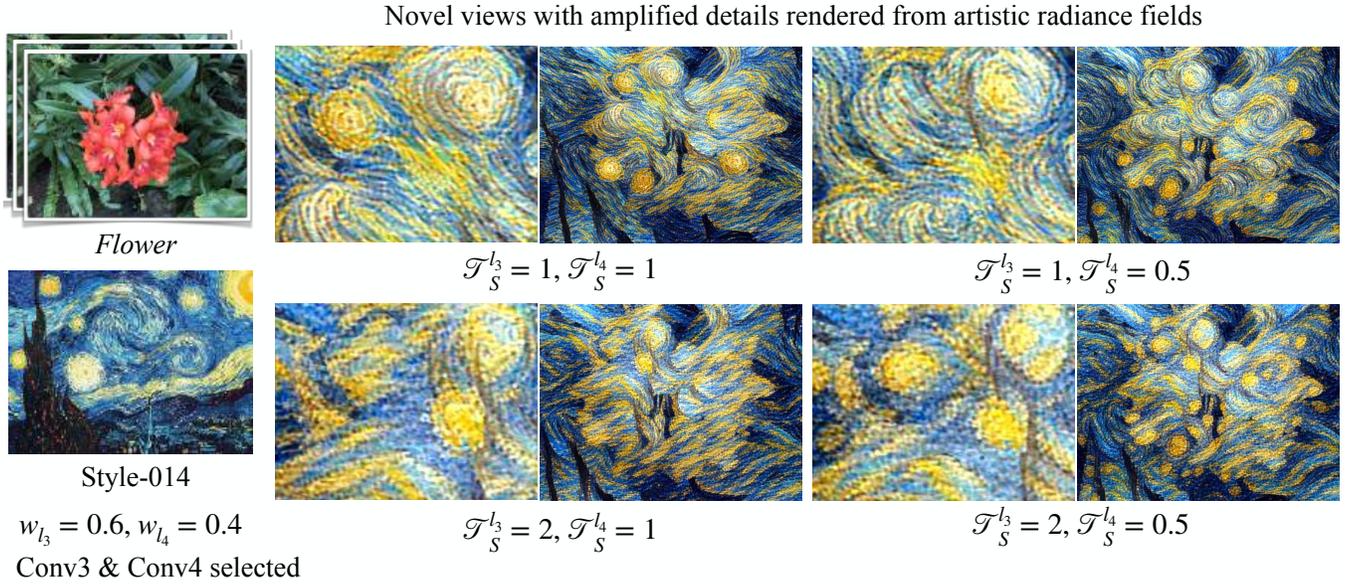
**Fig. 12:** Weighting across different receptive fields (convolution layers) achieves continuous and diversified scale control. The adjustment of weights can result in a continuous scaling of the style patterns on the petals.

### C.3.2. Combination of Receptive Field Weights:

We proceed to demonstrate the role of weight combinations for different receptive fields (Eq. 2). This enables the achievement of continuous style scaling control. Fig. 12 depicts the rendered novel views corresponding to six artistic (stylized) radiance fields, generated using combinations of receptive fields with varying weights. Each artistic radiance fields portrays a stylized 3D scene. In Fig. 12 (a), as the weights of Conv3 gradually increase from 0.1 to 0.9, while the weights of Conv2 decrease from 0.9 to 0.1, the style texture on the petals becomes progressively finer, scaling down. Similar observations can be made for the results in 12 (b). Increasing the weights of Conv4, which has a broader receptive field, also leads to style scaling down. Furthermore, it enhances the representation of coarse style patterns (circular shapes) present in the style image. In conclusion, the weight combinations for different receptive fields in our method are important, as they offer the advantages of flexible and continuous control over the style pattern scale.

### C.3.3. Combination of Style Image Sizes:

As shown in Fig. 11, when using the same receptive field for different style image sizes, distinct scale effects emerge. It is easy to infer that varying combinations of receptive fields with different style image sizes will offer increased flexibility. Fig. 13 provides further illustrations.



