# FPRF: Feed-Forward Photorealistic Style Transfer of Large-Scale 3D Neural Radiance Fields

**GeonU Kim**[1], **Kim Youwang**[2], **Tae-Hyun Oh**[1,2,3]
[1]Grad. School of Artificial Intelligence, POSTECH
[2]Dept. of Electrical. Engineering, POSTECH
[3] Institute for Convergence Research and Education in Advanced Technology, Yonsei University
gukim@postech.ac.kr, youwang.kim@postech.ac.kr, taehyun@postech.ac.kr

## Abstract

We present FPRF, a feed-forward photorealistic style transfer method for large-scale 3D neural radiance fields. FPRF stylizes large-scale 3D scenes with arbitrary, multiple style reference images without additional optimization while preserving multi-view appearance consistency. Prior arts required tedious per-style/-scene optimization and were limited to small-scale 3D scenes. FPRF efficiently stylizes large-scale 3D scenes by introducing a style-decomposed 3D neural radiance field, which inherits AdaIN's feed-forward stylization machinery, supporting arbitrary style reference images. Furthermore, FPRF supports multi-reference stylization with the semantic correspondence matching and local AdaIN, which adds diverse user control for 3D scene styles. FPRF also preserves multi-view consistency by applying semantic matching and style transfer processes directly onto queried features in 3D space. In experiments, we demonstrate that FPRF achieves favorable photorealistic quality 3D scene stylization for large-scale scenes with diverse reference images.

## Introduction

Large-scale 3D scene reconstruction is a longstanding problem in computer vision and graphics (Früh and Zakhor 2004; Snavely, Seitz, and Szeliski 2006; Pollefeys et al. 2008; Agarwal et al. 2009; Zhu et al. 2018; Tancik et al. 2022), which aims to build realistic 3D virtual scenes from a set of images. Recently, Neural Radiance Fields (Mildenhall et al. 2020; Barron et al. 2021, 2022; Fridovich-Keil et al. 2022; Jun-Seong et al. 2022; Müller et al. 2022; Fridovich-Keil et al. 2023) and its large-scale extensions (Tancik et al. 2022; Turki, Ramanan, and Satyanarayanan 2022; Zhenxing and Xu 2022) have shown remarkable progress in modeling coherent and photorealistic outdoor 3D scenes, suggesting promising future directions, *e.g.*, VR/AR applications. In this work, we take a step further and focus on a task, photorealistic style transfer (PST) for large-scale 3D scenes, *i.e.*, 3D scene PST.

The 3D scene PST task aims to transfer the visual styles of style reference images onto a large-scale 3D scene represented by a *neural radiance field*. Within this objective, the resulting stylized output is expected to be photorealistic and preserve the geometric structure of the original scene. Large-scale 3D scene PST has various applications, where it can enrich virtual 3D spaces of XR applications and realistically augment existing real-world autonomous driving datasets (Geiger, Lenz, and Urtasun 2012; Cordts et al. 2016).



| | UPST-NeRF (arXiv'22) | StyleRF (CVPR'23) | LipRF (CVPR'23) | **FPRF (Ours)** |
|---|---|---|---|---|
| Photorealistic | ✓ | ✗ | ✓ | ✓ |
| Feed-forward | ✓ | ✓ | ✗ | ✓ |
| Multi-reference | ✗ | ✗ | - | ✓ |
| Single-stage train | ✗ | ✗ | ✗ | ✓ |

Table 1: FPRF provides a photorealistic, feed-forward style transfer method for large-scale neural radiance fields, while supporting multiple style reference images in the stylization stage. Our model, FPRF, differs from competing methods, where it needs only an efficient single stage training, and it works with arbitrary style images in a feed-forward manner.

Recent studies (Chen et al. 2022; Zhang et al. 2023) have developed methods for 3D scene PST and shown plausible visual qualities. However, all of them require additional time-consuming learning and optimization stages even after scene reconstruction (Chen et al. 2022) or tedious per-style optimization steps to apply even just a single style to the scene (Zhang et al. 2023). More importantly, due to the aforementioned drawbacks, the previous 3D scene PST methods do not scale to large 3D scenes. This motivates us to develop an efficient PST method for large-scale 3D scenes that stylizes the whole 3D scene without exhaustive and time-consuming per-scene or per-style optimization.

In this work, we propose FPRF, an efficient and feed-forward PST for large-scale 3D scenes. To implement a feed-forward PST method for a large-scale 3D scene, we employ the adaptive instance normalization (AdaIN) method, which has shown efficient and promising style transfer results on various tasks (Huang and Belongie 2017; Huang et al. 2018; Gunawan et al. 2023; Wang et al. 2020; Aberman et al. 2020; Segu et al. 2020; Huang et al. 2022; Liu et al. 2023). Specifically, we propose a photorealistic *stylizable radiance field* consisting of a scene *content* field and a scene *semantic* field. Given a large set of photos of the target scene, we first train a grid-based scene content field to embed the scene geometry and content features that can be later decoded to a large-scale

radiance field of arbitrary styles. The scene semantic field is trained together to match proper local styles to the local scene. After obtaining the scene content field, our FPRF stylizes the whole 3D scene via AdaIN, that manipulates the scene content field with the style reference images' feature statistics in a feed-forward manner without any nuisance optimization.

In addition to efficiency, stylizing a large-scale scene requires dealing with diverse objects and contents that are hard to cover with a single reference image. Also, it is challenging to identify a single reference that can effectively encompass the entire semantics of a large-scale scene. Thus, extending the existing single reference-based methods *e.g.*, Chen et al. (2022), is not straightforward, due to diverse contents in the large-scale scene. To overcome this challenge, we introduce a style dictionary module and style attention for efficiently retrieving the style matches of each local part of the 3D scene from a given set of diverse style references. The proposed style dictionary consists of pairs of a local semantic code and a local style code extracted from the style references. To form a compact style dictionary, we cluster similar styles and semantics and use the centroids of the clusters as the dictionary's elements, notably reducing the computational complexity of the style retrieval. Using a style dictionary, we find semantic correspondences between the local semantic codes and the style semantic field.

