Ref-NPR: Reference-Based Non-Photorealistic Radiance Fields

Yuechen Zhang[1,2] Zexin He[1] Jinbo Xing[1] Xufeng Yao[1] Jiaya Jia[1,2]
[1]The Chinese University of Hong Kong [2]SmartMore

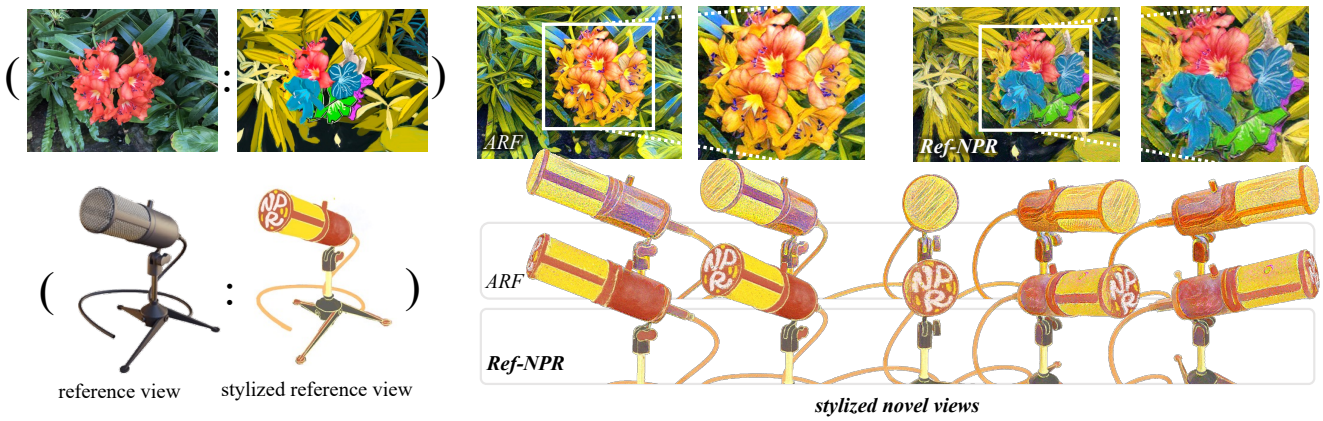

Figure 1. Given a pair of one reference view from a radiance field and its stylization, Ref-NPR propagates style more faithfully to novel views with semantic correspondence compared with state-of-the-art scene stylization method ARF [44].

Abstract

Existing 3D scene stylization methods employ an arbitrary style reference to transfer textures and colors as styles without establishing meaningful semantic correspondences. We present Reference-Based Non-Photorealistic Radiance Fields, i.e., Ref-NPR. It is a controllable scene stylization method utilizing radiance fields to stylize a 3D scene, with a single stylized 2D view taken as reference. To achieve decent results, we propose a ray registration process based on the stylized reference view to obtain pseudo-ray supervision in novel views, and exploit the semantic correspondence in content images to fill occluded regions with perceptually similar styles. Combining these operations, Ref-NPR generates non-photorealistic and continuous novel view sequences with a single reference while obtaining reasonable stylization in occluded regions. Experiments show that Ref-NPR significantly outperforms other scene and video stylization methods in terms of both visual quality and semantic correspondence. Code and data will be made publicly available.

1. Introduction

The past decade has witnessed a growing demand for stylizing and editing 3D scenes and objects in many fields such as augmented reality, game scene design, and digital artwork. Such tasks are traditionally achieved by creating 2D reference images and then converting them into stylized 3D textures. But it is common sense that utilizing cross-dimension correspondence directly is difficult, and professionals often spend a significant amount of time and effort to obtain stylized texture results similar to 2D reference schematics.

One key challenge in the 3D stylization problem is to make stylized results perceptually similar to the given style reference. Benefiting from radiance fields [3, 11, 29, 30, 38, 45], recent novel-view stylization methods [6, 9, 16, 17, 31, 44] greatly facilitate the style transfer from an arbitrary 2D style reference to 3D implicit representations, yet they do not allow creators to control generated results explicitly. For instance, with an arbitrary reference, we cannot specify the region to which the stylization should be applied, and it is hard to ensure the visual quality of the results. On the other hand, reference-based video stylization methods [18, 39] controllably generate stylized novel views with better semantic correspondence between content and style reference. But they suffer from divergence from the desired style when stylizing a frame sequence with unseen content, even under the assistance of stylized keyframes.

To tackle the above limitations, we therefore design a new paradigm to stylize 3D scenes from one stylized reference view and propose Reference-Based Non-Photorealistic

Radiance Fields (Ref-NPR), a controllable scene stylization approach that leverages the nature of volume rendering to maintain cross-view consistency and establishes semantic correspondence to transfer style in the entire scene.

Instead of using arbitrary style images as reference, Ref-NPR takes stylized views from radiance fields as reference to simultaneously achieve both flexible controllability and multi-view consistency. With such a stylized reference view, we design a reference-based ray registration process to project the 2D style reference into 3D space by utilizing the depth rendering of the radiance field, which provides pseudo-ray supervision to keep stylized novel views both geometrically and perceptually consistent with the stylized reference view. Further, to obtain semantic style correspondence in occluded regions, we perform template-based feature matching to use high-level semantic features as implicit style supervision. We then utilize such correspondence in the content domain to select style features in the given style reference, and use them to transfer style in the occluded regions. This way, the entire stylized 3D scene can be obtained from a single stylized reference view.

Ref-NPR generates satisfying stylized views with both semantic and geometric correspondences preserved, as shown in Fig. 1. Stylized novel views generated by Ref-NPR on various datasets are perceptually highly consistent with the given style reference while achieving impressive visual quality. We demonstrate that, with the same stylized view as reference, Ref-NPR qualitatively and quantitatively outperforms state-of-the-art scene stylization methods [31, 44].

To summarize, our contributions are threefold. 1) To introduce controllability into the scene stylization problem, we design a new paradigm to stylize 3D scenes from one stylized reference view. 2) We propose Ref-NPR, a controllable scene stylization approach that is composed of a reference-based ray registration process and a template-based feature matching scheme. 3) Ref-NPR successfully generates geometrically consistent novel-view stylizations with high-quality semantic correspondence. We have more comprehensive results and a demo video in the supplementary material for better demonstration.

2. Related Works

2.1. Stylization in 2D

Arbitrary style transfer is a classic computer vision problem under Non-Photorealistic Rendering (NPR) [13, 24]. Gatys et al. [12] first represented image style as a high-level feature extracted from a pre-trained deep neural network. Since then, parametric image style transfer methods [5, 19, 21, 23, 25, 26] have emerged to create high-quality stylized images efficiently. Confronted by a more challenging task, video stylization methods with arbitrary style input [4, 8, 33, 40, 41] mainly focus on keeping the temporal coherence to obtain continuous stylized frames. Though smoother, such stylized results lack interpretability and controllability, even if we provide stylized keyframes as reference.

Reference-based stylization methods stylize images with several content-related style reference images, where semantic correspondences are required. Multi-level semantic feature matching is utilized in [15, 27] to create a dense correspondence between the content image and the style reference with similar semantics. Explicit alignment like warping or content-aligned stylizing is then further applied in some other methods [18, 34, 35, 39] to ease the difficulty of the above feature matching. These methods are capable of getting controllable results by editing the style reference. However, when stylizing novel views, direct applying 2D reference-based methods cannot conduct stylization in unseen regions correctly.

2.2. Stylization in 3D

Reference-based style mapping. Even without using radiance fields, reference-based 3D stylization methods [1, 10, 14, 32] have already achieved satisfying results when stylizing novel views. However, obvious shortcomings exist in each of these methods. Texture Field [32] is trained on a limited ShapeNet [2] dataset, hence the stylization of 3D objects is only allowed within the trained categories. The StyLit family [10, 37] treats each view of a 3D object as a multi-channel 2D guidance and applies texture mapping, but suffers from flickering artifacts due to the lack of geometric prior. StyleProp [14] employs multi-view correspondence maps to get stylized novel views on a given 3D object, but only works limitedly on viewing directions around the reference view.

