ERF: Multi-Attention Fusion Artistic Radiance Fields

Kunze Jiang
Jiangxi Normal University
Ziyang Avenue, Nanchang
Nanchang, Jiangxi Province, China
Email: kunzejiang@jxnu.edu.com

Yufan Zhou
Zhejiang University
Yuhangtang Road No.388, Hangzhou,
Zhejiang Province, China
Email: yufan_zhou@zju.edu.cn

Chao Wu(✉)
Zhejiang University
Yuhangtang Road No.388, Hangzhou,
Zhejiang Province, China
Email: chao.wu@zju.edu.cn

***Abstract*—We propose a novel 3D scene stylization method called ERF(Multi-Attention Fusion Artistic Radiance Fields), which achieves artistic style transfer by combining 2D stylized images with the radiance fields of 3D scenes. ERF effectively captures high-frequency complex visual information from stylized images and transfers it to 3D scenes while maintaining multi-view consistency. To generate geometrically and semantically consistent novel view stylization effects while preserving the visibility of the original scene, we introduce the Multi-Scale Attention Module (EMA) and the CLIP module. Experimental results show that compared to existing models, ERF exhibits significantly higher stylization quality and detail expressiveness.**

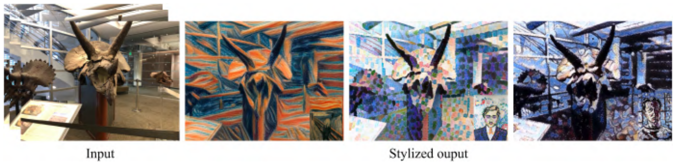

Fig. 1. The ERF achieves artistic style migration by combining 2D stylized images with the radiation field of the 3D scene. It not only stylizes local details, but also enhances global style migration by introducing EMA and CLIP modules, and effectively injects the overall style into the 3D scene through text and images, demonstrating higher stylization quality and detail expression than existing methods.

I. Introduction

To more accurately reconstruct scenes through implicit representation methods, the technology of Neural Radiance Fields (NeRF)[20] emerged. NeRF uses deep neural networks to learn the complex mappings from captured images to new image synthesis, effectively solving challenges in scene reconstruction and rendering. It can quickly and with high quality reconstruct scenes from videos or photos taken with a smartphone. Currently, NeRF has been widely applied and practically implemented in several fields, including Augmented/Virtual Reality[22] and digital human[23].

Style transfer is one of the most popular areas in computer vision, with most work focusing on transferring 2D artistic styles to ordinary images[24]. However, extending 2D style images to 3D scenes is a huge challenge[25]. This extension not only involves complex geometric transformations, but also requires ensuring style consistency with the three-dimensional space. Our proposed ERF is a novel 3D scene stylization method that achieves artistic style transfer by combining 2D stylized images with the radiance fields of 3D scenes. Our method supports reconstructing 3D stylized works with consistent spatial and temporal artistic styles from multiple viewpoints of 2D images.

The method Arf[9] applied nearest-neighbor feature matching(NNFM), which focuses on local image descriptions and captures unique local details. However, this method tends to overlook global style details, thereby reducing the quality of stylized rendering. To address this problem, we propose a multi-scale feature fusion method. This method processes the original image through a parallel structure input into the VGG network, combining local features from multiple input sources to capture spatial information at different scales. By integrating the outputs from different sub-networks, our approach learns more comprehensive global style features while preserving precise spatial structure details, thus enhancing the transfer and expression capabilities of global style features.

To address the problem of loss of stylized details, we extract the CLIP feature space using a pre-trained CLIP module. Specifically, we employ a method we call the mapping network, which transforms the CLIP feature space into content feature space and style feature space. The computed content loss helps ensure that important visual details and content features are preserved during the style transfer process. Additionally, through a weakly supervised approach, the CLIP loss assists the mapping network, making the style transfer process more controllable.

We have demonstrated that ERF can transfer 3D artistic features accurately from diverse and challenging 2D artistic images to various complex 3D scenes. Compared to previous techniques, our method achieves significant improvements in visual quality, and avoids the over-smoothing and blurriness that traditional methods may cause in stylized views. Additionally, our method consistently outperforms the baselines.