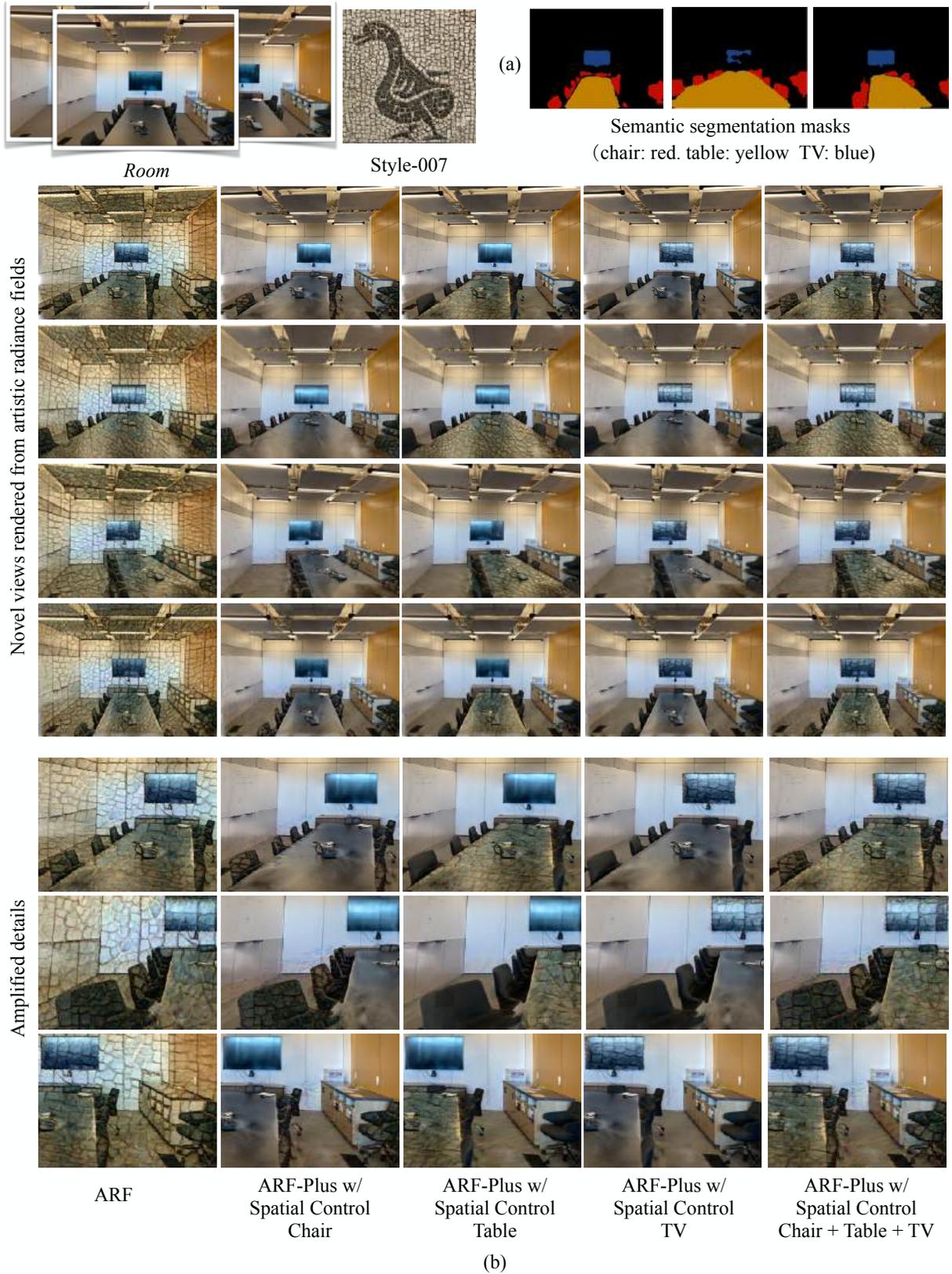
**Fig. 13:** Assigning selected receptive fields (conv3 and conv4) with different  $T_S^l$  can increase flexibility. The coarse pattern (moon-like circular shape) is shrunk while the width of stroke lines is scaled up.

#### C.4. Spatial Control

##### C.4.1. Qualitative Evaluation Results

Fig. 14 showcases additional rendered novel views of the scene *Room* under the spatial control via semantic segmentation masks. More experiment results of the scene *Horse* on various styles are shown in Fig. 15, which demonstrate that our spatial control approach exhibits remarkable performance across diverse styles. This is evident in the stylization of semantic objects within novel views rendered from various perspectives.