Stylizing radiance fields is a rising topic resulting from the recent popularity of radiance fields [6, 9, 29]. Huang et al. [16] is the first to apply scene stylization on implicit 3D representations. StylizedNeRF [17] then leverages a mutual learning strategy to get a stylized scene. Two recent methods, SNeRF [31] and ARF [44], focus on improving the stylization quality, while reducing GPU memory requirements by avoiding full-image supervision. However, as they do not require explicit correspondence modeling, the above methods are still unable to get controllable results, and the stylized novel view is often perceptually different from the style reference. To the best of our knowledge, a reference-guided controllable scene stylization method is still yet to exist. Given one stylized view as reference, our proposed Ref-NPR, in contrast to the aforementioned methods, acquires a 3D representation that is capable of generating stylized novel views with both geometric consistency and semantic correspondence to the style reference.

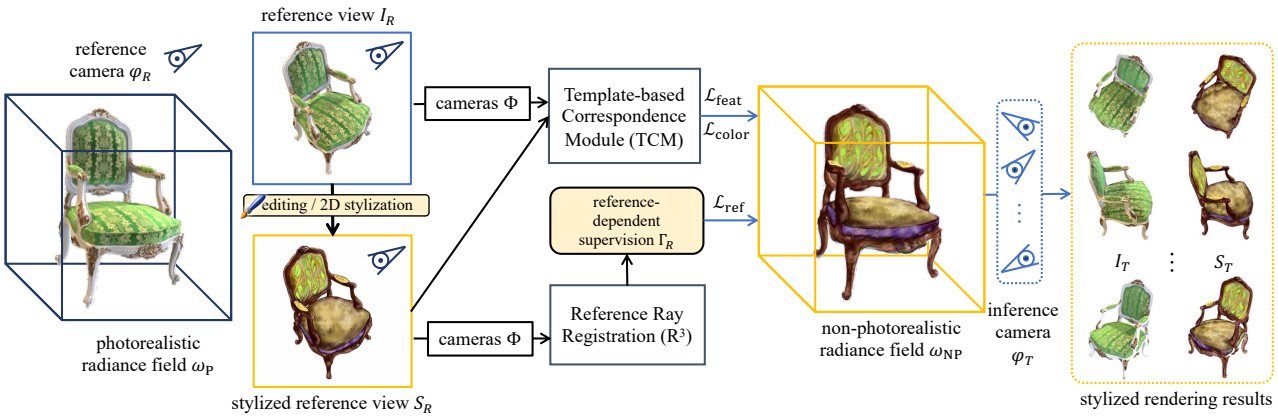

Figure 2. The workflow of **Ref-NPR**. Given a pre-trained photorealistic radiance field ω_{P}, we can provide a stylized reference view S_R to obtain reference-based supervisions $\mathcal{L}_{\text{ref}}$, $\mathcal{L}_{\text{feat}}$, and $\mathcal{L}_{\text{color}}$. Those loss constraints are used to optimize a non-photorealistic (NP) radiance field ω_{NP}. During inference, with such NP radiance field, stylized results S_T could be rendered from an arbitrary set of camera poses φ_T, corresponding to the original views I_T rendered by ω_{P}.

3. Ref-NPR

Ref-NPR aims to stylize a pre-trained photorealistic radiance field with the guidance of one or a few pairs of reference views and their corresponding stylizations. In this section, we illustrate the concrete process of Ref-NPR in Fig. 2 with a single reference as an example. Given a pre-trained photorealistic radiance field ω_{P}, a reference view I_R is rendered from one particular reference camera φ_R. By manually editing or by applying structure-preserving 2D-stylization algorithms like Gatys [12] or AdaIN [21], we can further obtain a stylized reference view S_R based upon the original reference I_R.

To propagate explicit supervision from the stylized reference view to novel views, we propose a Reference Ray Registration (R^3) process in Sec. 3.2 that produces a set of reference-dependent pseudo-rays Γ_R, which consists of the correlated rays produced between the reference camera φ_R and the set of training camera poses Φ. Furthermore, we use a Template-based Correspondence Module (TCM) in Sec. 3.3 to obtain implicit style supervision in occluded regions of the reference view. With explicit and implicit supervision provided by R^3 and TCM, a new non-photorealistic (NP) radiance field ω_{NP} is optimized, by which we have access to the stylized rendering results of arbitrary target views. The optimization details will be discussed in section Sec. 3.4.

3.1. Preliminary: Radiance Field Rendering

In volume rendering [20, 29], camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is an originated and directed linear function in 3D space defined by a camera origin $\mathbf{o} \in \mathbb{R}^3$ and a viewing direction $\mathbf{d} \in \mathbb{R}^3$ pointing from $\mathbf{o}$ to the center of a particular pixel of an image. Along each ray, N points are sampled as input to the radiance field ω, denoted by $\{\mathbf{r}(t_i)|i = 1 \dots N, t_i < t_{i+1}\}$. For each input sample, ω returns a density $\sigma_i = \sigma(\mathbf{r}(t_i))$ and a view-dependent color $\mathbf{c}_i = c(\mathbf{r}(t_i), \mathbf{d})$. In the discrete context, according to [29], the estimated accumulated pixel color $\hat{C}(\mathbf{r})$ of ray $\mathbf{r}$ is formulated as

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i(t_{i+1} - t_i)))\mathbf{c}_i, \qquad (1)$$

$$\text{where } T_i = \exp\Big(-\sum_{j=1}^{i-1} \sigma_j(t_{j+1} - t_j)\Big), \qquad (2)$$

which estimates the accumulated transmittance along the ray from t_1 to t_i. The radiance field is trained by directly minimizing the discrepancy between the predicted pixel color $\hat{C}(\mathbf{r})$ and the ground truth pixel color $C(\mathbf{r})$ for each ray, denoted by

$$\mathcal{L}_{\omega} = \sum_{\mathbf{r}} \|\hat{C}(\mathbf{r}) - C(\mathbf{r})\|_2^2. \qquad (3)$$

Based on the optimized radiance field, depth could be estimated by mapping ray $\mathbf{r}$ into an exact 3D position in the scene. To do this, a threshold σ_z is set on the accumulated density T_i calculated in Eq. (2), where the first sample to exceed this particular threshold is interpreted as the intersection point between the ray and the scene. We define the length of ray $\mathbf{r}$ as the distance between $\mathbf{o}$ and the intersection point, and denote it by $l(\mathbf{r})$. That is,

$$l(\mathbf{r}) = \min\{t_i | \sum_{j=1}^{i-1} \sigma_j(t_{j+1} - t_j) \geq \sigma_z\}. \qquad (4)$$

Then we write the intersection point corresponding to ray $\mathbf{r}$ as $\mathbf{x}(\mathbf{r}) = \mathbf{o} + l(\mathbf{r})\mathbf{d}$, which is our desired mapping.

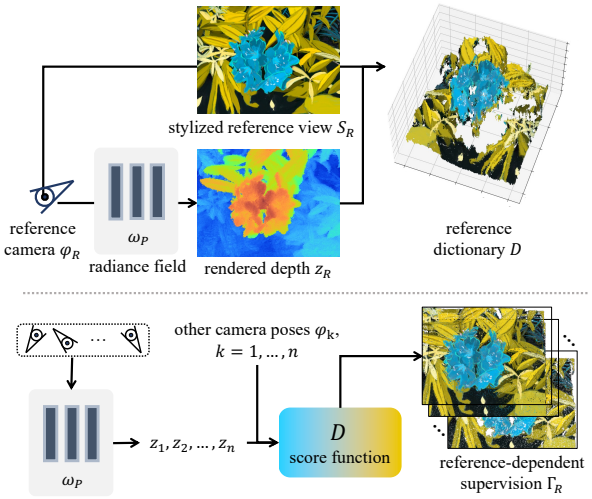

Figure 3. **Reference Ray Registration.** Given a stylized reference view S_R, along with its camera pose φ_R, a reference-based dictionary D is created by estimating pseudo-depths from ω_{P}. Then, a ray registration process is applied for each training camera pose to acquire a collection Γ_R containing pseudo-rays along with their assigned colors, as explicit supervision.

3.2. Reference Ray Registration

Compared with existing scene stylization methods [17, 31, 44], Ref-NPR is designed with the additional objective to establish the semantic consistency between novel views and the stylized reference view. Such an objective is shared by some scene modeling methods [7, 42, 43], where one commonly applied strategy is to utilize additional information outside of the 2D images, such as depth information. The extra 3D information introduced is notable for enhancing the rendering quality of the radiance field with limited training views. With radiance field ω_{P}, we therefore estimate the pseudo-depth information according to Eq. (4).