Our experiments demonstrate that FPRF obtains superior stylization qualities for both large/small-scale scenes with multi-view consistent and semantically matched style transfer. Furthermore, our model demonstrates versatility compared to the previous methods by stylizing 3D scenes with multiple style references, which is not feasible with other prior methods. Our main contributions are summarized as follows:

- We propose a stylizable radiance field where we can perform photorealistic style transfer in a feed-forward manner with an efficient single-stage training.
- We propose the style dictionary and its style attention for style retrieval, which allows us to deal with multiple style references via semantic correspondence matching.
- To the best of our knowledge, our work is the first multi-reference based 3D PST without any per-style optimization, which is scalable for large-scale 3D scenes.

## Related Work

Our task relates to large-scale neural scene reconstruction and photorealistic style transfer for large-scale 3D scenes via neural feature distillation.

**Large-scale neural 3D scene reconstruction.** Realistic 3D reconstruction of large-scale scenes has been considered an important task, which could be a stepping stone to achieve a comprehensive 3D scene understanding and immersive virtual reality. Recently, a few seminal studies (Tancik et al. 2022; Turki, Ramanan, and Satyanarayanan 2022; Zhenxing and Xu 2022) tackled the task by leveraging advances in neural radiance fields (Mildenhall et al. 2020), showing remarkable reconstruction quality for large-scale 3D scenes.

Prior arts mainly focused on decomposing the large-scale scene into smaller parts, *i.e.*, divide-and-conquer. Block-NeRF (Tancik et al. 2022) manually divided city-scale scenes into block-level, and Switch-NeRF (Zhenxing and Xu 2022) learns scene decomposition using a learnable gating network. Although they can reconstruct large scenes by decomposition, they take a long time to train and inference by using a structure composed of a large neural network (Mildenhall et al. 2020; Barron et al. 2021). Instead of adopting time-consuming MLP-based architecture, we leverage lightweight and easy-to-learn K-planes (Fridovich-Keil et al. 2023) for representing large-scale 3D scenes and build our system, FPRF. Specifically, using an efficient blockwise decomposed K-planes representation, we learn not only the scene geometry and radiance field, but also learn per-3D point high dimensional features for the large-scale scene, which we call the *stylizable radiance field*. With the stylizable radiance field and our multi-view consistent MLP color decoder, FPRF embeds compatibility for stylizing large-scale scenes with any style reference images while preserving the high-fidelity reconstruction quality for the original scene even without post-optimization.

**Photorealistic style transfer for 3D scenes.** 3D scene style transfer aims to stylize 3D scene according to the style of the reference image while preserving multi-view consistency. Recently, Chiang et al. (2022); Huang et al. (2022); Fan et al. (2022); Nguyen-Phuoc, Liu, and Xiao (2022); Chen et al. (2022); Zhang et al. (2022); Liu et al. (2023); Zhang et al. (2023) combine 3D neural radiance field with style transfer.

Among them, UPST-NeRF (Chen et al. 2022) and LipRF (Zhang et al. 2023) tackled PST on 3D neural radiance field. UPST-NeRF constructs a stylizable 3D scene with a hyper network, which is trained on stylized multi-view images of a single scene. They can stylize a trained scene in a feed-forward manner with arbitrary style. However, the model requires additional time-consuming ($>$10 hrs.) per-scene optimization after scene reconstruction. LipRF stylizes the reconstructed 3D scene with multi-view stylized images by 2D PST methods. They achieved 3D PST by leveraging Lipschitz network (Virmaux and Scaman 2018) to moderate artifacts and multi-view inconsistency caused by 2D PST methods. However, LipRF requires a time-consuming iterative optimization procedure for every unseen reference style. One of the artistic style transfer methods, StyleRF (Liu et al. 2023) leverages the 3D feature field to stylize 3D scenes with artistic style. They achieve feed-forward style transfer with distilled 3D feature field, however, it requires per-scene multi-stage training which spends more time than reconstruction ($>$5hrs.). Our model mitigates aforementioned inefficiency by constructing stylizable 3D radiance field with only a single training stage spending about 1 hr. Also, our work distinctively focuses on photorealistic stylization with the capability of referring multi-style references, while StyleRF focuses only on artistic stylization with a single reference. Table 1 compares the distinctiveness of our work with the prior work.

## Method

In this section, we introduce FPRF, a feed-forward Photorealistic Style Transfer (PST) method for large-scale 3D scenes. We first introduce a single-stage training of the stylizable

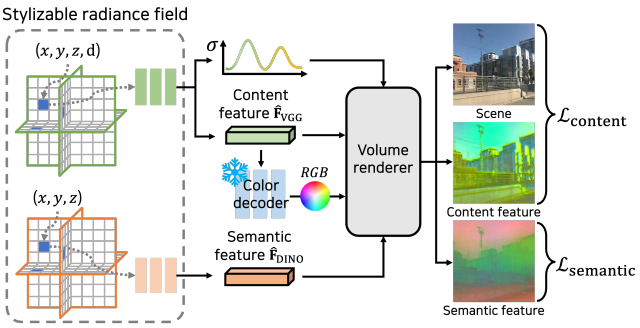


Figure 1: FPRF training stage. Given a set of scene images and corresponding VGG and DINO features, FPRF learns the stylizable radiance field. Stylizable radiance field embeds the geometry, radiance field, and semantic features of the scene. Note that FPRF only needs the original scene images during training, not the stylized images, while it can take arbitrary style images in the stylization stage.

radiance field using AdaIN. We further describe the scene semantic field for stylizing large-scale scenes with multiple reference images.

## Large-scale Stylizable Radiance Field

Our goal is to construct a 3D large-scale radiance field that can be stylized with reference images in a feed-forward manner while achieving photorealistic results after stylization. For the feed-forward style transfer, we employ adaptive instance normalization (AdaIN) (Huang and Belongie 2017), which is an efficient style reference-based style transfer method, where it linearly replaces the content image's feature statistics with a given style image's feature as:

$$\text{AdaIN}(\mathbf{F}(c), \mathbf{F}(s)) = \sigma_s \frac{(\mathbf{F}(c) - \mu_c)}{\sigma_c} + \mu_s, \qquad (1)$$

where $\mathbf{F}(c)$ and $\mathbf{F}(s)$ are semantic features extracted from content and style images by a pre-trained feature encoder, *e.g.*, a VGG encoder (Simonyan and Zisserman 2014). The feature statistics $\mu_*$ and $\sigma_*$ over spatial axes are the mean and standard deviation of the extracted features, respectively. The AdaIN layer is favorable to enable feed-forward style transfer. This property is particularly useful for dealing with large-scale 3D scenes and enables a fast and seamless stylization. To leverage AdaIN's statistic-based style transfer to the 3D scene, we propose to reconstruct a stylizable 3D neural radiance field by distilling the 2D features onto the field, called the stylizable radiance field.

