TeSTNeRF: Text-Driven 3D Style Transfer via Cross-Modal Learning

Jiafu Chen[1] , **Boyan Ji**[1] , **Zhanjie Zhang**[1] , **Tianyi Chu**[1] , **Zhiwen Zuo**[2] ,
Lei Zhao[1*] , **Wei Xing**[1*] and **Dongming Lu**[1]

[1]Zhejiang University
[2]Zhejiang Gongshang University

{chenjiafu, ji_by, cszzj, chutianyi, cszhl, wxing, ldm}@zju.edu.cn, zzw@zjgsu.edu.cn

Abstract

Text-driven 3D style transfer aims at stylizing a scene according to the text and generating arbitrary novel views with consistency. Simply combining image/video style transfer methods and novel view synthesis methods results in flickering when changing viewpoints, while existing 3D style transfer methods learn styles from images rather than texts. To address this problem, we for the first time design an efficient text-driven model for 3D style transfer, named TeSTNeRF, to stylize the scene using texts via cross-modal learning: we leverage an advanced text encoder to embed the texts in order to control 3D style transfer and align the input text and output stylized images in latent space. Furthermore, to obtain better visual results, we introduce style supervision, learning feature statistics from style images and utilizing 2D stylization results to rectify abrupt color spill. Extensive experiments demonstrate that TeSTNeRF significantly outperforms existing methods and provides a new way to guide 3D style transfer.

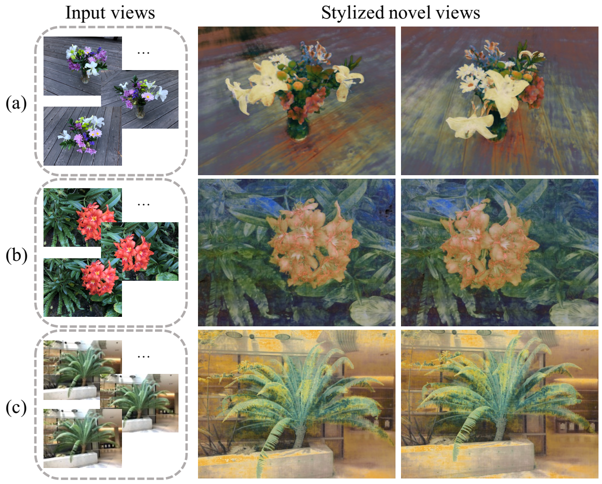

Figure 1: Results of text-driven 3D style transfer by our method. (a) is driven by text **"Paul Cezanne"**, (b) is driven by text **"Monet"**, and (c) is driven by text **"Van Gogh"**. Given a set of real photographs and a text, our method is capable of generating stylized novel views, which are consistent in 3D space.

1 Introduction

Given a collection of artworks, learning their internal expressions of art (such as color tones, strokes) and applying them to 3D scenes is a meaningful yet challenging task. From a practical point of view, artistic scene creations resembling style of various artists can be toured on visual reality (VR) and augmented reality (AR) devices, which provides users with a more intuitive understanding of artists' styles. One possible solution to 3D style transfer is to directly apply image/video stylization techniques to 3D scenes. However, these methods lack 3D scene perception. Without considering the underlying 3D structure, such methods may cause short-range flickering or long-range discontinuity when changing viewpoints.

The major challenge for 3D style transfer is to maintain consistency among different viewpoints to produce coherent results. Recently, Neural Radiance Field (NeRF) [Mildenhall *et al.*, 2020] has shown superior performance in reconstructing 3D objects and scenes and novel view synthesis. An implicit neural scene can be estimated from a few image observations. NeRF uses MLP to regress both volume density and color. Some subsequent works [Schwarz *et al.*, 2020; Niemeyer and Geiger, 2021; Gu *et al.*, 2022] focus on disentangling shape and appearance to manipulate them. Nevertheless, they are mainly designed for objects, which have difficulty in stylizing complex 3D scenes.

To address the problem, a number of works [Chiang *et al.*, 2022; Huang *et al.*, 2022; Zhang *et al.*, 2022; Nguyen-Phuoc *et al.*, 2022] investigate adopting NeRF for style transfer. Once trained, [Chiang *et al.*, 2022] is capable to transfer arbitrary style to 3D scenes, while [Zhang *et al.*, 2022] and [Nguyen-Phuoc *et al.*, 2022] need to re-train the network every time given a new style image. These methods use an image for style reference. However, to achieve artist-aware stylization, it is more convenient to use a text for reference, as one reference image is not representative enough. Thus, we tend to guide the stylization with texts instead of images.

In this paper, we propose TeSTNeRF, a novel method for text-driven 3D style transfer. On the basis of a pre-trained

[*]Corresponding authors.

NeRF where the 3D scene has already been reconstructed, we introduce latent codes to represent styles of different artists. Moreover, we leverage CLIP [Radford *et al.*, 2021], a large cross-modal vision-language model, to align the input text and output stylized images. We also enhance the stylization by learning style feature statistics from image collection corresponding to the text. To generate more harmonious stylized scenes, we utilize results from 2D stylization approaches to alleviate abrupt color spill. Our experiments show that the proposed approach can produce different results by using distinct texts, which are also view-consistent.