In summary, our contributions are:

We appreciate the support from the Ministry of Science and Technology, PRC, Zhejiang Provincial Government, and Zhejiang University.

1. Proposing a multi-scale feature fusion method that captures global style features through a parallel processing mechanism and VGG network, enhancing the detail expression.

2. Utilizing a pre-trained CLIP module and a mapping network to enhance the calculation of content loss, ensuring the preservation of important visual details during the style transfer process.

II. Related Work

A. Style transfer on NeRF

The methods NeRF, Arf and Ref-npr [8][9] rather than the traditional Gram method and instead adopt the NNFM loss function to compute style transfer losses, where this new method generates the target images via referencing any arbitrary style images. However, Desrf[10] points out that the generation performance of the aforementioned models is poor when artistic images have multiple styles. In this case, Desrf not only learns textures but also incorporates the geometric shapes of reference artworks into the 3D scenes. On the other hand, Stylerf[11] addresses the challenges of low geometric quality and insufficient stylization in NeRF style transfer by performing style transformations within NeRF's feature space. CoNeRF[6] introduces a mapping network to map the CLIP feature space to the style feature space. SNeRF[12], through alternating training of stylization optimization steps, addresses the potential issues of jittery artifacts in traditional methods during novel view synthesis and achieves zero-shot style transfer through feature space transformation. Stylizednerf[13] and Arf both employ VGG networks to extract and transform image features for stylization. The difference lies in StylizedNeRF's integration of 2D stylization capability and NeRF's 3D consistency through mutual learning between 2D and 3D.

B. Text-to-image style conversion

The paper on Contrastive Language-Image Pretraining (CLIP) [14]showcases impressive text-image matching capabilities and advanced feature extraction, leading to its widespread adoption in enhancing 3D scene editing within the field. CLIP-NeRF[15] introduces a decoupled conditional NeRF architecture, adjusting shape by learning deformation fields for positional encoding and deferring color adjustments to the volume rendering stage. Additionally, it designs two code mappers, taking CLIP embeddings as input and updating latent codes to reflect target edits. Unlike CLIP-NeRF, LENeRF[16] combines NeRF's 3D prior knowledge with CLIP's multimodal information, creating a framework capable of generating high-fidelity and fine-grained 3D editing results without additional datasets.

Blended-NeRF[17] creates the generation process via a text-image model, where this process incorporates new priors, enhancements, and volume blending techniques to seamlessly integrate new objects into existing NeRF scenes naturally and consistently. Similar to Blended-NeRF, concurrent work Blending-NeRF[18] also employs CLIP models for 3D scene editing. However, unlike Blended-NeRF, the primary goal of Blending-NeRF is to perform local edits on objects within the scene, such as modifying textures and removing parts of the original objects, rather than better integration of the inserted objects into the scene. LERF[19] proposes a method of embedding language into NeRF by rendering CLIP embeddings into training rays volumetrically, and then supervising these embeddings within training views to provide multi-view consistency and smooth underlying language spaces, thereby achieving precise segmentation of objects within the scene.

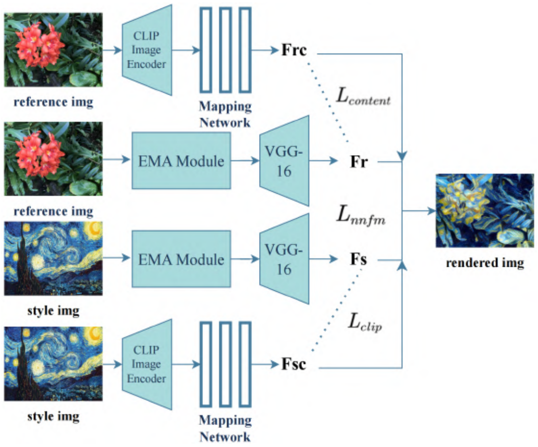

Fig. 2. The workflow of ERF. An EMA module and a CLIP module are used to extracted feature styles to the fusion radiance fields.