**Stylizable radiance field.** To build a multi-view consistent stylizable feature field, we apply multi-view bootstrapping (Simon et al. 2017) to our domain. We distill high-dimensional features obtained from 2D input images into neural feature fields that models the large-scale 3D scene. To represent a large-scale 3D scene, we extend K-planes (Fridovich-Keil et al. 2023) with the block-composition manner (Tancik et al. 2022), which is used for embedding volumetric scene geometry, radiance, and semantic features. Specifically, we design the stylizable radiance field with two tri-plane grids: scene *content* field and scene *semantic* field. The scene semantic field embeds semantic features of a scene, which will be discussed later in Sec. . The scene content field is responsible for embedding accurate scene geometry and appearance-related content features. Given a 3D scene point position $\mathbf{x}=(x, y, z)$ and a ray direction vector $\mathbf{d}$ as inputs, our scene content field outputs the density, and content feature of the query 3D point (see Fig. 1).

The original NeRF computes a pixel's RGB color $\hat{\mathbf{C}}(\mathbf{r})$ by accumulating the color $\hat{\mathbf{c}}_i$ and density $\sigma_i$ of sampled 3D points $\mathbf{x}_i$ along the ray $\mathbf{r}$. Correspondingly, we train to render high-dimensional features of 3D points in the scene to a pixel value $\hat{\mathbf{F}}(\mathbf{r})$ by accumulating features $\hat{\mathbf{f}}_i$ along the ray $\mathbf{r}$:

$$\hat{\mathbf{F}}(\mathbf{r}) = \sum_{i=1}^{K} \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) (1 - \exp(-\sigma_i \delta_i)) \hat{\mathbf{f}}_i, \qquad (2)$$

where $\delta_i$ denotes the distance between the sampled point $\mathbf{x}_i$ and the next sample point, $\exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$ is a transmittance to the point $\mathbf{x}_i$, and $1 - \exp(-\sigma_i \delta_i)$ represents the absorption by $\mathbf{x}_i$. We distill 2D image features to a 3D scene by minimizing the error between the volume-rendered features and the 2D features extracted from input images, which we call feature distillation. To distill highly-detailed features into the 3D scenes, we use the refined 2D features by Guided Filtering (He, Sun, and Tang 2012) before distillation.

**Generalizable pre-trained MLP color decoder.** The output of the scene content field requires a separate decoder to decode the stylized feature into the color. One may naïvely train such a decoder scene-by-scene, which suffers from a limited generalization to unseen colors and requires additional training for stylization. In the following, we present a scene-agnostic and pre-trained MLP color decoder $D_{\text{VGG}}$ compatible with AdaIN. Specifically, $D_{\text{VGG}}$ transforms the distilled VGG features into colors. To perform style transfer in a feed-forward manner, we pre-train $D_{\text{VGG}}$ with the diverse set of content images (Lin et al. 2014) and style images (Nichol 2016), which enable $D_{\text{VGG}}$ to be generalized to arbitrary style reference images. For that, we employ content loss $\mathcal{L}_c$ and the style loss $\mathcal{L}_s$, similar to Huang and Belongie (2017): $\mathcal{L}_{D_{\text{VGG}}} = \mathcal{L}_c + \lambda_s \mathcal{L}_s$. The main differences with the training process of AdaIN are threefold. First, we use the MLP architecture instead of the CNN upsampling layer inducing multi-view inconsistency. Also, we employ features from the shallower layer, `ReLU2_1` of VGGNet, that contain richer color information. Furthermore, the features from the input content images are upsampled to the pixel resolution for per-pixel decoding. When training the stylizable radiance field, we fix the pre-trained $D_{\text{VGG}}$ so that it preserves the knowledge about the distribution of VGG features from diverse images, which induces compatibility with AdaIN.

**Training scene content field.** We train the scene content field by optimizing the tri-plane grid features and MLP. In detail, for the input 3D point $\mathbf{x}_i$ and the view direction vector $\mathbf{d}$, the grid features of the scene content field are decoded into the 3D point density $\sigma_i$ and the content feature (see the green grid and MLP in Fig. 1). We further decode the content feature into RGB values using the pre-trained $D_{\text{VGG}}$. At the