In summary, our main contributions are threefold:

- To the best of our knowledge, we for the first time propose a novel approach to stylize 3D scene according to a given text via cross-modal learning, which can produce consistent novel views of high visual quality.
- We introduce latent codes to control styles of different artists, simplifying the representation of style domain. To establish the connection between texts and latent codes better, we adopt CLIP as the encoder.
- We conduct rectification of abrupt color spill utilizing 2D stylization results, which helps generate more harmonious visual results.

2 Related Work

2.1 Representing Scenes with Neural Field

In the past few years, rendering 3D scenes implicitly via neural networks [Jiang *et al.*, 2020; Genova *et al.*, 2020; Park *et al.*, 2019; Riegler and Koltun, 2020; Mildenhall *et al.*, 2020] has gained much concern. Among them, NeRF [Mildenhall *et al.*, 2020] has achieved incredibly high-quality results in reconstructing 3D objects and scenes. Given a few images, NeRF encodes a continuous neural radiance field to render photo-realistically novel views. The success of NeRF has inspired various follow-up works, extending NeRF to generative models [Schwarz *et al.*, 2020; Niemeyer and Geiger, 2021; Gu *et al.*, 2022], decomposing rendering of scenes [Boss *et al.*, 2021; Martin-Brualla *et al.*, 2021; Yang *et al.*, 2021], etc. We leverage NeRF as our backbone to learn volume density and view-dependent color for the scene, which can also be extended to better-quality NeRFs [Zhang *et al.*, 2020; Barron *et al.*, 2021].

2.2 2D Style Transfer

2D style transfer includes image style transfer and video style transfer. Image style transfer is to combine style feature from an image and content feature from another image, whose style is similar to the former and content the latter. Many works [Huang and Belongie, 2017; Li *et al.*, 2017] investigate approaches to perform the combination. To represent the full scope of artistic style, learning style from a collection of artworks, which forms a style domain, is another perspective of image style transfer. AST [Sanakoyeu *et al.*, 2018] manages to extract shared qualities among a group of artworks. DualAST [Chen *et al.*, 2021] develops a scheme to learn simultaneously both the holistic artist-style and the specific artwork-style via its proposed Style-Control Block. StyleBank [Chen *et al.*, 2017b] utilizes multiple convolution filter banks, each of which explicitly represents one style domain. Moreover, image translation approaches, like CycleGAN [Zhu *et al.*, 2017], can also deal with collection style transfer. Some other works [Kwon and Ye, 2022; Fu *et al.*, 2021]learn styles from text descriptions. TxST [Liu *et al.*, 2022] proposes to embed an image-text model.

Directly applying image style transfer techniques to video frame-by-frame usually causes instability and flickering. Video style transfer [Chen *et al.*, 2017a; Gao *et al.*, 2018] tackles this problem by introducing optical flow or aligning intermediate feature to constrain nearby video frames. MCCNet [Deng *et al.*, 2021] rearranges style representations based on content representations to make style patterns suitable for content structures. With the help of temporal regularization, ReReVST [Wang *et al.*, 2020] reconciles the contradiction between style transfer and temporal consistency.

2.3 3D Style Transfer

3D style transfer has higher requirements compared to video style transfer, since novel view synthesis is required in 3D scenes. Previous methods extend stylization to 3D scenes by representing 3D scenes with point clouds [Huang *et al.*, 2021] or meshes [Yin *et al.*, 2021]. Due to the success of NeRF, some approaches explore 3D style transfer based on NeRF. Stylizing-3D-Scene [Chiang *et al.*, 2022] develops a hypernetwork to control the appearance-related weights of the NeRF model. StylizedNeRF [Huang *et al.*, 2022] proposes a mutual learning strategy for NeRF and 2D stylization method. SNeRF [Nguyen-Phuoc *et al.*, 2022] deals with the memory limitation of training with a whole image in NeRF and improves the visual quality by alternating between stylization and NeRF training. ARF [Zhang *et al.*, 2022] aims to transfer detailed style features via matching features between the style image and the scene. These methods guide the stylization with images. Instead, we focus on realizing 3D style transfer under the guidance of texts.

2.4 CLIP-based NeRFs

Recently, a model based on Contrastive Language-Image Pretraining (CLIP) [Radford *et al.*, 2021] learns a latent space, which may be used to estimate the similarity between a text and an image and promote the development of cross-modal learning. The powerful representation learned by CLIP narrows the gap between texts and images. CLIP-NeRF [Wang *et al.*, 2022] leverages CLIP to flexibly control the 3D content through texts or images, providing an interactive approach to manipulate the shape and appearance of 3D objects. Dream Fields [Jain *et al.*, 2022] generates objects via NeRF with the help of CLIP. LaTeRF [Mirzaei *et al.*, 2022] uses CLIP to remove artifacts while extracting objects that is partly occluded from a scene. DFFs [Kobayashi *et al.*, 2022] distills the knowledge of CLIP-LSeg, a model for segmentation, into a 3D feature field to semantically decompose 3D scenes.

We propose to utilize CLIP for stylization in 3D scenes. Given texts, we use the text embeddings encoded by CLIP to manipulate which style domain the scene should be transferred to and align the input text and output stylized images in CLIP latent space.