III. Method

In this section, we propose a novel style transfer method applied in neural radiance fields(NeRF)[20], i.e., using 2D style images to transform the scene into the style of the images after reconstructing a 3D scene from any given set of photos. Building upon previous 3D stylization efforts, we employ encoders and attention mechanisms to enhance the stylization effect. Our main contribution is to introduce the Multi-Scale Attention Module (EMA) module and the CLIP Mapping Network module, both of which achieve outstanding stylization effects while maintaining the recognizability of the original scene. Fig. 2 illustrates an overview of our proposed method. We will provide a detailed introduction to ERF in this chapter.

A. Neural Radiance Field (NeRF)

To facilitate a better understanding of our approach, we provide a brief introduction to Neural Radiance Fields (NeRF). NeRF offer a novel approach to implicit representation, representing the pioneering use of deep learning methods for reconstructing 3D scenes.

Specifically, the reconstruction method of the radiance neural field involves emitting a ray $r(t)$ into the scene from any observational viewpoint, where the density $\sigma(x)$ can be understood as the probability of the ray terminating at position x. The color $C(r)$ of the camera ray $r(t) = o + td$, with near and far boundaries being t_n and t_f respectively, is:

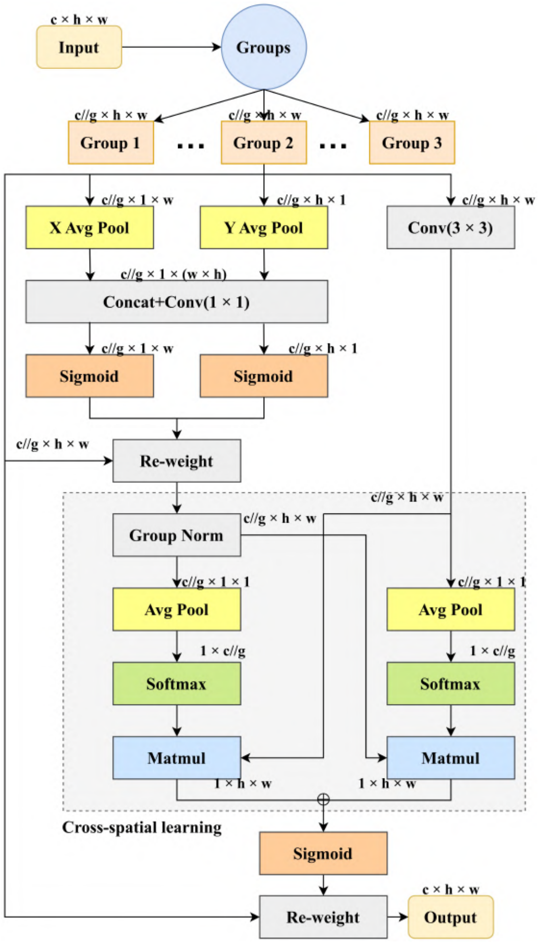

Fig. 3. The EMA splits the input feature map into groups to capture a range of semantic information. It applies parallel sub-networks with large receptive fields for multi-scale spatial feature extraction. This design allows for precise adjustment of channel importance and preservation of spatial details, which is vital for the effective transfer of artistic styles in 3D scene stylization.

$$C(r) = \int_{t_f}^{t_n} T(t)\sigma(r(t))c(r(t), d)dt \tag{1}$$

where $T(t) = exp - \int t_n t\sigma(r(s))ds$

The expression $T(t)$ represents the cumulative transmittance measured along a ray between t_n and t, signifying the likelihood that the ray traverses the distance from t_n to t unobstructed by any particles.

B. Attention-based feature extraction

Artworks often possess unique stylistic forms. Currently, pretrained neural networks can effectively capture these stylistic features and apply the styles of artworks to other 2D images. Additionally, there is ongoing research exploring the application of style transfer methods to the 3D vision. While using VGG16[20] for style transfer on 2D images is a favorable choice, it faces challenges in the 3D domain, as it can only aggregate global information and may not capture image details accurately. To address this challenge, we introduce the Multi-Scale Attention Module (EMA) mechanism. Specifically, the attention module integrates local features from multiple input sources, enhancing model performance through parallel processing and self-attention mechanisms. While maintaining excellent feature representation capabilities, this mechanism improves the model's effectiveness in style transfer tasks.