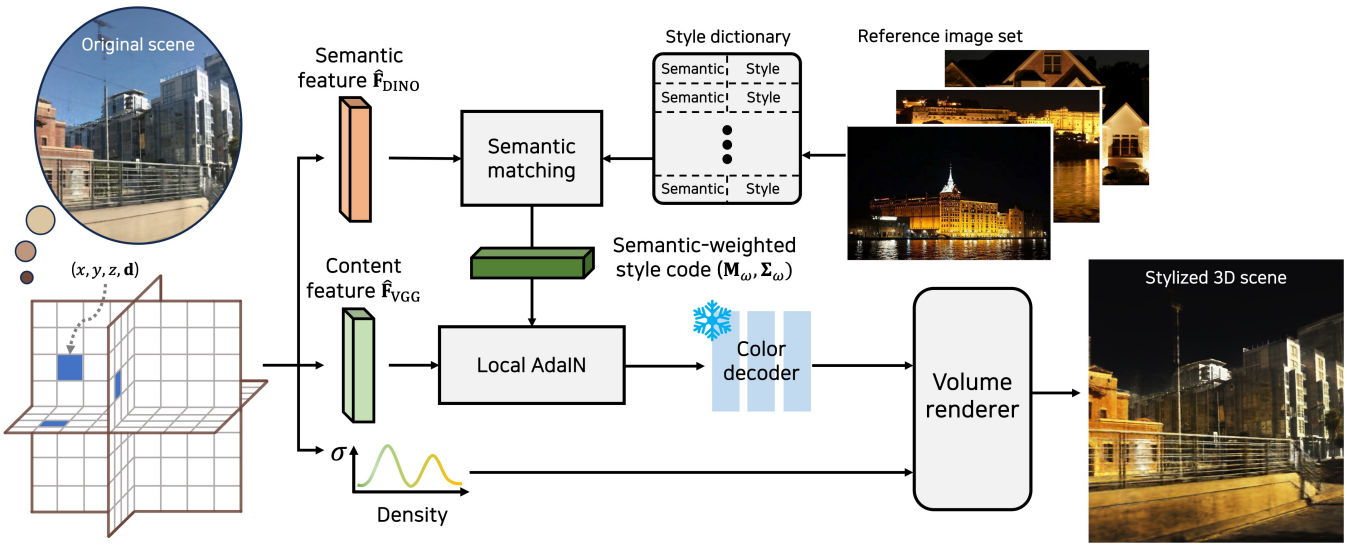


Figure 2: FPRF stylization stage. Given the optimized stylizable radiance field and the set of arbitrary reference images, we stylize the large-scale 3D scene via our novel semantic-aware local AdaIN. We compose a style dictionary consisting of local semantic codes and local style codes pairs extracted from the clustered reference images. Using semantic features from the stylizable radiance as a query, we find the corresponding local semantic features and retrieve the paired local style codes. Using the retrieved semantic-style code pairs, we perform semantic matching and local AdaIN, then finally render the stylized colors.

end, we render the content feature $\hat{\mathbf{F}}_{\text{VGG}}(\mathbf{r})$ and the scene color $\hat{\mathbf{C}}(\mathbf{r})$ via volume rendering (Eq. (2)).

To perform AdaIN, we guide the content feature map $\hat{\mathbf{F}}_{\text{VGG}}$ to follow VGG feature distribution by feature distillation. We distill the ground truth features $\mathbf{F}_{\text{VGG}}(\mathbf{I})$ obtained from input images $\mathbf{I}$ via pre-trained VGGNet, by minimizing the error between $\mathbf{F}_{\text{VGG}}(\mathbf{I})$ and the volume rendered features $\hat{\mathbf{F}}_{\text{VGG}}(\mathbf{r})$. Note that the ground truth VGG feature maps are upsampled to pixel resolution with guided filtering. Since the scene content field needs to reconstruct the accurate scene geometry and appearance, we compute the photometric loss for the volume rendered color $\hat{\mathbf{C}}(\mathbf{r})$. Also, typical regularization losses (Fridovich-Keil et al. 2023) are employed. The total loss function for training the scene content field is as below:

$$\mathcal{L}_{\text{content}} = \sum_{\mathbf{r}\in\mathcal{R}} \|\hat{\mathbf{F}}_{\text{VGG}}(\mathbf{r}) - \mathbf{F}_{\text{VGG}}(\mathbf{I}, \mathbf{r})\|_2^2 + \lambda_{\text{RGB}} \sum_{\mathbf{r}\in\mathcal{R}} \|\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{I}, \mathbf{r})\|_2^2 + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \quad (3)$$

where $\mathcal{R}$ is the set of sampled rays in each training batch, and $\mathbf{F}_{\text{VGG}}(\mathbf{I}, \mathbf{r})$ and $\mathbf{C}(\mathbf{I}, \mathbf{r})$ denote ground truth VGG features and RGB values of the pixels correspond to the ray $\mathbf{r} \in \mathcal{R}$.

Note that we keep $D_{\text{VGG}}$ frozen after its pre-train stage, *i.e.*, $D_{\text{VGG}}$ is fixed during the training stage of the scene content field. This differs from StyleRF (Liu et al. 2023), which needs to fine-tune a CNN-based decoder for each scene.

**Feed-forward stylization using the scene content field.** After training the scene content field, we can perform PST with an arbitrary style image in the stylization stage. In other words, we train the scene content field once and perform PST on a trained radiance field in a feed-forwards manner, *i.e.*, without per-style or per-scene optimization.

Given a reference image $\mathbf{I}_s$, we start with extracting the VGG feature $\mathbf{F}_{\text{VGG}}(\mathbf{I}_s)$. Then we stylize 3D content features $\hat{\mathbf{F}}_{\text{VGG}}$ as $\sigma_s(\hat{\mathbf{F}}_{\text{VGG}} - \mu_c)/\sigma_c + \mu_s$, where $\mu_s$ and $\sigma_s$ are the mean and standard deviation of $\mathbf{F}_{\text{VGG}}(\mathbf{I}_s)$. By decoding the stylized content features to RGB values using the pre-trained color decoder $D_{\text{VGG}}$, we can render a stylized 3D scene. The mean and standard deviation of $\hat{\mathbf{F}}_{\text{VGG}}(\mathbf{r})$ are keep tracked with the moving average during training (Liu et al. 2023).

## Multi-reference Image 3D Scene PST via Semantic Matching and Local AdaIN

With the trained scene content field and AdaIN, we can efficiently transfer the style of images to the 3D radiance field. However, it often fails to produce satisfying results when it comes to large-scale 3D scenes: AdaIN only allows a single reference image which often cannot cover all components in the large-scale scene. To overcome this limitation, we propose a multi-reference based 3D scene PST by semantically matching the radiance field and multiple reference images. As shown in the Fig. 2, the process involves two steps. First, we compose a style dictionary containing local semantic-style code pairs obtained from semantically clustered reference images. Then, we perform a semantic-aware style transfer by leveraging the semantic correspondence between the 3D scene and each element of the composed style dictionary.

**Reference image clustering.** To stylize a 3D scene with multiple style reference images, we consider a set of reference images, $\mathcal{I}_s = \{\mathbf{I}_s^i\}_{i=1,\dots,N}$, where $N$ denotes the number of reference images we use. We compose a compact style dictionary $\mathcal{D}$ with the reference images by clustering them

with similar styles and semantics. We first extract semantic feature maps to cluster the reference images according to semantic similarity. We employ DINO (Caron et al. 2021) as a semantic feature encoder, which can be generalized to various domains by being trained on large-scale datasets in a self-supervised manner. We then apply K-means clustering to the extracted semantic feature map $\mathbf{F}_{\text{DINO}}(\mathbf{I}_s^i)$, and obtain semantically correlated $M$ number of clusters $\mathcal{S}=\{\mathbf{S}^{ij}\}_{i=1,\dots,N}^{j=1,\dots,M}$ from each reference image $\mathcal{I}_s$.