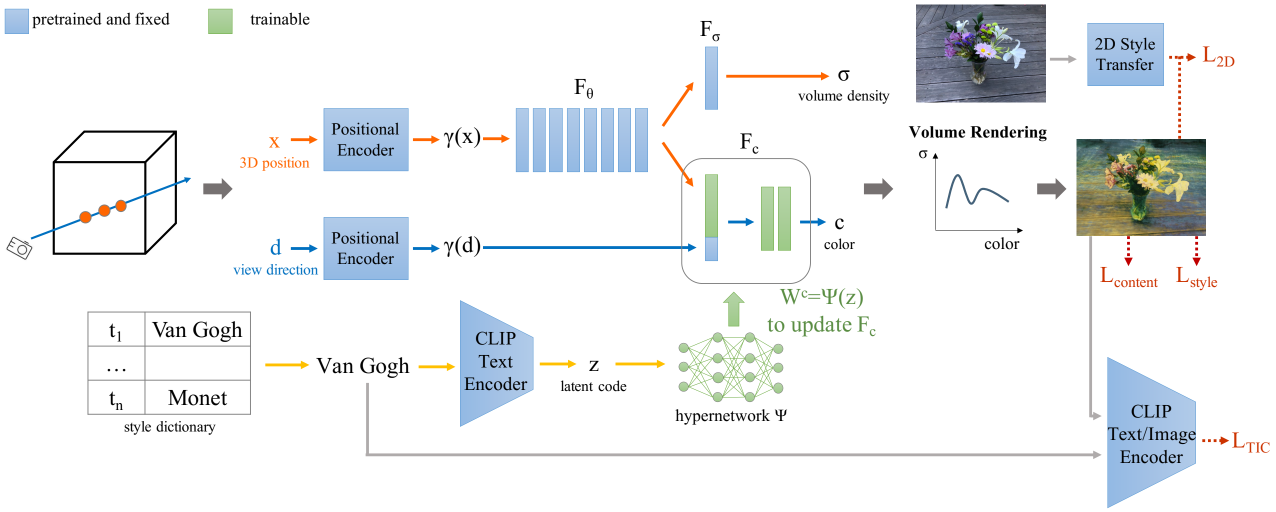

Figure 2: TeSTNeRF model overview. Based on pre-trained NeRF reconstructing the scene, we perform stylization using a text from the style dictionary. The text is then projected to the latent space of CLIP, guiding hypernetwork Ψ to predict the parameters W^c of F_c. Finally, stylized images are generated via volume rendering. The objective functions $L_{content}$, L_{style}, L_{2D}, and L_{TIC} are used for constraining the generated results.

3 Preliminary

NeRF [Mildenhall *et al.*, 2020] adopts multi-view images of a 3D scene to optimize the underlying continuous radiance field, using MLP to output volume density σ and color $\mathbf{c}$ given a point coordinate $\mathbf{x}$ and its view direction $\mathbf{d}$. The whole networks can be disentangled into three parts: F_θ, F_σ, and F_c. Practically, volume density and color can be calculated separately as:

$$\sigma = F_\sigma(F_\theta(\mathbf{x})), \ \mathbf{c} = F_c(F_\theta(\mathbf{x}), \mathbf{d}). \tag{1}$$

During volume rendering, a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is cast from the center $\mathbf{o}$ of the camera along the direction $\mathbf{d}$ through a pixel in the image. According to the volume rendering equation, each pixel's color is integrated along the ray as:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt, \tag{2}$$

$$\text{where } T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds), \tag{3}$$

where t_n and t_f represent the near and far bounds of the ray. The image taken from a specific viewpoint is finally generated by aggregating its pixels according to Eq. 2 and 3.

4 Method

Our goal is to stylize a scene generated by a set of photos from different viewpoints. We carry out the implementation based on pre-trained implicit reconstruction of the scene using NeRF. During stylization, we fix the parameters of F_θ and F_σ to maintain the geometry of the scene, and only optimize F_c to change the appearance. Existing methods [Chiang *et al.*, 2022; Huang *et al.*, 2022; Zhang *et al.*, 2022; Nguyen-Phuoc *et al.*, 2022] achieve stylization under the guidance of style images. Our proposed TeSTNeRF manages to stylize the scene with texts as conditional inputs. Please refer to Fig. 2 for an overview of our proposed framework.

4.1 Cross-modal Text-driven Style Transfer

Given paintings from different artists, it is not a difficult task for experienced people to distinguish which painting belongs to whom. Thus, the artists' name can be used as a high-level representation, denoting different style domains. Inspired by TxST [Liu *et al.*, 2022], we leverage a latent space with different latent codes to control different style domains. Although it is possible to distinguish style domains with one-hot encoding, one-hot encoding may lead to poor performance since its representation lacks concrete meaning in latent space. To directly handle a text t as input, we project t into CLIP text latent space. With the powerful representation of CLIP's text embedding, the latent code z guides the optimization of hypernetwork Ψ, which predicts the parameters W^c to update F_c. In this way, we utilize CLIP embedding as style condition to manipulate the appearance of a scene.

To ensure the generated images containing the same content as the origin images of the scene, we introduce content loss. Features produced by pre-trained VGG-19 network [Simonyan and Zisserman, 2014] ϕ can effectively capture intrinsic representation of an image, which is often used in image style transfer. Denoting $I_{content}$ as the ground truth image and I_{out} as the stylized image, we compute the content loss as:

$$L_{content} = \|\phi_4(I_{content}) - \phi_4(I_{out})\|_2, \tag{4}$$

where ϕ_4 denotes the relu4_1 layer in VGG-19.