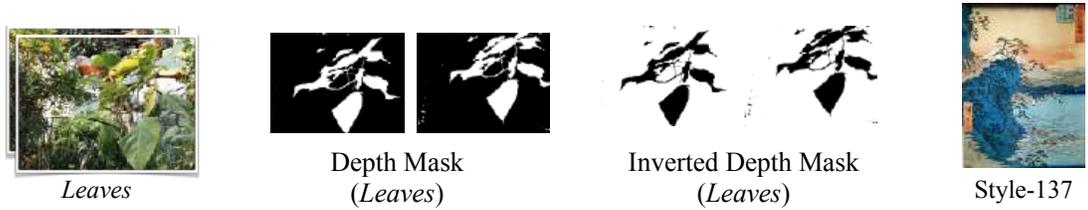
As the source of our spatial control masks can be derived from various segmentation methods, we observed that, for forward-facing scenes, employing binary segmented depth maps is also effective in identifying significant object regions. This is based on the assumption that, within the user’s field of view, the main objects that grab their interest typically have less depth than the surroundings. Since the reconstructed photo-realistic radiance fields (used as the stylization input) contains depth information for each view, the depth map can be directly generated. Depth spatial masks are generated by applying Otsu’s binary segmentation to select the optimal threshold. Fig. 16 presents the outcomes of spatial control using depth segmentation masks. The results further underscore the efficacy of our spatial control approach, as evidenced by the proficient stylization of selected regions, irrespective of whether employing segmentation masks or their inverted counterparts. In both scenarios, effective stylization can be applied to either the chosen background regions or the object (foreground) regions.



**Fig. 14:** Qualitative results of spatial control with semantic segmentation masks: style transfer on the forward-facing scene - *Room*. Our ARF-Plus with spatial control effectively stylizes specific semantic objects - chair, table, and TV - within the scene.



**Fig. 15:** Qualitative results of spatial control with semantic segmentation masks: style transfer on the 360-degree scene - *Horse*. Our ARF-Plus with spatial control exhibits superior performance in stylizing the designated semantic object - horse - within the scene.



Novel views rendered from artistic radiance fields with amplified details



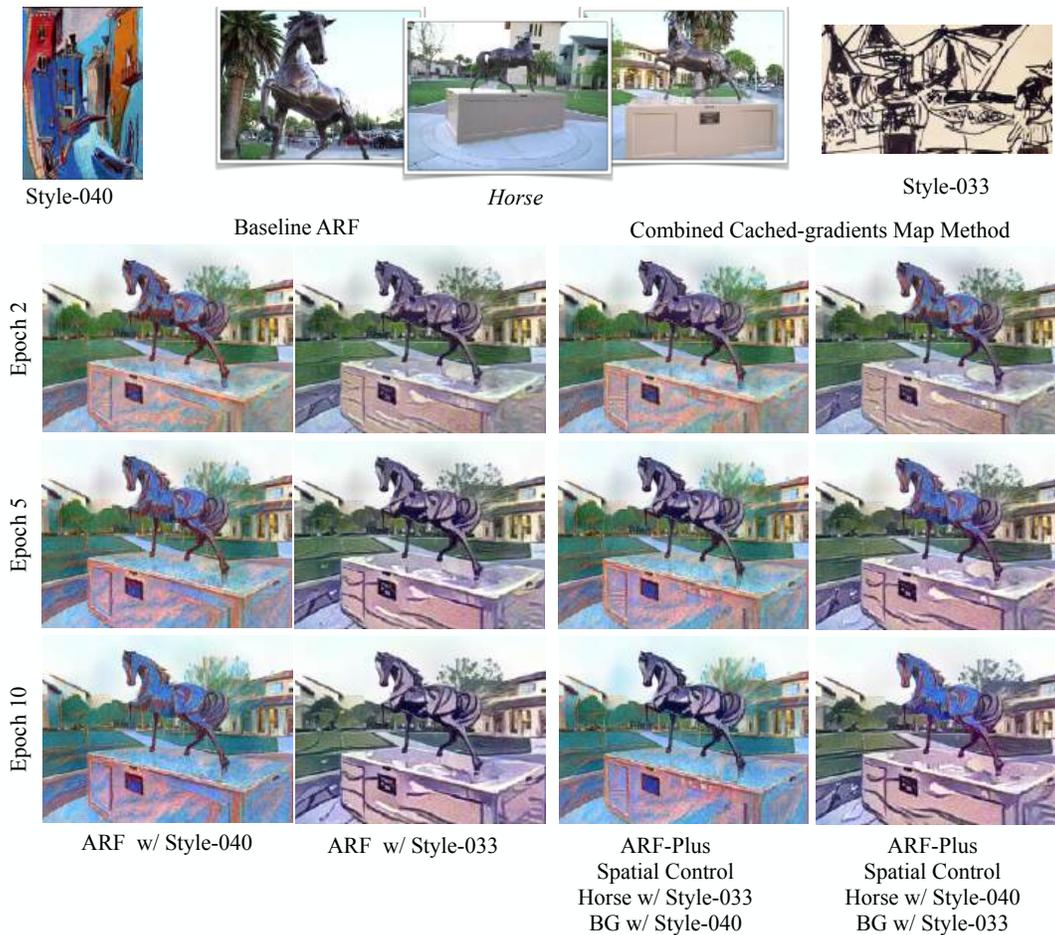
Novel views rendered from artistic radiance fields with amplified details