Reference Ray Registration (R^3) is designed to leverage the aforesaid pseudo-depth information to acquire reference-related novel view supervision, as illustrated in Fig. 3. With the property of depth rendering, pixels in the stylized reference view S_R are mapped to 3D space by estimating the lengths of their respective rays. We therefore construct a reference dictionary D, where the element at index (x, y, z) is a collection of all the rays terminating in the voxel with the same index, based on a quantization operator $Q(\cdot)$, which maps 3D positions to their corresponding voxels. Formally, we define this reference dictionary by

$$D_{(x,y,z)} = \{\mathbf{r}_i \in \varphi_R \,|\, Q(\mathbf{x}(\mathbf{r}_i)) = (x, y, z)\}, \tag{5}$$

in which φ_R is the reference camera and $\mathbf{x}(\mathbf{r}_i)$ is the intersection point of $\mathbf{r}_i$ estimated from the pseudo-depth $l(\mathbf{r}_i)$ according to Eq. (4). By splitting 3D space into discrete voxels, each entry in D may be mapped with multiple rays or none at all, depending on the estimated pseudo-depths. For each ray $\mathbf{r}_i \in \varphi_R$, we use $\hat{C}_R(\cdot)$ to denote the stylized color according to the stylized reference view S_R.

Now that S_R is propagated to 3D space with such reference dictionary, we thus register each ray $\mathbf{r}_j \in \Phi$ from the training views as a pseudo-ray $\hat{\mathbf{r}}_j \in \varphi_R$ in a best-matching manner, through minimizing the Euclidean distance of two corresponding intersection points in 3D space. Further, to avoid the over-matching problem coming from the large gap of ray directions, we deploy an additional constraint that the angle spanned between the directions of two matched rays should not exceed a certain threshold θ, i.e., $\angle(\mathbf{d}_{\mathbf{r}_i}, \mathbf{d}_{\mathbf{r}_j}) < \theta$. We formulate this ray registration process as

$$\hat{\mathbf{r}}_j = \underset{\substack{\mathbf{r_i} \in D_{(x,y,z)}, \\ \angle(\mathbf{d}_{\mathbf{r}_i}, \mathbf{d}_{\mathbf{r}_j}) < \theta}}{\arg\min} \|\mathbf{x}(\mathbf{r}_i) - \mathbf{x}(\mathbf{r}_j)\|_2, \tag{6}$$

$$\text{where } Q(\mathbf{x}(\mathbf{r}_j)) = (x, y, z), \mathbf{r}_j \in \Phi. \tag{7}$$

Ray registration essentially finds a ray in the dictionary whose intersection point drops into the same voxel $D_{(x,y,z)}$ as the ray of interest. Eventually, we construct reference-dependent pseudo-ray supervision as Γ_R by collecting each validated, registered ray $\mathbf{r}_j$ and assign its color in accordance to the corresponding reference ray $\hat{\mathbf{r}}_j$. We formally define such a collection of reference-dependent pseudo-rays and their stylized colors with the aforementioned $\hat{C}_R(\cdot)$ as

$$\Gamma_R = \{(\mathbf{r}_j, \hat{C}_R(\hat{\mathbf{r}}_j)) \,|\, \mathbf{r}_j \in \Phi \cup \varphi_R, \hat{\mathbf{r}}_j \neq \varnothing\}, \tag{8}$$

where Φ is the collection of all accessible camera poses.

3.3. Template-Based Semantic Correspondence

Though R^3 could effectively generate pseudo-ray supervision around the given reference camera pose φ_R, yet it fails to register reference rays to occluded regions under such a camera, especially for 360° scenes in [22, 29]. Fortunately, with the stylized reference view S_R based on source content I_R, we may establish a content-style mapping as a template, to broadcast such reference style to novel views with semantically similar contents. We therefore introduce a Template-based Correspondence Module (TCM) to utilize this content-style correspondence and construct a semantic correlation within the content domain.

As illustrated in Fig. 4, for each content domain view I rendered from ω_{P} under some certain camera pose $\varphi \in \Phi$, we obtain its high-level semantic feature map F_I from a pre-trained semantic feature extractor (e.g., VGG16 [36]). Similarly, we denote the extracted feature maps for content reference I_R and style reference S_R by F_{I_R} and F_{S_R}, respectively, and use a superscript to index each element in the feature maps. We therefore construct our desired guidance feature F_G for further supervision, described by a search-and-replace process on 2D position (i, j) as

$$F_G^{(i,j)} = F_{S_R}^{(i^*,j^*)}, \tag{9}$$

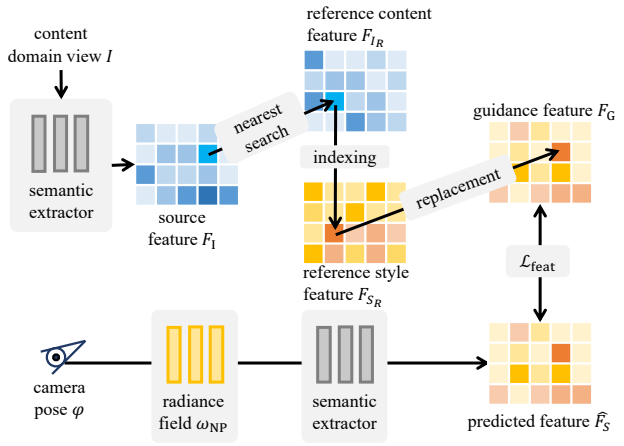

Figure 4. Illustration of the **Template-based Correspondence Module (TCM)**. To get the implicit style supervision of one view, the content domain view I is passed to a semantic extractor to obtain its semantic feature F_I. Then a nearest search replacement process is conducted to replace the reference feature according to the semantic correspondence in the content domain. The resulting guidance feature F_G serves as implicit supervision to optimize the NP radiance field ω_{NP}.

$$\text{where } (i^*, j^*) = \arg\min_{i',j'} dist\big(F_I^{(i,j)}, F_{I_R}^{(i',j')}\big). \tag{10}$$

Here we use $dist(\mathbf{a}, \mathbf{b})$ to denote the distance between two feature vectors $\mathbf{a}$ and $\mathbf{b}$, which is proved to be effective [23, 44] by taking the form of cosine distance

$$dist(\mathbf{a}, \mathbf{b}) = 1 - \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}, \tag{11}$$

when evaluating semantic features.

Finally, for a stylized view S rendered from the NP radiance field ω_{NP} under the camera pose φ, we conduct implicit feature-level supervision by forcing its semantic feature $\hat{F}_{\text{S}}$ to imitate the aforesaid guidance feature F_G.

3.4. Ref-NPR Optimization

With the previously discussed collection Γ_R and guidance feature F_G as supervision, we optimize Ref-NPR and obtain the stylized scene representation ω_{NP}.

In each training iteration, we sample a subset from Γ_R and denote the set of reference-dependent rays as N_s. For each sampled reference ray $\mathbf{r}_k \in N_s$, $\hat{C}_R(\hat{\mathbf{r}}_k)$ is the assigned color of the corresponding pseudo-ray $\hat{\mathbf{r}}_k$, as defined in Eq. (6), and we further denote $\hat{C}_{\text{NP}}(\mathbf{r}_k)$ to be the stylized color rendered from ω_{NP}. This explicit supervision is formulated as the reference loss

$$\mathcal{L}_{\text{ref}} = \frac{1}{|N_s|} \sum_{\mathbf{r}_k \in N_s} \|\hat{C}_{\text{NP}}(\mathbf{r}_k) - \hat{C}_R(\hat{\mathbf{r}}_k)\|_2^2. \tag{12}$$

As for implicit supervision, the discrepancy between F_G and $\hat{F}_{\text{S}}$ should be minimized, as discussed in Sec. 3.3. To maintain the original content structure during such implicit stylization, we also minimize the mean squared distance between content feature F_I and stylized feature $\hat{F}_{\text{S}}$, according to [12]. This implicit supervision is formulated as the feature loss

$$\mathcal{L}_{\text{feat}} = \frac{1}{N} \sum_{i,j}^{N} \big(dist(F_G^{(i,j)}, \hat{F}_S^{(i,j)}) + \lambda' \|F_I^{(i,j)} - \hat{F}_S^{(i,j)}\|_2^2\big), \tag{13}$$

where λ' is a balancing factor.