The EMA attention mechanism first divides the input feature map along the channel dimension into multiple sub-feature groups to capture different semantic information. It employs parallel sub-networks to capture multiscale spatial information. The process of extracting attention weight descriptors involves three parallel pathways, with two in the 1x1 branch and one in the 3x3 branch. The 1x1 branch includes two 1D global average pooling operations for encoding information, followed by the aggregation of two channel attention maps within each group through multiplication. The 3x3 branch uses a single 3x3 convolutional kernel to capture multiscale feature representations. The structure of EMA is depicted in Fig. 3.

Additionally, the EMA mechanism incorporates a cross-space information aggregation strategy to manage feature interactions. These mechanisms effectively enhance the capability of extracting features from style images and original scene features, thus improving fusion effects.

Afterwards, we use VGG to extract feature images. To further ensure that the scenes remain consistent from multiple perspectives, we employ the NNFM mentioned by Arf as the loss function.

We denote P_{style} as the art image, and P_{view} as an image rendered from the radiance field at a selected viewpoint. The VGG feature map M_{style} and M_{view} are extracted for P_{style} and P_{view}, and in particular, we define $P_{view}(i, j)$ as the feature vector at pixel location (i, j) of the feature map M_{view}. The vector $P_{style}(i, j)$ can be defined analogously. In this case, NNFM loss can be written as:

$$\ell_{\text{nfm}}(P_{\text{view}}, P_{\text{style}}) = \frac{1}{N} \sum_{i,j} \min_{i',j'} D(P_{\text{view}}(i, j), P_{\text{style}}(i', j')), \tag{2}$$

where N is the number of pixels in M_{style}, and $D(\eta, \xi)$ is the corresponding cosine distance of two vectors η, ξ where

$$D(\eta, \xi) = 1 - \frac{\eta^T \xi}{\|\eta\| \|\xi\|}. \tag{3}$$

So for each feature in P_{style}, our goal is to minimize its cosine distance (Eq. (3)) to its nearest viewpoint in the style image's VGG feature space (P_{style}).

Note that the optimization result of (eq2) will lead to Overstylization and then make it difficult to identify the content. To address this problem, we add a penalizing function where

$$\ell = \ell_{\text{nfm}}(P_{\text{view}}, P_{\text{style}}) + \delta \ell_2(P_{\text{view}}, P_{\text{original}}). \tag{4}$$

In this equation, δ is the trade-off parameter, and $P_{original}$ is the content of the image in the original scene.

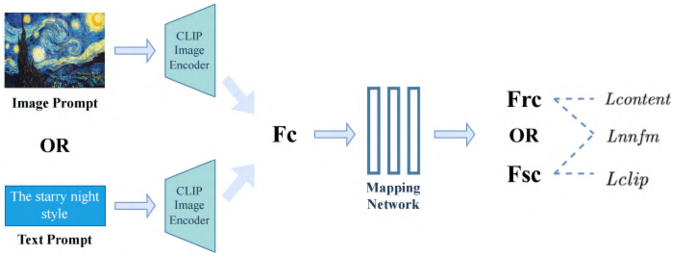

Fig. 4. Feature extraction based on CLIP workflow

C. Feature extraction based on CLIP

CLIP has achieved great success in mapping images and texts to a shared embedding space. Notably, CLIP can extract high-level semantic information from images. In this paper, we use CLIP's encoder as the feature extractor for the second part. By projecting the features of the scene image into the subspace of the style image features, we enhance the effect of style transfer. The CLIP image encoder E_i extracts the feature vector F_c from the style image I_a and we have

$$F_c = E_i(I_a). \tag{5}$$

Subsequently, we utilize the mapping module f_m to establish a correspondence between the CLIP text-image feature vectors F_c and the style feature representation:

$$F_{rc} \quad OR \quad F_{sc} = f_m(F_c) \tag{6}$$

Note that the styles can be represented by mean μ_1 and standard-deviation σ_1, and we have

$$(\sigma_1, \mu_1) = F_{sc}. \tag{7}$$

To refine the CLIP branch, we incorporate style feature loss to evaluate the coherence between the mapping module's output and the VGG branch's style features, i.e.,

$$L_{clip} = \|F_s^v - F_s^c\|_2^2 \tag{8}$$

where

$$F_s^v = (\sigma_v, \mu_v), \quad F_s^c = (\sigma_c, \mu_c). \tag{9}$$

CLIP extracts features from training images and integrates these features to derive multi-space representations, enabling the 3D selector to learn the capability for pixel-level feature querying.