**Fig. 16:** Qualitative results of spatial control with depth segmentation spatial masks: style transfer on forward-facing scenes - *Leaves* and *orchids*. Our ARF-Plus with spatial control effectively stylizes certain spatial areas according to spatial masks.

### C.4.2. Intermediate Training Results

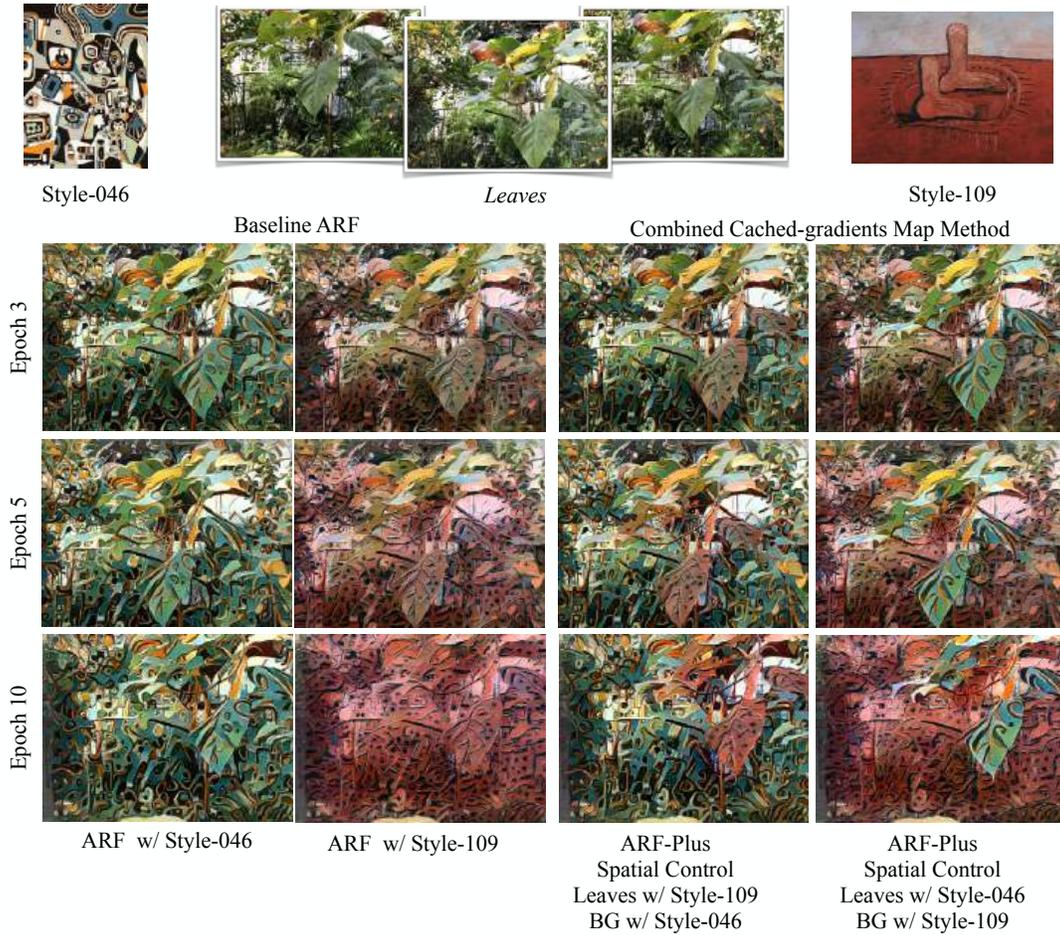
In Fig. 17 and Fig. 18, the intermediate results of multiple-style spatial control are demonstrated. The training epoch number of this method is identical to the baseline ARF. It shows that our multiple styles spatial control method - Combined Cached-gradients Map - simultaneously optimizes and updates gradients in different regions, regardless of whether the spatial mask originates from semantic segmentation or depth map segmentation. The style of each selected region changes at every epoch.