Cross-modal learning aims to effectively utilize the correlation of different modal contents for modeling. Since CLIP has made a significant breakthrough in evaluating the similarity of image-text pairs, we carry out our cross-modal learning based on CLIP. For an image and a text, the similarity of their features is proportional to the probability of the text being associated with the image. Thus. we define a text-image cross-modal loss L_{TIC} to shorten the CLIP-space distance between

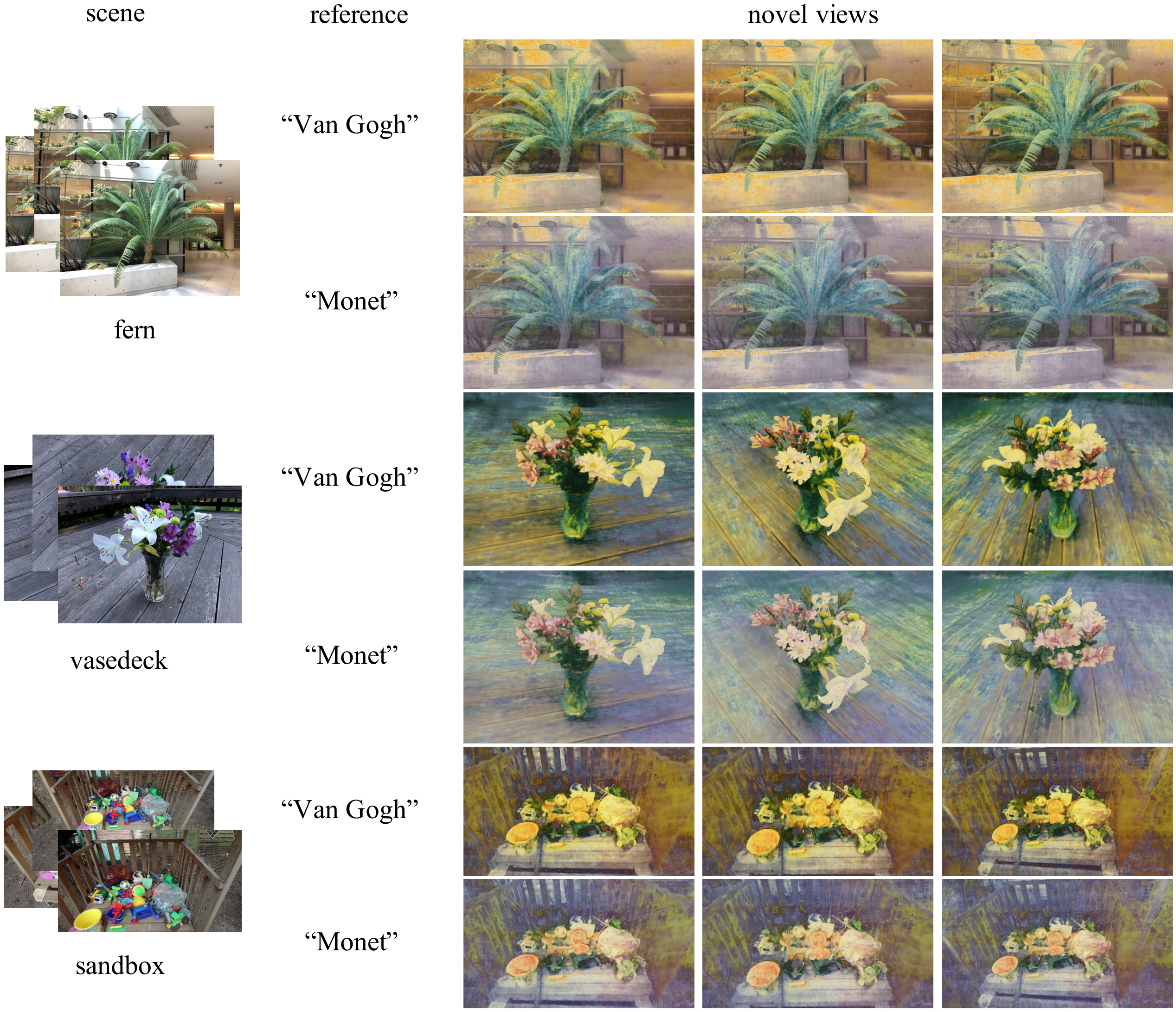

Figure 3: Qualitative results generated by our proposed TeSTNeRF. The first column shows various scenes. The second column shows the texts, which are taken as the style references. The rest of the images are novel views of difference styles in corresponding scenes generated by our model.

the input text t and the output stylized image I_{out}:

$$L_{TIC} = 1 - \langle \zeta_i(I_{out}),\ \zeta_t(t) \rangle, \tag{5}$$

where ζ_i and ζ_t are pre-trained CLIP image encoder and text encoder, respectively, and $\langle , \rangle$ denotes cosine similarity. Through training with $L_{content}$ and L_{TIC}, we are capable to generate stylized scene corresponding to the input text.

4.2 Style Supervision

Although cross-modal learning can produce stylized results, the results are not visually-pleasant enough, as shown in Fig. 6. To tackle this problem, we introduce style loss to assist in better learning the style feature, which is usually used in image style transfer. We utilize a collection of style images corresponding to the text, where a style image is randomly chosen each iteration. Style loss L_{style} measures the error between feature statistics of the style image I_{style} and the stylized image I_{out} as:

$$\begin{aligned} L_{style} = & \sum_i \|\mu(\phi_i(I_{style})) - \mu(\phi_i(I_{out}))\|_2 \\ & + \sum_i \|s(\phi_i(I_{style})) - s(\phi_i(I_{out}))\|_2, \end{aligned} \tag{6}$$

where μ and s are channel-wise mean and standard deviation, respectively. ϕ_i denotes a layer in VGG-19 used to compute the style loss. In our experiments, we use relu1_1, relu2_1, relu3_1, and relu4_1 layers.