However, as discussed in [23, 44], optimizing the cosine distance between feature vectors cannot effectively eliminate color mismatches. To address this issue, we transfer the average color in a patch by a coarse color-matching loss

$$\mathcal{L}_{\text{color}} = \frac{1}{N} \sum_{i,j}^{N} \|\bar{C}_{\text{NP}}^{(i,j)} - \bar{C}_{\text{R}}^{(i^*,j^*)}\|_2^2, \tag{14}$$

in which $\bar{C}_{\text{NP}}^{(i,j)}$ is the ω_{NP}-rendered average color of the patch at feature-level index (i, j), and $\bar{C}_R^{(i^*,j^*)}$ is the average color of reference patch matched by minimizing feature distance, as described in Eq. (10). Since semantic features are extracted at the image level, considering the memory limitation caused by back-propagation, we follow the gradient cache strategy in [44] and optimize ω_{NP} patch-wisely.

Ultimately, the overall objective is $\mathcal{L}_{\text{NP}} = \lambda_f \mathcal{L}_{\text{feat}} + \lambda_r \mathcal{L}_{\text{ref}} + \lambda_c \mathcal{L}_{\text{color}}$, where $\lambda_{(\cdot)}$ are the balancing factors. Once ω_{NP} is optimized, we may consider it to be a normal radiance field and render stylized novel views with arbitrary camera poses.

4. Experiments

4.1. Implementation Details

Ref-NPR is based on the ARF codebase [44] and uses Plenoxels [11] as the radiance field for scene representation. We follow Plenoxels's training scheme to obtain the photorealistic radiance field ω_{P}. As we do not expect a view-dependent color change in stylized scenes, following [17, 44], we discard view-dependent rendering and apply a view-independent fitting on training views for two epochs before optimizing ω_{NP}. Then, content domain views I are rendered from ω_{P} after this view-independent training. In addition, to keep the same geometrical structure of the scene, we do not optimize the density function $\sigma(\mathbf{r}(t_i))$ in ω_{NP} [11, 44].

In R^3, we set reference dictionary D as a cube containing 256^3 voxels. To parallelize the registration, we store at most 8 rays at each entry $D_{(x,y,z)}$. The angle constraint of directions in Eq. (6) is empirically set to $\cos(\theta) > 0.6$ for ray registration. For each training step, we set the number of pseudo-ray samples to be $|N_s| = 10^6$, with half

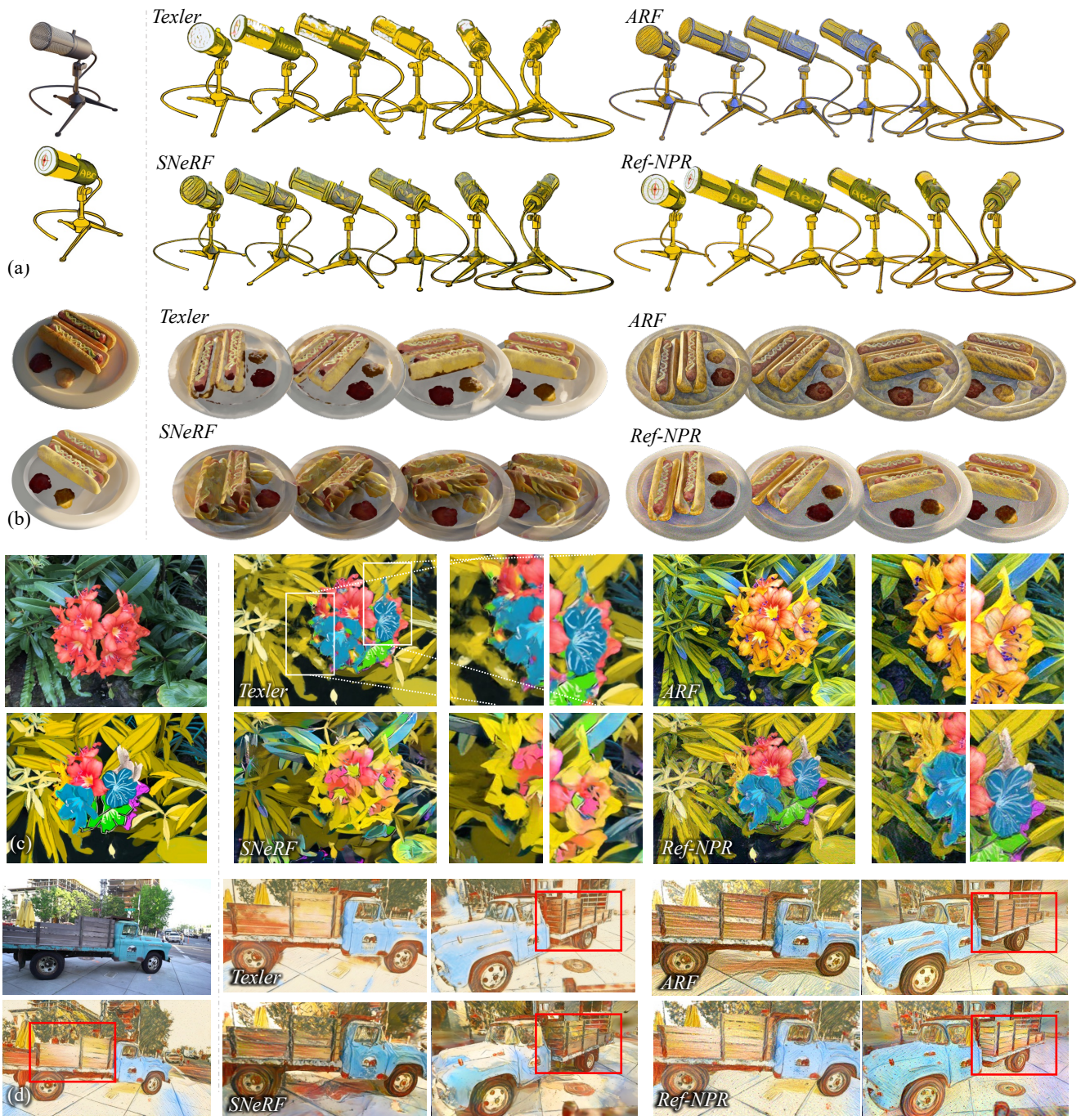

Figure 5. Qualitative comparisons in novel-view stylization. For each example, we provide the reference view (top) and its corresponding stylized reference view (bottom) on the left. We compare Ref-NPR with other methods in Synthetic (a), (b) [29], LLFF (c) [28], zoomed-in on the flower and the occluded regions, and T&T (d) [22]. Semantic consistencies are highlighted.

of them from φ_R and the other half from rays registered in correlated views. In TCM, we use VGG16 [36] as the semantic feature extractor. We concatenate features that have passed through the activation layers in stages 3 and 4 ($relu\_3\_^*$, $relu\_4\_^*$), and use them for $\mathcal{L}_{\text{feat}}$. Balancing factors among loss terms are set to $\lambda' = 5 \times 10^{-3}$, $\lambda_c = 5$, and $\lambda_r = \lambda_f = 1$. We train each scene on one NVIDIA 3090 GPU for 10 epochs. Before training the final 3 epochs, for a smoother content update, we replace ω_{P} with a frozen ω_{NP} as the content view generator in TCM, and minimize $\mathcal{L}_{\text{ref}}$ and $\mathcal{L}_{\text{feat}}$ only, with $\lambda_f = 0.2$ and $\lambda' = 0$.

Figure 6. Ablations on the effectiveness of respective loss components.

4.2. Datasets

We validate our Ref-NPR on three commonly used datasets. *Synthetic* [29] is a well-defined synthetic dataset for 3D objects. The stylization performance on Synthetic is not reported in many scene stylization methods [16, 17, 31, 44]. It is challenging to transfer a full-image style reference to an object with a foreground mask. However, stylization on synthetic datasets is meaningful since it indicates the capability to stylize human-designed 3D models [14]. To avoid overflow of the stylized view on the mask boundary, we apply a 2D morphological erosion on the foreground mask in R^3. *LLFF* [28] is a real-world high-resolution dataset for novel view synthesis. We choose the resolution as 4× downsampled, following [11, 44]. *Tanks and Temples (T&T)* [22] is a 360° scene dataset for novel-view synthesis and scene reconstruction with 200 to 300 high-resolution training views for each scene. In R^3, apart from the reference view itself, we only register rays in Plenoxels's foreground model [11] to make the dictionary more compact. We follow the official split of training and testing for all datasets above.