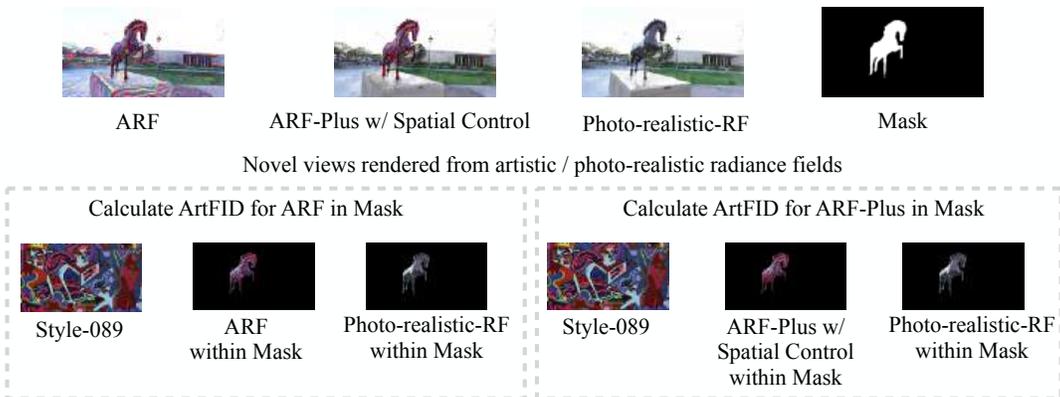
**Fig. 17:** Multiple styles spatial control (semantic segmentation mask) with Combined Cached-gradients Map: intermediate validation results rendered from fixed viewpoints during the training process.

### C.4.3. ArtFID Mask-In Evaluation

Regarding quantitative assessment, we provide a more detailed explanation of ArtFID’s usage within specific regions as follows. The situation with spatial control is rather unique, as only certain regions within each rendered image are stylized. Directly applying ArtFID to the entire image would fail to capture the stylized details within a specific region. As illustrated in Fig. 19, we propose the Mask-In evaluation method, which is specifically designed to accommodate spatial control. It is important to note that the spatial mask used during training (to specify the selected area for stylization) cannot be utilized in evaluation. The generated views in the evaluation phase have novel perspectives, which are different from training viewpoints. The spatial masks for evaluation are generated from selected novel views rendered from photo-realistic radiance fields (photo-realistic-RF). Then spatial masks are then applied to the views of the photo-realistic radiance fields and the stylized radiance field (with the same selected viewpoints). The processed results are used as inputs  $X_c$  and  $X_g$  for ArtFID (Eq. 11). The ARF results within the mask provide specific indicators for the designated regions.



**Fig. 18:** Multiple styles spatial control (depth segmentation mask) with Combined Cached-gradients Map: intermediate validation results rendered from random viewpoints during the training process.