We obtain local style codes from the clusters by extracting another feature map $\mathbf{F}_{\text{VGG}}(\mathbf{I}_s^i)$ from each reference image with VGGNet (Simonyan and Zisserman 2014). Then we obtain the local style code, mean $\mu_s^{ij}$ and standard deviation $\sigma_s^{ij}$, from VGG features $\mathbf{F}_{\text{VGG}}(\mathbf{I}_s^{ij}) \in \mathbf{S}^{ij}$ assigned to each cluster. The centroid $\bar{\mathbf{f}}_{\text{DINO}}(\mathbf{I}_s^{ij})$ of each clustered semantic feature and the assigned local style code $(\mu_s^{ij}, \sigma_s^{ij})$ compose a key-value pair for the style dictionary $\mathcal{D}$, as $\mathcal{D}_{ij}=\{\bar{\mathbf{f}}_{\text{DINO}}(\mathbf{I}_s^{ij}):(\mu_s^{ij}, \sigma_s^{ij})\}_{i=1,\dots,N}^{j=1,\dots,M}$. With this compact style dictionary, we can efficiently perform local style transfer using multiple reference images by semantically matching the clusters with the 3D scene.

**Scene semantic field.** To semantically match the 3D scene and the reference image clusters, we design and learn an auxiliary 3D feature grid, called scene semantic field. The scene semantic field contains semantic features of the 3D scene (Kobayashi, Matsumoto, and Sitzmann 2022; Tschernezki et al. 2022; Kerr et al. 2023). To distill semantic features to the scene semantic field, we optimize the tri-plane features and MLP (orange grid and MLP in Fig. 1) by minimizing the error between rendered features $\hat{F}_{\text{DINO}}(\mathbf{r})$ and features extracted from the input images by DINO as follows:

$$\mathcal{L}_{\text{semantic}} = \sum_{\mathbf{r}\in\mathcal{R}} \|\hat{\mathbf{F}}_{\text{DINO}}(\mathbf{r}) - \mathbf{F}_{\text{DINO}}(\mathbf{I}, \mathbf{r})\|_1, \tag{4}$$

where $\mathbf{F}_{\text{DINO}}(\mathbf{I}, \mathbf{r})$ denotes ground truth DINO features matched to ray $\mathbf{r}$. The density $\sigma$ from the scene content field is used for volume rendering, and $\mathcal{L}_{\text{semantic}}$ does not affect the learning of the density. We do not query view direction as input, in order to preserve multi-view semantic consistency. Also, for constructing a fine-grained scene semantic field, the ground truth DINO feature maps are refined by guided filtering (He, Sun, and Tang 2012). The guided filtering enables consistent distillation of semantic features, resulting in clean and photorealistic stylized outputs (see Fig. 3).

**Semantic correspondence matching & Local AdaIN.** When rendering the stylized 3D scene, we use two semantic feature matrices for computing semantic correspondence between the 3D scene and the elements of the style dictionary $\mathcal{D}$. The first one is $\hat{\mathbf{F}}_{\text{DINO}} \in \mathbb{R}^{K\times C_D}$, obtained from the scene semantic field, where $K$ is the number of queried 3D points, and $C_D$ is the channel size of the semantic feature, *i.e.*, DINO feature. The other one is $\bar{\mathbf{F}}_{\text{DINO}}(\mathbf{I}_s) \in \mathbb{R}^{T\times C_D}$ comprising keys $\bar{\mathbf{F}}_{\text{DINO}}(\mathbf{I}_s^{ij})$ of $\mathcal{D}$, which are the centroids of the $T$ clusters. Given semantic feature matrices, $\hat{\mathbf{F}}_{\text{DINO}}$ and $\bar{\mathbf{F}}_{\text{DINO}}(\mathbf{I}_s)$, we compute a cross-correlation matrix $\mathbf{R}$ as:

$$\mathbf{R} = \hat{\mathbf{F}}_{\text{DINO}}\, \bar{\mathbf{F}}_{\text{DINO}}(\mathbf{I}_s)^\top. \tag{5}$$

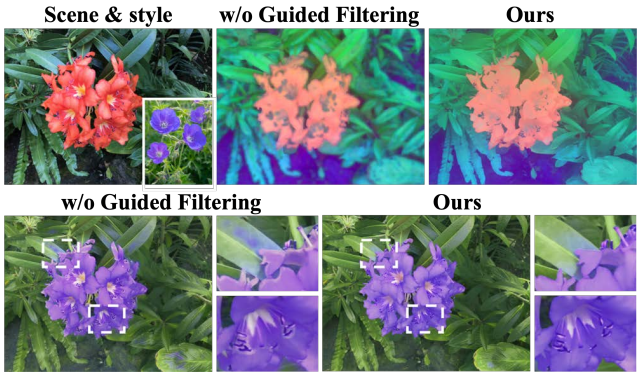


Figure 3: Effects of guided filtering on semantic features. [Top] Given the trained 3D scene and the reference image (left), we visualize the learned stylizable radiance field without (mid) / with (right) guided filtering. The learned semantic features are much sharper when guided filtering is applied. [Bottom] The stylizable radiance field shows degraded stylization results if learned without guided filtering, *e.g.*, blurry boundaries (left), higher stylization quality when learned with guided filtering (right).

For mapping local styles of the reference images according to $\mathbf{R}$, we compose two matrices, $\mathbf{M}(\mathbf{F}_{\text{VGG}}(\mathbf{I}_s)) \in \mathbb{R}^{T\times C_V}$ and $\mathbf{\Sigma}(\mathbf{F}_{\text{VGG}}(\mathbf{I}_s)) \in \mathbb{R}^{T\times C_V}$, comprising of the local style codes $(\mu_s^{ij}, \sigma_s^{ij})$ from the style dictionary, where $C_V$ denotes the channel size of VGG feature map. We compute the matrix form of semantic-weighted style codes $(\mathbf{M}_w, \mathbf{\Sigma}_w)$ as follows:

$$\begin{aligned} \mathbf{M}_w &= \mathbf{R}^{\text{S}}\mathbf{M}(\mathbf{F}_{\text{VGG}}(\mathbf{I}_s)) \\ \mathbf{\Sigma}_w &= \mathbf{R}^{\text{S}}\mathbf{\Sigma}(\mathbf{F}_{\text{VGG}}(\mathbf{I}_s)), \end{aligned} \tag{6}$$

where $\mathbf{R}^{\text{S}} = \texttt{Softmax}(\mathbf{R})$ demonstrates style attention assigned to the queried 3D point features. The semantic-weighted style code $(\mathbf{M}_w, \mathbf{\Sigma}_w)$ are assigned to each 3D point according to the style attention. We feed these semantic-weighted style codes to AdaIN layer as $\mathbf{\Sigma}_w(\hat{\mathbf{F}}_{\text{VGG}} - \mu_c)/\sigma_c + \mathbf{M}_w$, and perform volumetric rendering to obtain final stylized scene renderings. Note that this semantic-aware local AdaIN preserves the multi-view stylized color consistency by directly measuring semantic correspondence between reference images and features on the scene semantic fields (Kobayashi, Matsumoto, and Sitzmann 2022).