4.3. Comparisons

We compare our Ref-NPR with two recent scene stylization methods: ARF [44] and a reimplemented SNeRF on Plenoxels [11, 31]. Besides, we include a reference-based video stylization method by Texler et al. [39] for a more comprehensive comparison.

Qualitative comparison. Fig. 5 shows qualitative comparisons with competitive video and scene stylization approaches. Texler et al. [39] synthesizes novel views with a proper low-level color distribution. However, suffering from limited correspondence in the whole scene, it lacks geometrical coherence and results in flickering effects, as shown in Fig. 5(a). ARF [44] and SNeRF [31] both maintain geometric consistency in novel views, yet they are plagued by the missing content-style correlation, whether the style reference is manually created (Fig. 5(a-c)) or generated by a 2D-neural stylization method [23] in Fig. 5(d). In comparison, Ref-NPR significantly improves geometric and semantic-style consistency in each example. Moreover, benefiting from TCM, Ref-NPR fills occluded regions with perceptually reasonable visual contents. A more detailed comparison can be found in the supplementary materials.

Ref-LPIPS ↓	Consist.	Synthetic	LLFF	T&T
Texler [39]	✗	0.1556	0.3591	0.6709
ARF [44]		0.1883	0.5608	0.6376
SNeRF [31]	✓	0.1845	0.5693	0.6826
Ref-NPR		0.1711	0.3802	0.6327

Table 1. Reference-related novel view LPIPS on 8 examples.

Quantitative comparison. Style transfer methods usually conduct user preference surveys as a subjective evaluation metric. For reference-based stylization, the visual quality can be quantitatively evaluated by the Perceptual Similarity (LPIPS) [46] between the stylized reference view and the top 10 nearest novel views under the test camera poses. We also evaluate the ability to maintain the cross-view geometric consistency for each method, where Texler [39] does not utilize any geometric information, as shown in Tab. 1. Meanwhile, compared with state-of-the-art scene stylization methods, Ref-NPR achieves cross-view geometric consistency with a higher perceptual similarity.

4.4. Ablation Studies

We conduct a series of ablation studies to analyze Ref-NPR's validity. Firstly, a module-wise ablation is done to validate the effectiveness of supervision components in R^3 and TCM. As shown in Fig. 6 (a), when we remove the pseudo-ray supervision $\mathcal{L}_{\text{ref}}$, detailed textures related to the style reference are lost. To address the missing texture problem, it is natural to cast rays in φ_R as additional supervision. However, implicit volume rendering does not guarantee cross-view consistency on the surface. As shown in Fig. 6 (b), detailed textures in novel views are still partially missing if pixel-level supervision is only applied on

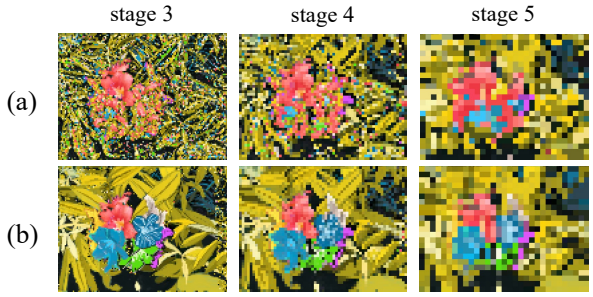

Figure 7. Patch-wise replacement results on features from the last three stages of VGG backbone. (a) Matching with the style reference directly. (b) Matching with the content reference (TCM).

φ_R. Moreover, in TCM, if we discard feature supervision $\mathcal{L}_{\text{feat}}$, a photo-realistic colorization result will occur in the occluded regions, as shown in Fig. 6 (c). On the other hand, a lack of color supervision $\mathcal{L}_{\text{color}}$ as shown in Fig. 6 (d) may cause a color mismatch in the occluded region. In the full Ref-NPR model, the combination of R^3 and TCM compensates for the above shortcomings and fully utilizes the merits of each module.

We also validate the effectiveness of the patch-wise matching scheme in TCM. As shown in Fig. 7 (a), a direct match with the stylized view often fails to get desired correspondence due to the domain gap in the semantic feature space. Conversely, in Fig. 7 (b), TCM matches features within the same content domain. Hence the semantic correspondence is preserved at each level of semantic features.

5. Discussions

Adapt to general style transfer. Although Ref-NPR itself requires a stylized reference view, it can be extended to use arbitrary style images as reference. As shown in Fig. 8, we use three different 2D stylization methods [12, 21, 23] to generate three reference views of the same scene, each with a slightly different stylization. Then we feed them into Ref-NPR to render three sets of stylized novel views, each preserving the characteristics of the corresponding style reference. This extension enables Ref-NPR to work with arbitrary style reference images and provides a faster, more flexible solution compared with other scene stylization methods [16, 17, 31, 44].

Multi-reference. As semantic content in a single stylized view may be deficient to stylize a large-scale scene, it is essential to extend Ref-NPR adaptively for multiple style reference. This can be easily accommodated by registering rays using all stylized reference views in R^3, and expanding the capacity of styles and content features in TCM. Fig. 9 gives an example of multi-reference input in the scene Playground [22]. With two additional stylized views, better feature matching is obtained from richer style content.

Limitations. Although Ref-NPR exhibits powerful versatility, it fails when no meaningful semantic correspondence can be found in the reference view, as shown in Fig 10. Besides, feature matching may fail when stylizing objects with intricate geometric structures.

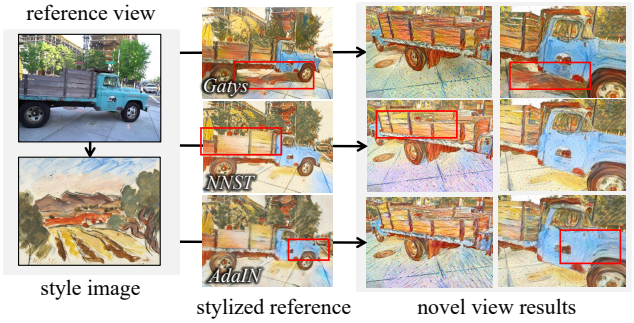

Figure 8. The pipeline of Ref-NPR naturally extends to arbitrary style reference. Images are cropped for a better presentation. Method-related style textures are highlighted.

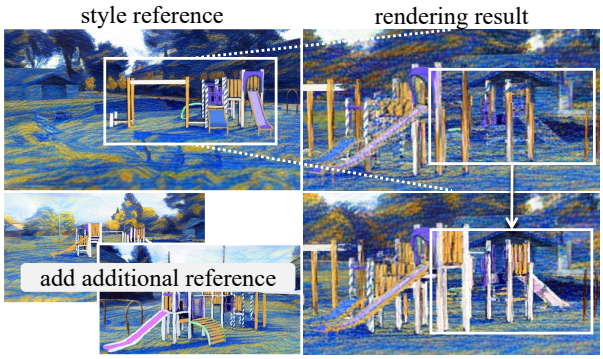

Figure 9. Multi-reference results of a 360° scene.

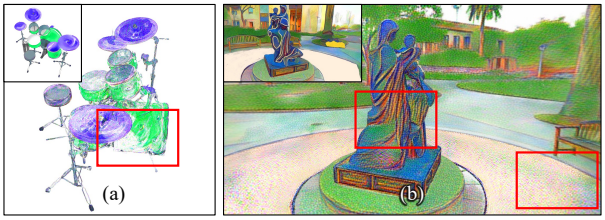

Figure 10. Failure cases of Ref-NPR. Feature matching may fail when (a) stylizing intricate geometric structures, or (b) stylizing a large scene with a single reference.

6. Conclusion

In this work, we propose a new paradigm in controllable 3D scene stylization and present Ref-NPR, a novel framework for controllable non-photorealistic 3D scene stylization based on radiance fields. Powered by a stylized reference view, Ref-NPR generates geometrically consistent novel-view stylizations with high-quality semantic correspondence. We hope Ref-NPR can be utilized as a tool for professional visual content creators to maximize the efficiency of human creativity.

A. Supplementary Materials

We provide a document and a video as supplementary materials for Ref-NPR to help readers better understand the motivation and performance of our method. In this document, we first discuss the implementation details in Appendix B, and provide visualizations in Appendix C to better illustrate proposed modules in Ref-NPR, followed by more examples and visualizations in Appendix D to demonstrate the performance and controllability of Ref-NPR.