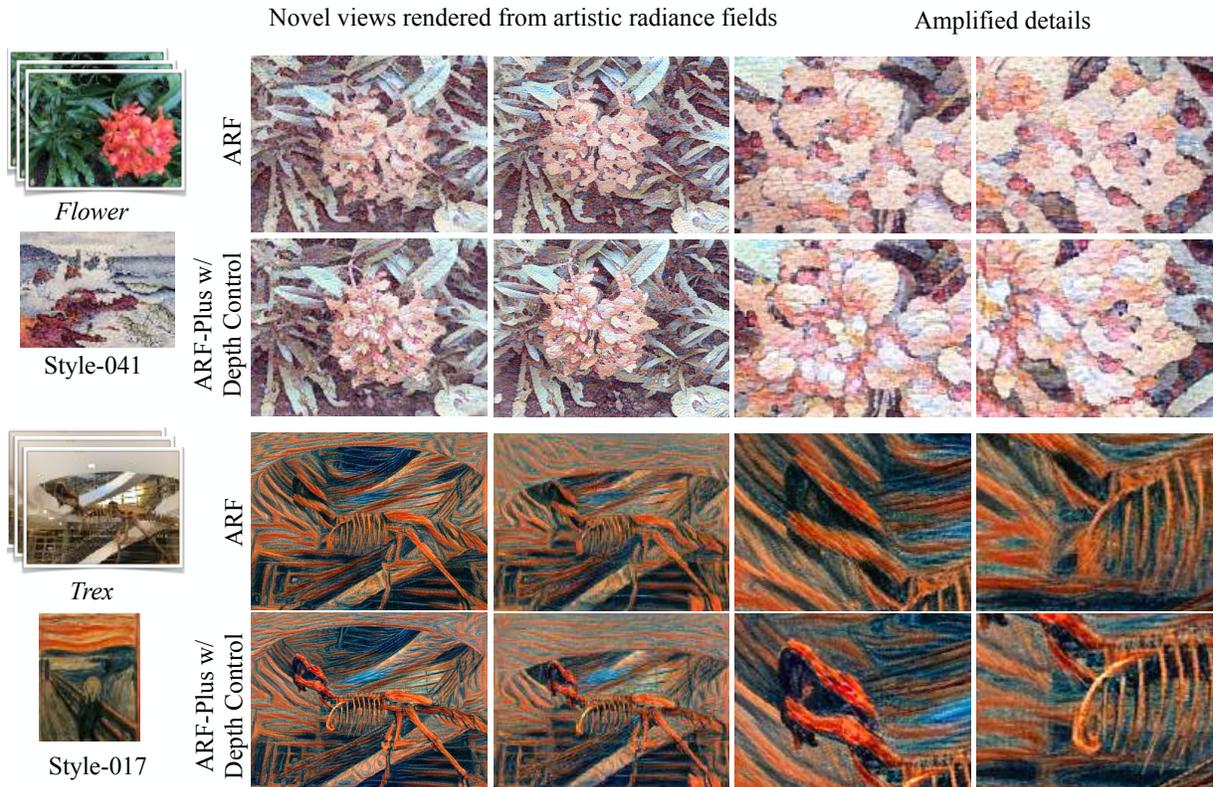


**Fig. 19:** Our proposed Mask-In evaluation method for ARF-Plus spatial control.

## C.5. Depth Enhancement Control

### C.5.1. Qualitative Evaluation Results

Fig. 20 provides visual contrasts between our depth enhancement control method and the baseline ARF. Qualitative findings substantiate our approach’s successful retention of perceptual depth across diverse viewpoints. The flower’s petals in the *Flower* scene generated by our depth control method better retain the depth of the scenes. The petals of the entire flower contain more perceptible depth. After applying our depth enhancement control method, the skeleton outline of the trex in *Trex* is more prominent, and the depth details of the head are also more pronounced.



**Fig. 20:** Qualitative results of depth enhancement control: style transfer on the forward-facing scenes - *Flower* and *Trex*. Our method, ARF-Plus with depth enhancement control, demonstrates superior performance in preserving perceptible depth across varying views - the petals in *Flower*, and the head and skeleton in *Trex*.

### C.5.2. Quantitative Evaluation Results

Table 3 displays the quantitative results on the forward-facing scenes. Due to the preservation of perceptual depth, the stylized scenes appear to be more similar to the original photo-realistic views than the baseline, as seen by our depth enhancement control method’s lower ContentDist scores. One interesting finding is that our depth-aware control method always generates better StyleFID and ArtFid scores on Style-017 and Style-022. This suggests that the depth control result in a smaller distance between feature distributions on the style image and the stylized scene views. We note this issue as future work and hope to find an answer through collaborative research with peers in the field.