We found that computing $\mathbf{M}_w$ and $\mathbf{\Sigma}_w$ on the feature map resolution without clustering (Gunawan et al. 2023) is inefficient for the volume rendering framework where iterative rendering of rays is inevitable. Our clustering enables efficient style transfer, especially for cases using multiple reference images. We highlight that photorealistic scene stylization results are obtained when $\sim$10 clusters are used. It is worth noting that the clustering process takes no more than 1 *sec.* for each reference image. Using the small number of clusters allows us to avoid iterative concatenation of high-dimensional reference image features and effectively reduces the computational cost of matrix multiplication.

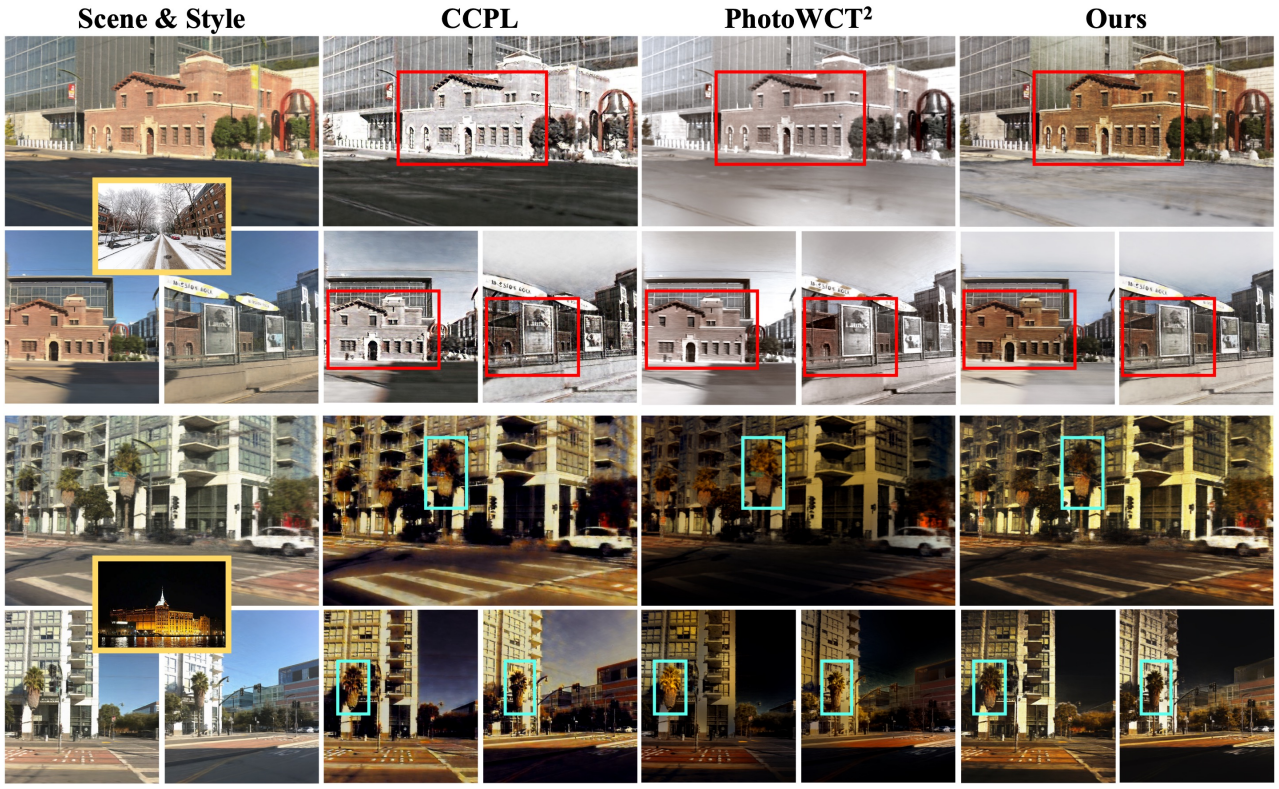


Figure 4: Multi-view appearance consistency on the San Francisco Mission Bay dataset (Tancik et al. 2022). FPRF preserves multi-view appearance consistency even in extreme viewpoint change, while 2D PST methods (Wu et al. (2022); Chiu and Gurari (2022)) produce inconsistent colors of the same building as the viewpoint changes.