In addition to this document, we provide a video to better present the results and comparisons.

B. Technical Details

Implementation details. Two worth-noting details may affect the visual quality of stylization results when implementing Ref-NPR.

- Before computing image-level loss terms ($\mathcal{L}_{\text{color}}$ and $\mathcal{L}_{\text{feat}}$), for LLFF [28] and T&T [22] dataset, we downsample both stylized and content views by 2x to speed up the calculation of patch-wise feature distance.
- Different from the implicit feature loss $\mathcal{L}_{\text{feat}}$, in order to get a high-level semantic color mapping for the color-matching loss $\mathcal{L}_{\text{color}}$, we evaluate distances between features extracted by the last stage (i.e., stage 5) of VGG backbone [36]. Besides, when calculating $\mathcal{L}_{\text{color}}$, we exclude the position of interest (i, j) where the semantic feature is not close enough to any feature in the reference view, to avoid over-matching. Such a constraint of the feature distance for valid position (i, j) is formulated as

$$\min_{i',j'} dist(F_I^{(i,j)}, F_{I_R}^{(i',j')}) < 0.4\,. \tag{15}$$

Details of comparison. For Texler [39], we conduct our experiments based on the official implementation. As the reference view can be arbitrarily selected, desired continuous views with high-quality temporal coherence in the test view sequence might not exist. Hence we only use the RGB image sequence as input. We follow the default settings in the original paper and train each scene for 30,000 iterations. For SNeRF [31], we re-implement it based on Plenoxels [11], and choose Gatys [12] as the stylization method. For each training view, we train the stylization step for 10 iterations and train the stylization of the whole scene for 10 epochs.

Quantitative comparison. In Sec. Sec. 4.3, we provide a reference-based perceptual similarity metric for evaluation. Tab. 2 reports detailed LPIPS scores for each scene. Note that scene-wise LPIPS scores vary a lot. We speculate that fluctuations in results are due to the significant difference in camera poses between the reference view and all other test views. Besides, Texler [39] gets a slightly better reference-related LPIPS. Still, it does not get satisfying results when the camera pose diverges far from the reference camera φ_R, which is depicted in Fig. 14 and the supplementary video.

reference view training views with registered rays

Figure 11. Two examples to visualize registered rays in R^3. We paste pseudo-rays on content images in the first example for a better presentation.

C. Method Visualizations

Reference ray registration. Fig. 11 gives two concrete examples of how ray registration provides supervision in reference-dependent areas. Rays related to the stylized reference S_R are projected to each training view to provide pseudo-ray supervision.

Template-based feature matching. Except for explicit supervision in R^3, the implicit supervision provided by TCM is essential to occluded regions. Fig. 12 shows two examples of patch-wise replacement results. For guidance feature F_G, we select VGG features at stages 3 and 4. Since the patch-wise semantic feature is a high-level representation for each patch, the receptive field is much larger than the corresponding image patch.

Conversely, directly using patch replacement results at the same stages for the color supervision $\mathcal{L}_{\text{color}}$ may result in a color mismatch problem, as highlighted in Fig. 11. This

Ref-LPIPS ↓	Geo. Consist.	Chair	Ficus	Hotdog	Mic	Flower	Horn	Truck	Playground	Average
Texler [39]	✗	0.167	0.120	0.216	0.119	0.230	0.488	0.667	0.675	**0.335**
ARF [44]		0.185	0.123	0.300	0.146	0.619	0.502	0.683	0.592	0.394
SNeRF [31]	✓	0.188	0.129	0.283	0.138	0.646	0.492	0.702	0.663	0.405
Ref-NPR		0.164	0.122	0.273	0.126	0.289	0.471	0.669	0.596	**0.339**

Table 2. Reference-related novel view LPIPS for each test scene.

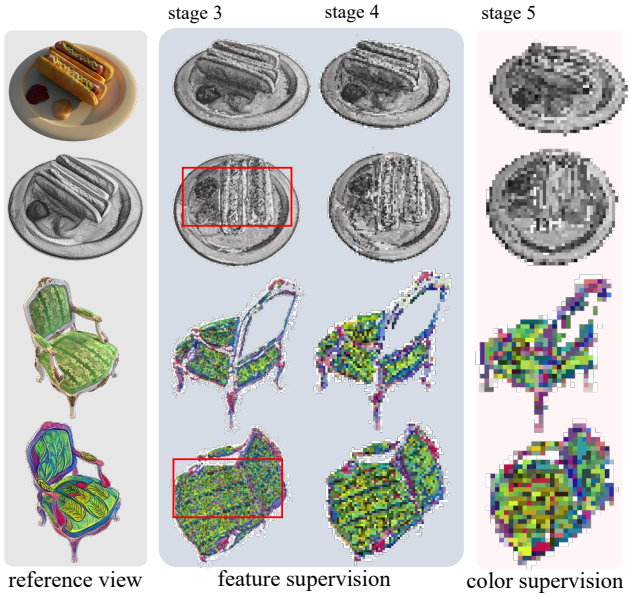

Figure 12. Two examples of patch-wise replacement on VGG feature at the last three stages to visualize the semantic correspondence. Color mismatch problems in shallow semantic features are highlighted.

problem is mainly caused by the receptive field difference between the feature patch and the image patch. Hence, as mentioned in Appendix B, we evaluate feature distances at the last VGG stage for color-matching supervision.

Loss balancing ablation. In addition to the ablation studies on the microphone example provided in Sec. Sec. 4.4, we conduct another ablation on the scene flower to discuss the effectiveness of color-matching loss $\mathcal{L}_{\text{color}}$ and the smooth content update strategy, which is described in Sec. Sec. 4.1.

For the same content view in Fig. 13 (a), the color mismatch problem would exist in occluded regions when we remove the color-matching loss $\mathcal{L}_{\text{color}}$, as shown in Fig. 13 (b). In Fig. 13 (c), we find that the stylized view without applying the smooth update strategy leads to occluded regions being under-stylized, which implies that the quality of semantic correspondence in the original content domain needs to be enhanced by TCM. A full model in Fig. 13 (d) clearly shows a satisfying stylization result in terms of both color and style.

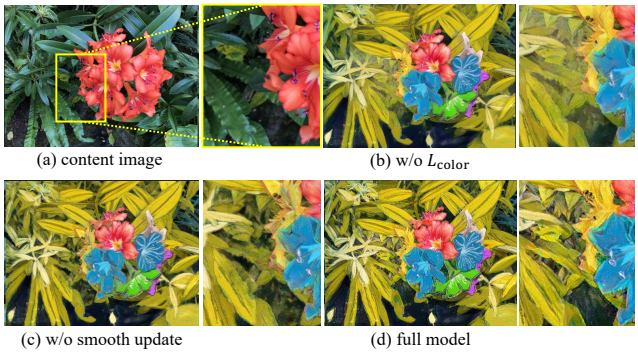

Figure 13. Ablation on the color-matching loss and smooth update strategy. The occluded region is zoomed in.

D. Results

More comparisons. Fig. 14 offers two additional examples to compare our method with [31, 39, 44]. As discussed in Sec. Sec. 4.3, Texler can generate novel-view stylized results with a proper color distribution, but consistent results with the reference stylized view can be only obtained under the condition that the test camera pose is around the reference. More specifically, it fails to generate reasonable style in the occluded regions and has some flickering or ghosting artifacts in a continuous sequence. Two scene stylization methods [31, 44] are unable to find a desired style mapping to the entire scene. Neither in the reference-related regions nor the occluded regions. By contrast, results generated by Ref-NPR keep both semantic correspondence and geometric consistency with the reference view.

Flexibility & controllability. In Sec. Sec. 5, we show the ability of Ref-NPR to adapt with an arbitrary image as reference. Fig. 15 gives two examples to demonstrate the flexibility of Ref-NPR, where the stylized reference view is generated by selecting one stylized view from ARF for each scene. In Fig. 15 (a), we manually edit the selected view and take it as the style reference. Ref-NPR faithfully reproduces the textures in the edited regions. Meanwhile, as shown in Fig. 15 (b), our method can reproduce the original novel-view stylizations by ARF through feeding in a stylized view as reference, which requires high-quality semantic correspondence.