**Table 3:** Quantitative results of depth enhancement control: style transfer on the forward-facing scenes. ContentDist ↓ measures content preservation. StyleFID ↓ measures style matching. ArtFID ↓ [26] combines the metrics ContentDist and StyleFID to evaluate the overall visual effect of style transfer. The best results are in **bold**.

Style	Scene	ArtFID ↓		ContentDist ↓		StyleFID ↓	
		ARF	ARF-Plus w/ Depth Control	ARF	ARF-Plus w/ Depth Control	ARF	ARF-Plus w/ Depth Control
Style-001 	Leaves	<b>40.7593</b>	46.1252	0.5609	<b>0.5557</b>	<b>25.1122</b>	28.6497
	Flower	<b>46.1371</b>	46.8988	0.5695	<b>0.5534</b>	<b>28.3964</b>	29.1908
	Trex	<b>46.7243</b>	52.0687	0.6465	<b>0.6114</b>	<b>27.3781</b>	31.3131
	Horns	<b>46.8514</b>	48.5809	0.6169	<b>0.5631</b>	<b>27.9754</b>	30.0794
Style-017 	Leaves	32.8128	<b>31.4048</b>	0.6294	<b>0.6044</b>	19.1385	<b>18.5738</b>
	Flowers	35.0658	<b>31.2184</b>	0.6166	<b>0.5892</b>	20.6908	<b>18.6444</b>
	Trex	31.0811	<b>27.8044</b>	0.6817	<b>0.6506</b>	17.4823	<b>15.8447</b>
	Horns	31.6769	<b>29.0361</b>	0.5938	<b>0.5678</b>	18.8755	<b>17.5198</b>
Style-022 	Leaves	36.6079	<b>32.7389</b>	0.6208	<b>0.5699</b>	21.5859	<b>19.8536</b>
	Flower	46.7435	<b>41.1923</b>	0.6853	<b>0.6473</b>	26.7367	<b>24.0061</b>
	Trex	39.9837	<b>38.3577</b>	0.7348	<b>1.5834</b>	22.0485	<b>21.7852</b>
	Horns	35.4226	<b>34.5354</b>	0.6606	<b>0.6154</b>	<b>20.3306</b>	20.3795
Style-041 	Leaves	48.4786	<b>48.2329</b>	0.6318	<b>0.6045</b>	<b>28.7086</b>	29.0618
	Flowers	<b>59.2806</b>	60.9881	0.6908	<b>0.6706</b>	<b>34.0610</b>	35.5072
	Trex	51.5526	<b>49.8702</b>	0.6425	<b>0.6299</b>	30.3871	<b>29.5977</b>
	Horns	<b>46.7634</b>	49.9683	0.5992	<b>0.5918</b>	<b>28.2413</b>	30.3908

## D. LIMITATIONS AND FUTURE WORK

### D.1. Limitations

The limitations of our work include: 1) Due to limited time and computational resources, we primarily conducted experiments on Plenoxels [32], which is known for its faster optimization (e.g., training the same 3D scene where Plenoxel takes only 11 minutes while NeRF [6] requires 1 day). 2) We did not explore other semantic segmentation or depth prediction models to potentially improve the performance of spatial control and depth-aware control. The reason is that our primary focus lies in validating the fundamental principles underlying our control methods. 3) We did not extensively experiment or conduct in-depth analysis on the combination of multiple controls and the use of multiple styles. This is because of our desire to prioritize the analysis of each method’s effectiveness. Conducting experiments specifically on a single style allows for a better understanding of the strengths and limitations of each control method.

### D.2. Future Work

A direct direction to explore is scaling control when dealing with multiple blended styles. Since different styles have varying strengths and characteristics, it would be more convenient and practical to have an algorithm that can automatically compute suitable parameters for different coarse or fine pattern scaling. Moreover, it is worth trying to build an interactive panel that can dynamically display a preview of the scaling effects on a 3D scene.

Another promising direction to explore is the integration of the-state-of-art techniques for scene understanding with our perceptual control methods. Currently, our ARF-Plus framework utilizes depth information and the positional information of semantic objects in the scene, but it does not take into account additional types of information such as scene composition, categories, and other aspects. By incorporating a more comprehensive range of scene information, we can enhance the perceptual control capabilities of our technology and further improve the quality of stylization.