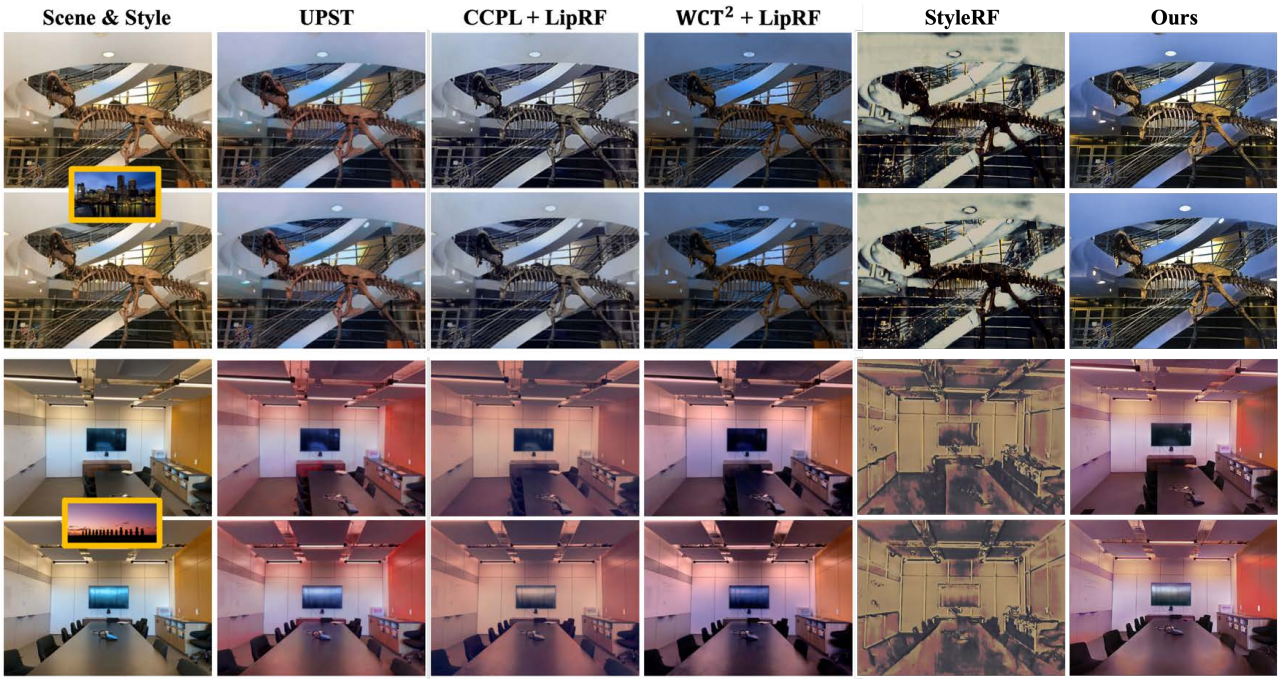


Figure 5: PST quality comparison on the LLFF dataset (Mildenhall et al. 2019). Compared to the competing 3D PST methods, FPRF stylizes the radiance field in a photorealistic manner by transferring the diverse color of the reference image while preserving the original images' naturalness and vividness.

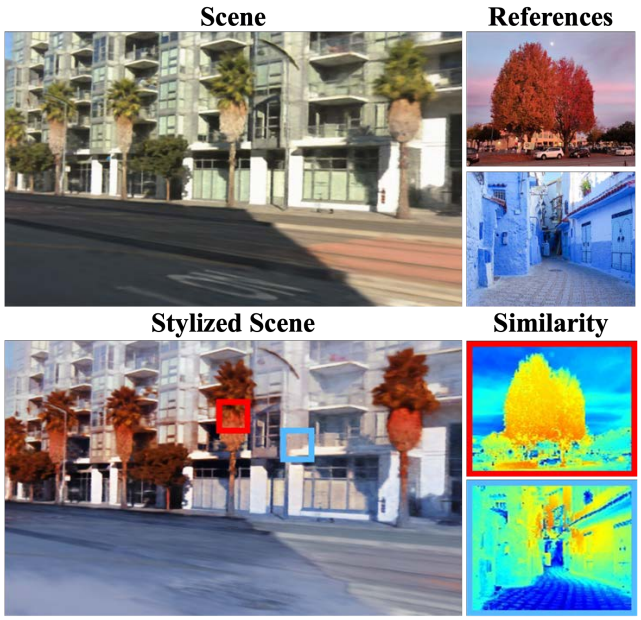


Figure 6: Multi-reference style transfer. FPRF stylizes the 3D radiance field with multiple reference images. Each heatmap shows the similarity between the semantic features of a highlighted patch and the reference image. Our model comprehends the semantic relationship of a large-scale 3D scene and matches the scene with the reference images.

## Experiments

In this section, we demonstrate the qualitative results of FPRF on the large- and small-scale scenes, and ablation studies.

**PST on large-scale scenes.** For large-scale scenes, we consider the San Francisco Mission Bay dataset (Tancik et al. 2022), a city scene dataset consisting of about 12,000 images recorded by 12 cameras. Since we propose the first method aiming for large-scale 3D scene PST, no 3D PST method supports large-scale datasets. Therefore, we compare FPRF against two competing 2D PST methods, CCPL (Wu et al. 2022) and PhotoWCT$^2$ (Chiu and Gurari 2022). Figure 4 illustrates a qualitative comparison on large-scale scenes. We observe that 2D PST methods fail to preserve multi-view color consistency under wide-range view changes, *e.g.*, they obtain different colors of the same building and sky as the viewpoints change. Also, they stylize images without understanding of semantic relation between the content image and the reference image. Instead, our FPRF elaborately transfers styles by reflecting the semantic correspondence between scene and reference images. Semantic matching can also preserve multi-view consistency by directly measuring semantic correspondence between reference images and 3D points in the 3D scene semantic field directly.

**PST on small-scale scenes.** For small-scale scenes, we consider LLFF (Mildenhall et al. 2019), which includes eight real forward-facing scenes. We compare our model with two competing 3D PST methods; LipRF (Zhang et al. 2023), UPST-NeRF (Chen et al. 2022), and the most recent feed-forward 3D *artistic* style transfer method, StyleRF (Liu et al. 2023). As shown in Fig. 5, our method transfers the diverse colors of the style reference image well while maintaining the texture fidelity of the original scene. While UPST-NeRF effectively retains the structure of the scene, the stylized scene's color style differs from the reference. LipRF stabilizes the artifacts caused by 2D PST methods with the Lipschitz network, but it tends to oversmooth the scene and loses the color diversity of the reference image. Noticeably, StyleRF fails to obtain photorealistic results and loses the detailed structure of the original scene. Note that UPST-NeRF and StyleRF require per-scene optimization after scene reconstruction for style transfer, and LipRF needs per-style optimization. On the contrary, our model achieves feed-forward PST after an efficient single-stage learning of the stylizable radiance field.

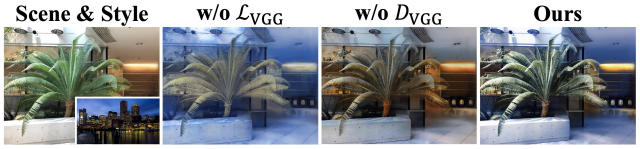


Figure 7: Ablation studies for the VGG feature distillation loss and the pre-trained color decoder. "w/o $\mathcal{L}_{\text{VGG}}$" trains model without the VGG feature distillation loss $\mathcal{L}_{\text{VGG}}$. "w/o $D_{\text{VGG}}$" replaces the pre-trained color decoder $D_{\text{VGG}}$ to an MLP which is trained from scratch during scene optimization.

**Multi-reference style transfer.** Figure 6 shows the style transfer result using multiple reference images. Our model effectively selects suitable styles from multiple references with scene semantic field. Semantic similarity is computed by multiplying features from the scene semantic field with DINO (Caron et al. 2021) feature maps extracted from reference images. The similarity map clearly shows that our model comprehends the accurate semantic relationship between scenes and reference images.