This module consists of two steps. First, we define the mapping from the CLIP space to the style space. Then, we introduce a loss function to calculate the difference between our output and the style features. The integration of these two parts significantly enhances the effect of style transfer.

IV. Experiments

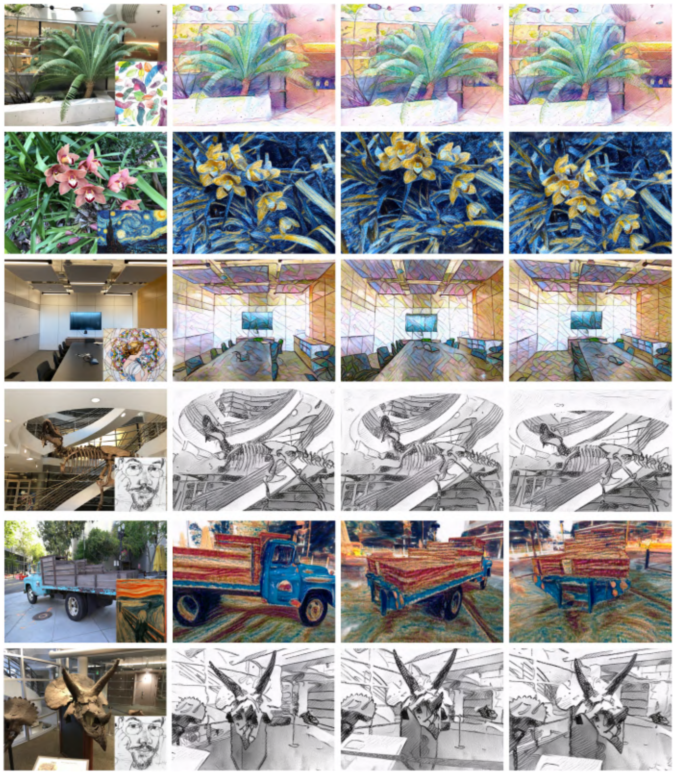

Fig. 5. Results of our ERF in different scenes with different styles.

The goal of this section is to give a thorough evaluation of our ERF method. The details of the experiments and the corresponding results, see Fig. 5, are given in Section IV.A . In Section IV.B, we compared our method with the current state-of-the-art method in 3D scene stylization as in REF-NPR[8], ARF[9], and SNeRF[12]. Also, we demonstrated the effectiveness of our dilated samples.

A. Implementation Details

We use Plenoxels[3] as the based radiation field. We reference method ARF and extract content features from the layer 3 conv of the pre-trained VGG16 model. Also, we use Lstyle with the same settings as in NNST[4] and ARF, as well as the color preservation matrix from ARF.

After updating the style migration network, we observed that setting the content loss parameter $\lambda = 0.1$, Style loss parameter $\mu = 20$, and setting the EMA's $factor = 3$, $channels = 3$ can achieve the best results. All experiments were conducted on a single RTX3090.