## E. REFERENCES

- [1] Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, and Ming-Hsuan Yang, "Learning to stylize novel views," in *ICCV*, 2021.
- [2] Fangzhou Mu, Jian Wang, Yicheng Wu, and Yin Li, "3d photo stylization: Learning to generate stylized novel views from a single image," in *CVPR*, 2022.
- [3] Kangxue Yin, Jun Gao, Maria Shugrina, Sameh Khamis, and Sanja Fidler, "3dstylenet: Creating 3d shapes with geometric and texture style variations," in *ICCV*, 2021.
- [4] Lukas Höllein, Justin Johnson, and Matthias Nießner, "Stylemesh: Style transfer for indoor 3d scene reconstructions," in *CVPR*, 2022.
- [5] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely, "Arf: Artistic radiance fields," in *ECCV*, 2022.
- [6] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Communications of the ACM*, 2021.
- [7] 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 *WACV*, 2022.
- [8] 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 *CVPR*, 2022.
- [9] Leon A Gatys, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman, "Preserving color in neural artistic style transfer," in *arXiv preprint arXiv:1606.05897*, 2016.
- [10] 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 *ECCV*, 2018.
- [11] Max Reimann, Benito Buchheim, Amir Semmo, Jürgen Döllner, and Matthias Trapp, "Controlling strokes in fast neural style transfer using content transforms," in *The Visual Computer*, 2022.
- [12] Zhifeng Yu, Yusheng Wu, and Tianyou Wang, "A method for arbitrary instance style transfer," in *arXiv preprint arXiv:1912.06347*, 2019.
- [13] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman, "Controlling perceptual factors in neural style transfer," in *CVPR*, 2017.
- [14] Zhiwen Fan, Yifan Jiang, Peihao Wang, Xinyu Gong, Dejia Xu, and Zhangyang Wang, "Unified implicit neural stylization," in *ECCV*, 2022.
- [15] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao, "Snerf: stylized neural implicit representations for 3d scenes," in *arXiv preprint arXiv:2207.02363*, 2022.
- [16] Yaosen Chen, Qi Yuan, Zhiqiang Li, Yuegen Liu, Wei Wang, Chaoping Xie, Xuming Wen, and Qien Yu, "Upst-nerf: Universal photorealistic style transfer of neural radiance fields for 3d scene," in *arXiv preprint arXiv:2208.07059*, 2022.
- [17] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa, "Plenotrees for real-time rendering of neural radiance fields," in *ICCV*, 2021.
- [18] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su, "Tensorf: Tensorial radiance fields," in *ECCV*, 2022.
- [19] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao, "Nerf-art: Text-driven neural radiance fields stylization," in *Trans Vis Comput Graph*, 2023.
- [20] Yuechen Zhang, Zexin He, Jinbo Xing, Xufeng Yao, and Jiaya Jia, "Ref-npr: Reference-based non-photorealistic radiance fields for controllable scene stylization," in *CVPR*, 2023.
- [21] Xiao-Chang Liu, Ming-Ming Cheng, Yu-Kun Lai, and Paul L Rosin, "Depth-aware neural style transfer," in *NPAR*, 2017.
- [22] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song, "Neural style transfer: A review," in *Trans Vis Comput Graph*, 2019.
- [23] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," in *PAMI*, 2020.
- [24] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," in *TOG*, 2019.
- [25] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," in *TOG*, 2017.
- [26] Matthias Wright and Björn Ommer, "Artfid: Quantitative evaluation of neural style transfer," in *DAGM GCPR*, 2022.

- [27] Jonathan Long, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation,” in *CVPR*, 2015.
- [28] Xueting Liu, Wenliang Wu, Huisi Wu, and Zhenkun Wen, “Deep style transfer for line drawings,” in *AAAI*, 2021.
- [29] Kibeom Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, and Hyeran Byun, “Domain-aware universal style transfer,” in *ICCV*, 2021.
- [30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, “Image quality assessment: from error visibility to structural similarity,” in *TIP*, 2004.
- [31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *CVPR*, 2018.
- [32] Sara Fridovich-Keil and et al., “Plenoxels: Radiance fields without neural networks,” in *CVPR*, 2022.