Previous domain-based 2D style transfer approaches [Sanakoyeu *et al.*, 2018; Zhu *et al.*, 2017] generate globally color harmonious results due to their effective learning from a large scale of images. They learn a mapping between a source domain and a target domain,

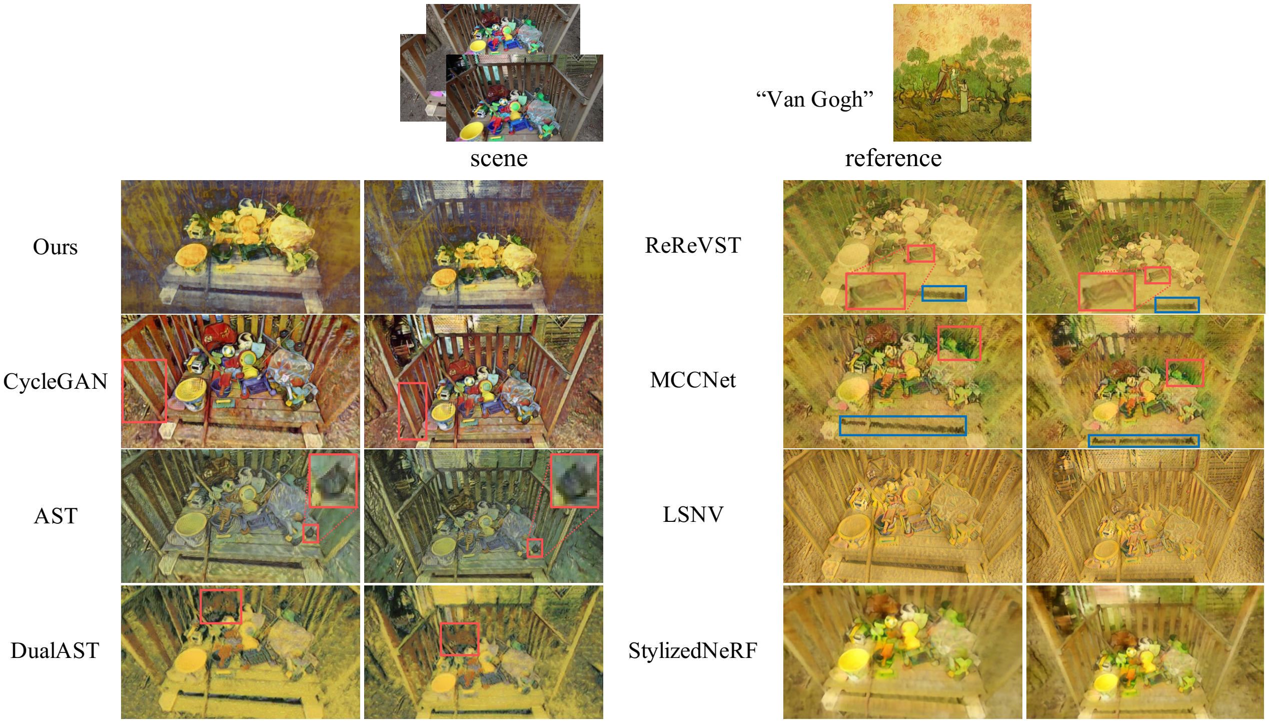

Figure 4: Qualitative comparisons. Our method takes the text "Van Gogh" as the reference style. For CycleGAN and AST, we use the provided models pret-rained on a collection of Van Gogh's artworks to generate the stylized results of novel views. For the rest methods, we choose a reference style image as shown in the first row to guide the stylization. Inconsistencies are highlighted in red boxes.

generating results that are indistinguishable from the target domain. To rectify color spill when training with limited data in a scene, we leverage 2D stylization to supply additional supervision. We adopt CycleGAN [Zhu *et al.*, 2017] to provide the 2D stylization supervision. Specifically, we utilize their pre-trained collection style transfer models. It should be noted that CycleGAN is a representative domain-based style transfer method, which may be replaced by other advanced methods. 2D stylization supervision is introduced via computing L_1 loss between the corresponding result generated by 2D style transfer method I_{2D} and 3D output stylized image I_{out}:

$$L_{2D} = \|I_{2D} - I_{out}\|_1. \tag{7}$$

The loss L_{2D} constrains the output of 3D style transfer to alleviate the influence of color spill from irrelevant object in the scene, thus leads to better visual quality, as we will later demonstrate in Fig. 6.

4.3 Total Loss

In summary, the content loss $L_{content}$ is to make sure that the generated images contain the same content as the origin images of the scene, while L_{TIC} is utilized to align the input text and output images in a pre-trained latent space. Meanwhile, the style loss L_{style} is to guide the model to learn from a specific collection of images corresponding to the input text, and L_{2D} is leveraged to mitigate the abrupt color spill, yielding high-quality stylization results. The final loss function L used to train our model is:

$$\begin{aligned} L = \lambda_{content} L_{content} + \lambda_{TIC} L_{TIC} \\ + \lambda_{style} L_{style} + \lambda_{2D} L_{2D}, \end{aligned} \tag{8}$$

where the constants $\lambda_{content}$, λ_{TIC}, λ_{style}, and λ_{2D} are hyper-parameters of the model.