**Ablation studies.** Figure 7 shows the effects of the VGG feature distillation loss and the pre-trianed color decoder. The VGG feature distillation loss helps preserve the original scene content and improve the quality of style transfer, by guiding the content feature to follow VGG feature distribution. The generalizability of the pre-trained color decoder allows the model to get arbitrary style reference images as input.

## Conclusion

In this paper, we present FPRF, a novel stylizable 3D radiance field aiming large-scale 3D scene photorealistic style transfer (PST). FPRF allows feed-forward PST after only a single-stage training by leveraging AdaIN. FPRF also supports multi-reference style transfer, which allows stylizing large-scale 3D scenes which consist of diverse components.

The current limitation is that the semantic matching performance of our model is bounded by the capability of the semantic image encoder, DINO. Nonetheless, since our model can utilize any semantic encoder for constructing the scene semantic field, the performance of our model stands to benefit from the emergence of more advanced models.

## Acknowledgements

This work was supported by the LG Display (2022008004), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00290,Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense; No.RS-2023-00225630, Development of Artificial Intelligence for Text-based 3D Movie Generation; No.2021-0-02068, Artificial Intelligence Innovation Hub; No. 2019-0-01906, Artificial Intelligence Graduate School Program(POSTECH)).

## References

Aberman, K.; Weng, Y.; Lischinski, D.; Cohen-Or, D.; and Chen, B. 2020. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4): 64–1.

Agarwal, S.; Snavely, N.; Simon, I.; Seitz, S. M.; and Szeliski, R. 2009. Building Rome in a day. In *IEEE International Conference on Computer Vision (ICCV)*.

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 5855–5864.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *IEEE International Conference on Computer Vision (ICCV)*.

Chen, Y.; Yuan, Q.; Li, Z.; Xie, Y. L. W. W. C.; Wen, X.; and Yu, Q. 2022. 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.

Chiu, T.-Y.; and Gurari, D. 2022. Photowct2: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fan, Z.; Jiang, Y.; Wang, P.; Gong, X.; Xu, D.; and Wang, Z. 2022. Unified implicit neural stylization. In *European Conference on Computer Vision*, 636–654. Springer.

Fridovich-Keil, S.; Meanti, G.; Warburg, F. R.; Recht, B.; and Kanazawa, A. 2023. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; and Kanazawa, A. 2022. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5501–5510.

Früh, C.; and Zakhor, A. 2004. An automated method for large-scale, ground-based city model acquisition. *International Journal of Computer Vision*, 60: 5–24.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gunawan, A.; Kim, S. Y.; Sim, H.; Lee, J.-H.; and Kim, M. 2023. Modernizing Old Photos Using Multiple References via Photorealistic Style Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12460–12469.

He, K.; Sun, J.; and Tang, X. 2012. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(6): 1397–1409.

Huang, X.; and Belongie, S. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In *IEEE International Conference on Computer Vision (ICCV)*.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, 172–189.

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.

Jun-Seong, K.; Yu-Ji, K.; Ye-Bin, M.; and Oh, T.-H. 2022. Hdr-plenoxels: Self-calibrating high dynamic range radiance fields. In *European Conference on Computer Vision*, 384–401. Springer.

Kerr, J.; Kim, C. M.; Goldberg, K.; Kanazawa, A.; and Tancik, M. 2023. Lerf: Language embedded radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*.

Kobayashi, S.; Matsumoto, E.; and Sitzmann, V. 2022. Decomposing nerf for editing via feature field distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, K.; Zhan, F.; Chen, Y.; Zhang, J.; Yu, Y.; El Saddik, A.; Lu, S.; and Xing, E. P. 2023. StyleRF: Zero-shot 3D Style Transfer of Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8338–8348.

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. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision (ECCV)*.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4): 1–15.

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

Nichol, K. 2016. Painter by numbers, Wikiart, 2016. *URL https://www. kaggle. com/c/painter-by-numbers/overview*.

Pollefeys, M.; Nistér, D.; Frahm, J.-M.; Akbarzadeh, A.; Mordohai, P.; Clipp, B.; Engels, C.; Gallup, D.; Kim, S.-J.; Merrell, P.; et al. 2008. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78: 143–167.

Segu, M.; Grinvald, M.; Siegwart, R.; and Tombari, F. 2020. 3dsnet: Unsupervised shape-to-shape 3d style transfer. *arXiv preprint arXiv:2011.13388*.

Simon, T.; Joo, H.; Matthews, I.; and Sheikh, Y. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Snavely, N.; Seitz, S. M.; and Szeliski, R. 2006. Photo tourism: Exploring photo collections in 3D. In *SIGGRAPH Conference Proceedings*.

Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.; Srinivasan, P.; Barron, J. T.; and Kretzschmar, H. 2022. Block-NeRF: Scalable Large Scene Neural View Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tschernezki, V.; Laina, I.; Larlus, D.; and Vedaldi, A. 2022. Neural Feature Fusion Fields: 3D distillation of self-supervised 2D image representations. In *International Conference on 3D Vision (3DV)*, 443–453. IEEE.

Turki, H.; Ramanan, D.; and Satyanarayanan, M. 2022. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12922–12931.

Virmaux, A.; and Scaman, K. 2018. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31.

Wang, J.; Wen, C.; Fu, Y.; Lin, H.; Zou, T.; Xue, X.; and Zhang, Y. 2020. Neural Pose Transfer by Spatially Adaptive Instance Normalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, Z.; Zhu, Z.; Du, J.; and Bai, X. 2022. CCPL: contrastive coherence preserving loss for versatile style transfer. In *European Conference on Computer Vision*, 189–206. Springer.

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, Z.; Liu, Y.; Han, C.; Pan, Y.; Guo, T.; and Yao, T. 2023. Transforming Radiance Field with Lipschitz Network for Photorealistic 3D Scene Stylization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20712–20721.

Zhenxing, M.; and Xu, D. 2022. Switch-NeRF: Learning Scene Decomposition with Mixture of Experts for Large-scale Neural Radiance Fields. In *International Conference on Learning Representations (ICLR)*.

Zhu, S.; Zhang, R.; Zhou, L.; Shen, T.; Fang, T.; Tan, P.; and Quan, L. 2018. Very large-scale global sfm by distributed motion averaging. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4568–4577.