Except for the local editing and scene stylization reproducing, the controllability of Ref-NPR can also be rep-

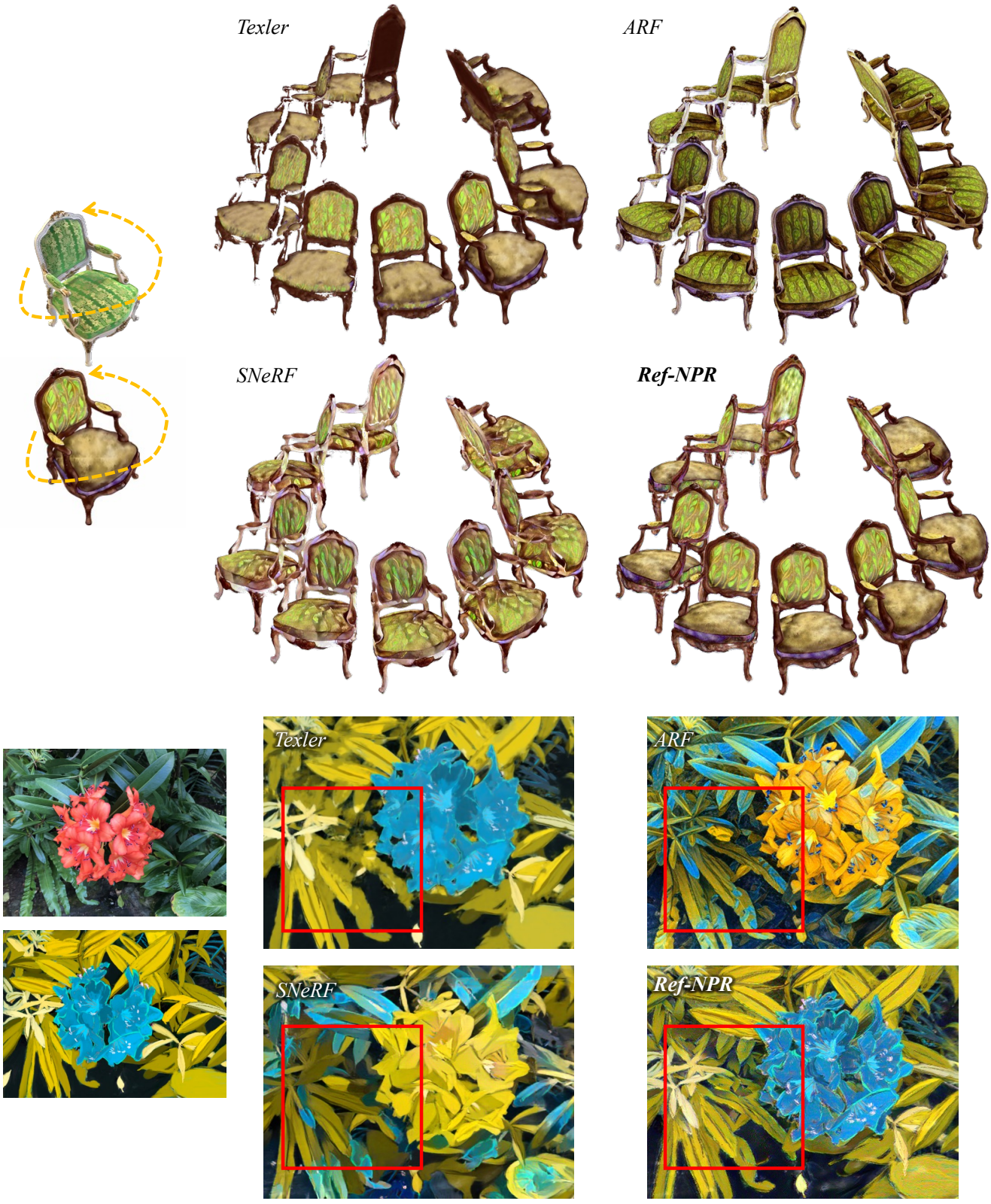

Figure 14. Additional examples for qualitative comparisons.

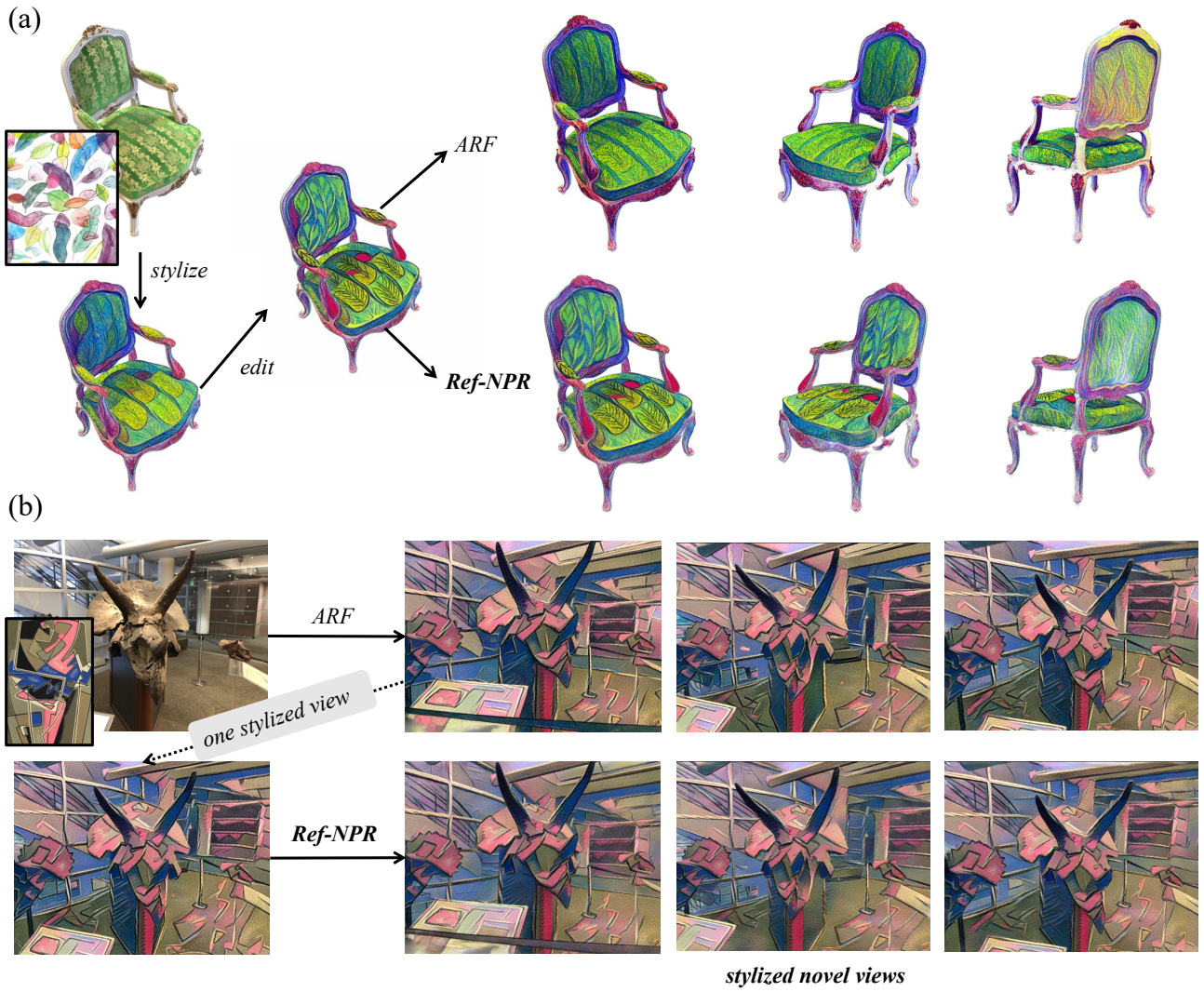

Figure 15. Examples to show the flexibility of Ref-NPR: (a) reference editing based on a stylized view, and (b) reproducing novel-view stylization given one stylized view generated by ARF [44] as reference.

resented by adapting scene stylization to various styles. Fig. 16 shows two examples of applying multiple styles to the same scene. Ref-NPR is capable of producing a faithful stylization result for each style owing to the modeling of cross-view semantic correspondence.