5 Experiment

5.1 Implementation Details

Following NeRF [Mildenhall *et al.*, 2020], we use separate coarse and fine sampling strategies, both sampling 64 points each ray. To represent a larger region with a limited amount of rays, we generate a $K \times K$ patch $\mathcal{P}$ containing pixel (u, v) using stride c:

$$\mathcal{P} = \{(u + cx,\ v + cy)\}, \tag{9}$$

where $x, y \in \left\{-\frac{K}{2}, ..., \frac{K}{2} - 1\right\}$.

Specifically, we set the patch size $K = 100$ for all experiments. The loss weights in Eq. 8 are set to $\lambda_{content} = 0.1$, $\lambda_{TIC} = 1$, $\lambda_{style} = 1$, $\lambda_{2D} = 0.3$. Our proposed method is optimized using Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and its learning rate starts from $1e^{-4}$. The proposed TeST-NeRF is trained for 1500 iterations on a single NVIDIA RTX 3090 GPU.

Datasets

We utilize scenes provided by NeRF [Mildenhall *et al.*, 2020] and FVS [Riegler and Koltun, 2020]. All of them are captured using handheld cameras in real-scenes.

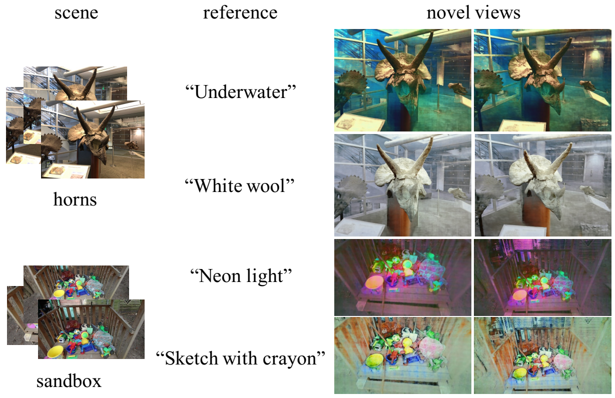

Figure 5: More stylization results generated by TeSTNeRF.

Baselines

We compare TeSTNeRF to SOTA image stylization methods, video stylization methods and 3D stylization methods. Since implementing image style transfer first and then reconstructing the scene is not stable and fails sometimes, which is caused by inconsistent stylization across different views (we also provide the failure case in our supplementary video), we compare our method to the following 3 categories of schemes:

- Novel View Synthesis→Image Style Transfer: we perform novel view synthesis using NeRF, and stylize each new view using CycleGAN, AST, and DualAST.
- Novel View Synthesis→Video Style Transfer: we perform novel view synthesis along a smooth camera path using NeRF, gather the results as a video, and then stylize the video using ReReVST and MCCNet.
- 3D Style Transfer→Novel View Synthesis: we perform 3D scene stylization and then synthesize novel view. Specifically, we conduct a point-cloud-based method LSNV, NeRF-based methods StylizedNeRF and ARF.

For CycleGAN and AST, we use their provided models pre-trained on the corresponding collection of artworks. For the rest methods, we choose a style image from the artist's domain corresponding to the given text.

5.2 Qualitative Results

In Fig. 3, we show our stylization results based on different texts in various scenes. It can be seen that the stylized scenes contain both view-consistent visual quality and satisfying text-driven style controllability.

To validate the superiority of our method, we compare our results with baselines, as shown in Fig. 4. We show results generated by different methods on the scene, Sandbox.

Image style transfer approaches, e.g., CycleGAN, AST, and DualAST, are able to produce high quality stylized results each viewpoint. However, they produce obvious inconsistency artifacts, as highlighted in red boxes.

Although video style transfer approaches, e.g., ReReVST and MCCNet, are able to maintain short-range consistency, they struggle to deal with long-range instability. Moreover, video style transfer methods lack perception of the spatial information in the scene, resulting in violation of geometry of the scene. Please see their failures as reconstructing the sandbox with jagged edges in blue boxes in Fig. 4, while their inconsistencies are highlighted in red boxes as well.

Methods	Fern	Orchids	Vasedeck	Average
CycleGAN	0.0251	0.0184	0.0453	0.0296
AST	0.0315	0.0277	0.0379	0.0307
DualAST	0.0252	0.0247	0.0791	0.0430
ReReVST	0.0058	0.0097	0.0285	0.0147
MCCNet	0.0124	0.0226	0.0679	0.0343
LSNV	0.0173	0.0535	0.0378	0.0362
StylizedNeRF	0.0123	0.0387	0.0178	0.0229
ARF	0.0096	0.0127	0.0152	0.0125
Ours	**0.0043**	**0.0057**	**0.0118**	**0.0073**

Table 1: Short-range consistency comparison. We compare the short-range consistency using warped distance score (the lower the better). **Best** and second best results are marked.

Methods	Fern	Orchids	Vasedeck	Average
CycleGAN	0.0326	0.0326	0.0617	0.0423
AST	0.0370	0.0356	0.0427	0.0384
DualAST	0.0263	0.0309	0.0692	0.0421
ReReVST	**0.0099**	0.0280	0.0373	0.0251
MCCNet	0.0181	0.0431	0.0952	0.0521
LSNV	0.0322	0.1614	0.0231	0.0722
StylizedNeRF	0.0590	0.0720	0.0572	0.0627
ARF	0.0357	0.0392	0.0310	0.0353
Ours	0.0108	**0.0190**	**0.0203**	**0.0167**

Table 2: Long-range consistency comparison. We compare the long-range consistency using warped distance score (the lower the better). **Best** and second best results are marked.

3D style transfer approaches, e.g., LSNV, StylizedNeRF, ARF, and ours, aim to stylize the holistic scene, therefore they all generate view consistent results. However, LSNV cannot capture the reference style well, which affects its visual quality. StylizedNeRF and ARF utilize NeRF as their geometry representation, which is the same as our method. StylizedNeRF and ARF learn style from a reference image, while our method takes in a text as the style reference.

In comparison, we utilize a text instead of an image as the reference style. We produce both view-consistent and high-quality visual results. We encourage readers to have a look at the supplementary video, which shows more intuitive differences between our method and baselines, especially in consistency and stability when changing viewpoints.

We also conduct experiments on more texts than just the artist's name. In this case, we only use $L_{content}$ and L_{TIC} in Eq. 8 to train the network. The result is shown in Fig. 5.

5.3 Quantitative Results

Consistency Measurement

Following the measurement in LSNV, we use a warped LPIPS metric [Zhang *et al.*, 2018] to measure the consistency across different views. Firstly, we utilize FlowNetS [Dosovitskiy *et al.*, 2015] to compute the optical flow from a ground truth image I_x to another I_y. Then a warped mask M is generated according to the optical flow. Finally, we warp the corresponding stylized images $\hat{I}_x$ to $\hat{I}_y$ and calculate their distance

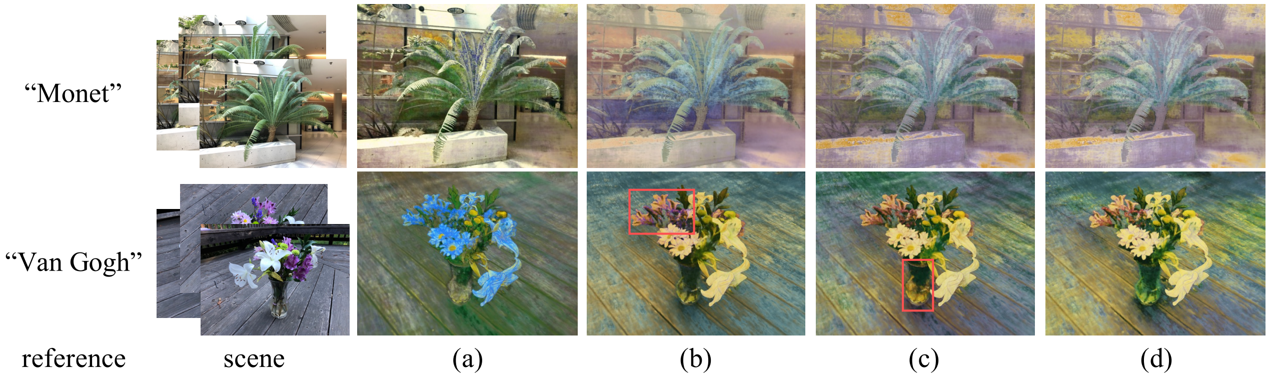

Figure 6: Ablation results. (a) The result of TeSTNeRF without L_{style} and L_{2D}. (b) The result of TeSTNeRF without L_{TIC} and L_{2D}. (c) The result of TeSTNeRF without L_{2D}. (d) The result of full TeSTNeRF.

Methods	Visual Quality		Temporal Consistency	
	mean↑	variance↓	mean↑	variance↓
CycleGAN	5.94	6.14	6.82	4.67
AST	6.88	5.03	6.78	4.49
DualAST	5.87	5.48	5.19	6.20
ReReVST	6.06	4.43	6.42	5.46
MCCNet	5.42	4.04	5.76	3.50
LSNV	6.32	4.01	6.85	2.39
StylizedNeRF	6.61	4.95	7.04	5.90
ARF	6.89	4.25	7.52	1.72
Ours	**7.23**	**3.95**	**8.32**	**1.38**

Table 3: User study. We used "Van Gogh" and "Monet" as text input for style transfer on scene Orchids. We invited users to score the stylized results both in visual quality and consistency, where 10 denotes excellent performance, and 1 denotes poor performance.

along with M. The distance score is formulated as:

$$E(\hat{I}_x, \hat{I}_y) = LPIPS(M \odot Warp(\hat{I}_x, \hat{I}_y)), \qquad (10)$$

where $\odot$ denotes element-wise multiplication.

We compare our method with baselines on three scenes, Fern, Orchids, and Vasedeck, reporting average warped distance score on texts "Van Gogh" and "Monet". For short-range consistency, we randomly choose 20 adjacent novel views $(\hat{I}_t, \hat{I}_{t+1})$ from each scene, as shown in Tab. 1. In Tab. 2, we show the long-range consistency score, randomly choosing 20 frame pairs $(\hat{I}_t, \hat{I}_{t+7})$ from each scene. From these two tables, we observe that TeSTNeRF outperforms baselines in short-range consistency in all scenes, and performs the best or second best as for long-range consistency.

User Study

We also conduct a user study for subjective evaluation. We invite 27 male and 23 female participants to score stylized scenes in visual quality and temporal consistency. We first show a text denoted an artist with three representative artworks from the artist as reference. We carry out the study on scene Orchids, providing its stylized result of each method together with a ground truth video for easy comparison. The participants are required to score the visual quality and temporal consistency, where 10 denotes excellent performance, and 1 denotes poor performance. The results are shown in Tab. 3. We observe that our method outperforms other methods both in visual quality and temporal consistency.