References

[1] Sai Bi, Nima Khademi Kalantari, and Ravi Ramamoorthi. Patch-based optimization for image-based texture mapping. *ACM Trans. Graph.*, 36(4), 2017. 2

[2] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 2

[3] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, 2022. 1

[4] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent online video style transfer. In *ICCV*, pages 1105–1114, 2017. 2

[5] Tian Qi Chen and Mark W. Schmidt. Fast patch-based style transfer of arbitrary style. *ArXiv*, abs/1612.04337, 2016. 2

[6] Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, and Wei-Chen Chiu. Stylizing 3d scene via implicit representation and hypernetwork. In *CVPR*, pages 1475–1484, 2022. 1, 2

[7] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster

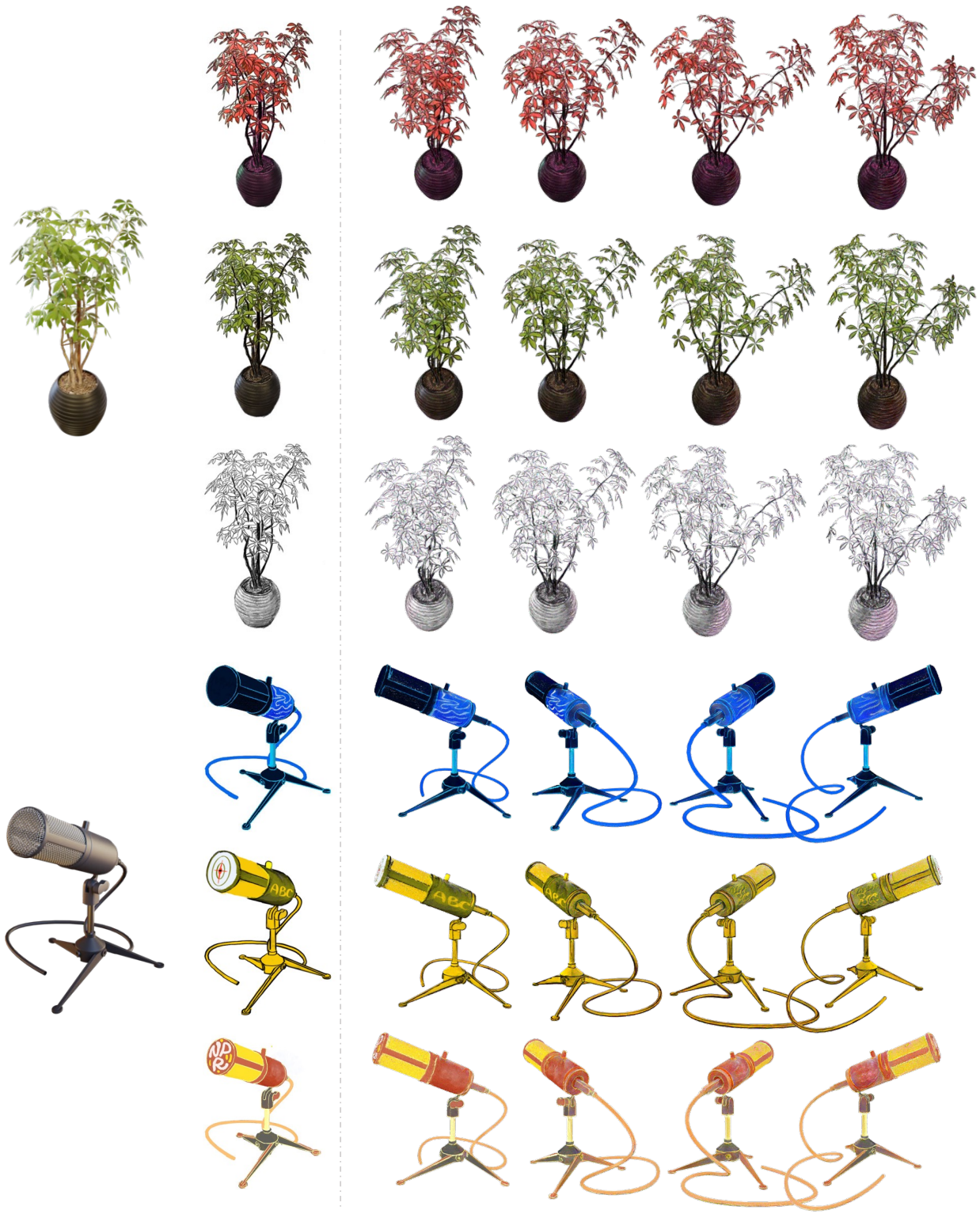

Figure 16. Examples to show the controllability of Ref-NPR. Stylized novel-view rendering results are satisfactory with references in different styles.

training for free. In *CVPR*, June 2022. 4

[8] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. Arbitrary video style transfer via multi-channel correlation. In *AAAI*, volume 35, pages 1210–1217, 2021. 2

[9] Zhiwen Fan, Yifan Jiang, Peihao Wang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Unified implicit neural stylization. In *ECCV*, pages 636–654. Springer, 2022. 1, 2

[10] Jakub Fišer, Ondřej Jamriška, Michal Lukáč, Eli Shechtman, Paul Asente, Jingwan Lu, and Daniel Sỳkora. Stylit: illumination-guided example-based stylization of 3d renderings. *ACM Trans. Graph.*, 35(4):1–11, 2016. 2

[11] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 1, 5, 7, 9

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 2, 3, 5, 8, 9

[13] Bruce Gooch and Amy Gooch. *Non-photorealistic rendering*. AK Peters/CRC Press, 2001. 2

[14] Filip Hauptfleisch, Ondrej Texler, Aneta Texler, Jaroslav Krivánek, and Daniel Sỳkora. Styleprop: Real-time example-based stylization of 3d models. In *Computer Graphics Forum*, volume 39, pages 575–586. Wiley Online Library, 2020. 2, 7

[15] Mingming He, Jing Liao, Dongdong Chen, Lu Yuan, and Pedro V Sander. Progressive color transfer with dense semantic correspondences. *ACM Trans. Graph.*, 38(2):1–18, 2019. 2

[16] Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, and Ming-Hsuan Yang. Learning to stylize novel views. In *ICCV*, 2021. 1, 2, 7, 8

[17] 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. 1, 2, 4, 5, 7, 8

[18] Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sỳkora. Stylizing video by example. *ACM Trans. Graph.*, 38(4):1–11, 2019. 1, 2

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 2

[20] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM Trans. Graph.*, 18(3):165–174, 1984. 3

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 3, 8

[22] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36(4), 2017. 4, 6, 7, 8, 9

[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, 5, 7, 8

[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. *TVCG*, 19(5):866–885, 2012. 2

[25] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *NeurIPS*, 30, 2017. 2

[26] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *ECCV*, pages 453–468, 2018. 2

[27] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *ACM Trans. Graph.*, 36(4):120, 2017. 2

[28] 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. *ACM Trans. Graph.*, 2019. 6, 7, 9

[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 3, 4, 6, 7

[30] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 1

[31] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *arXiv preprint arXiv:2207.02363*, 2022. 1, 2, 4, 7, 8, 9, 10

[32] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2019. 2

[33] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos and spherical images. *IJCV*, 126(11):1199–1219, 2018. 2

[34] Ahmed Selim, Mohamed Elgharib, and Linda Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM Trans. Graph.*, 35(4):1–18, 2016. 2

[35] YiChang Shih, Sylvain Paris, Connelly Barnes, William T Freeman, and Frédo Durand. Style transfer for headshot portraits. *ACM Trans. Graph.*, 2014. 2

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 4, 6, 9

[37] Daniel Sỳkora, Ondřej Jamriška, Ondrej Texler, Jakub Fišer, Michal Lukáč, Jingwan Lu, and Eli Shechtman. Styleblit: Fast example-based stylization with local guidance. In *Computer Graphics Forum*, volume 38, pages 83–91. Wiley Online Library, 2019. 2

[38] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, pages 8248–8258, June 2022. 1

[39] Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menglei Chai, Sergey Tulyakov, and Daniel Sýkora. Interactive video stylization using few-shot patch-based training. *ACM Trans. Graph.*, 39(4):73, 2020. 1, 2, 7, 9, 10

[40] Wenjing Wang, Shuai Yang, Jizheng Xu, and Jiaying Liu. Consistent video style transfer via relaxation and regularization. *TIP*, 29:9125–9139, 2020. 2

[41] Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. Ccpl: Contrastive coherence preserving loss for versatile style transfer. In *ECCV*, 2022. 2

[42] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *ECCV*, 2022. 4

[43] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *CVPR*, pages 5438–5448, 2022. 4

[44] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields, 2022. 1, 2, 4, 5, 7, 8, 10, 12

[45] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 1

[46] 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. 7