5.4 Ablation Studies

In this section, we explore each component's effect in TeSTNeRF and validate their importance by ablation studies.

With and without style loss. In this study, we explore the effect of including style loss L_{style}. We introduce L_{style} to learn the style feature statistics. The model trained with only $L_{content}$ and L_{TIC} shows stylized result, but its visual quality is dissatisfying, as shown in Fig. 6(a). It demonstrates that introducing style feature statistics is necessary.

With and without text-image cross-modal loss. We introduce L_{TIC} to align the CLIP-space embeddings between the text and the stylized image. In Fig. 6(b), the model trained with only $L_{content}$ and L_{style} presents color supersaturation, with color not existing in the artist domain and in this image especially on flowers highlighted in the red box. Regarding the benefits of aligning texts and images, we conclude that text-image cross-modal loss is worth adopting.

With and without 2D loss. As shown in Fig. 6(c), when training without L_{2D}, there exits color spill from the desk to the vase, highlighted in the red box. To rectify the color spill, we leverage results from 2D stylization to supply additional supervision. From Fig. 6(d), we can see that applying this supervision helps resist with the color spill and provide better visual results.

6 Conclusion

In this paper, we investigate text-driven 3D style transfer and propose TeSTNeRF to stylize scenes. Specifically, we utilize CLIP to encode the input text to a latent code, guiding a hypernetwork to predict the parameters related to appearance of the scene in NeRF model and conduct cross-modal learning by aligning the input text and output results in CLIP latent space. Moreover, adopting style supervision improves the stylized visual quality. The experiments on 3D scenes demonstrate the effectiveness of TeSTNeRF for 3D style transfer according to given texts.

Acknowledgements

This work was supported in part by the projects No. 2021YFF0900604, 2020YFC1523202-2, 19ZDA197, LY21F020005, 2021009, 2019011, 2022C01222, Research and Application Demonstration of Key Technologies in Digital Archaeology of Ningbo Prehistoric Sites, Key Technologies and Product Research and Development for Cultural Relics Protection and Trading Circulation, MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University), National Natural Science Foundation of China (62172365).

References

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

[Boss *et al.*, 2021] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021.

[Chen *et al.*, 2017a] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent online video style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1105–1114, 2017.

[Chen *et al.*, 2017b] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1897–1906, 2017.

[Chen *et al.*, 2021] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Dualast: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 872–881, 2021.

[Chiang *et al.*, 2022] Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, and Wei-Chen Chiu. Stylizing 3d scene via implicit representation and hypernetwork. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1475–1484, 2022.

[Deng *et al.*, 2021] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1210–1217, 2021.

[Dosovitskiy *et al.*, 2015] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[Fu *et al.*, 2021] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven image style transfer. *arXiv preprint arXiv:2106.00178*, 2021.

[Gao *et al.*, 2018] Chang Gao, Derun Gu, Fangjun Zhang, and Yizhou Yu. Reconet: Real-time coherent video style transfer network. In *Asian Conference on Computer Vision*, pages 637–653. Springer, 2018.

[Genova *et al.*, 2020] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020.

[Gu *et al.*, 2022] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022.

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[Huang *et al.*, 2021] Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, and Ming-Hsuan Yang. Learning to stylize novel views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13869–13878, 2021.

[Huang *et al.*, 2022] Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18342–18352, 2022.

[Jain *et al.*, 2022] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022.

[Jiang *et al.*, 2020] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020.

[Kobayashi *et al.*, 2022] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[Kwon and Ye, 2022] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022.

[Li *et al.*, 2017] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

[Liu *et al.*, 2022] Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, and Vicky Kalogeiton. Name your style: An arbitrary artist-aware image style transfer. *arXiv preprint arXiv:2202.13562*, 2022.

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

[Mildenhall *et al.*, 2020] 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 *European conference on computer vision*, pages 405–421. Springer, 2020.

[Mirzaei *et al.*, 2022] Ashkan Mirzaei, Yash Kant, Jonathan Kelly, and Igor Gilitschenski. Laterf: Label and text driven object radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 20–36. Springer, 2022.

[Nguyen-Phuoc *et al.*, 2022] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

[Niemeyer and Geiger, 2021] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.

[Park *et al.*, 2019] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

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

[Riegler and Koltun, 2020] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision*, pages 623–640. Springer, 2020.

[Sanakoyeu *et al.*, 2018] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. A style-aware content loss for real-time hd style transfer. In *proceedings of the European conference on computer vision (ECCV)*, pages 698–714, 2018.

[Schwarz *et al.*, 2020] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.

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

[Wang *et al.*, 2020] Wenjing Wang, Shuai Yang, Jizheng Xu, and Jiaying Liu. Consistent video style transfer via relaxation and regularization. *IEEE Transactions on Image Processing*, 29:9125–9139, 2020.

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

[Yang *et al.*, 2021] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13779–13788, 2021.

[Yin *et al.*, 2021] Kangxue Yin, Jun Gao, Maria Shugrina, Sameh Khamis, and Sanja Fidler. 3dstylenet: Creating 3d shapes with geometric and texture style variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12456–12465, 2021.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

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

[Zhang *et al.*, 2022] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 717–733. Springer, 2022.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.