B. Comparison

In Fig. 6, we compare our results with current state-of-the-art 3D scene stylization methods. For SNeRF, the synthesized texture is not refined enough and cannot convey complex textures such as the distorted sky in Van Gogh's Starry Night. Also, it cannot synthesize its stylistic features such as the petals of a flower still reference to the color of the image. For Ref-NPR, the style migration of arbitrary styles is very bad, almost like the effect of a stylized image. For ARF, although the result looks more transparent, it learns more about the average texture of the whole image, which is not what we expect. In contrast, our model remembers a globalized style features through the EMA module and achieves semantic

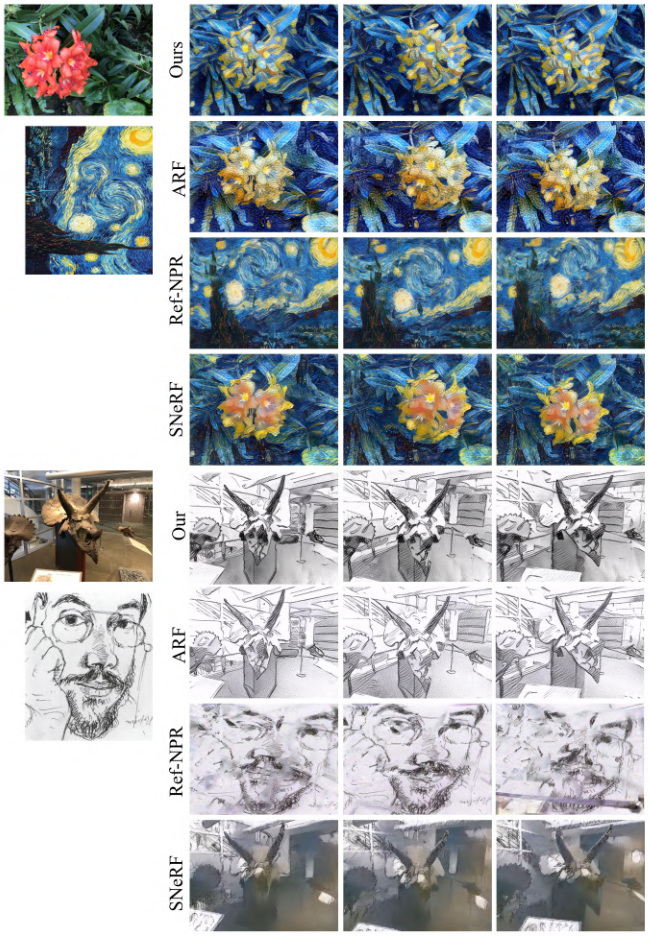

Fig. 6. Comparison Results of our ERF, ARF, Ref-NPR, and SNeRF

weak supervision through the CLIP module with zero-style migration, which has better style migration results on the overall image.

C. Ablation

In order to verify the effectiveness of our proposed EMA and CLIP, we evaluated our method's performance by removing either the EMA or CLIP modules. We observed that without the CLIP module, the resulting model exhibited significant fuzziness, possibly due to the absence of L_{clip} and $L_{content}$. Conversely, upon removing the EMA module, although the images became clearer, we noted a decrease in stylization quality This decrease indicates the importance of the EMA module in maintaining global style coherence.

V. Discussion

Our ERF method has experimentally demonstrated its effectiveness in 3D scene stylization, being able to learn and synthesize novel perspectives with geometric and semantic consistency from 2D stylized images. However, we found in our experiments that the EMA module may lead to loss of stylistic details in some cases. This suggests that we need to further explore how to balance the relationship between the global style migration and the local detail preservation in future work. We also realize that there is some ambiguity in the text-image feature space of CLIP. In future work, we plan to explore more accurate text-image feature mapping methods to improve the accuracy and controllability of text-based style migration.

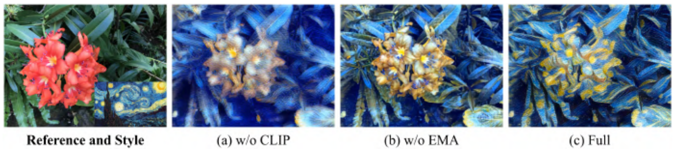

Fig. 7. Our full ERF and the ERF w/o CLIP, the ERF w/o EMA

VI. Conclusion

In this paper, we propose a novel 3D scene stylization method called ERF, which achieves artistic style transformation by combining the radiance fields of a 2D stylized image and a 3D scene. The ERF method utilizes Nearest Neighbor Feature Matching Loss to transfer complex high-frequency visual details from a 2D stylized image to a 3D scene with consistency across multiple viewpoints. In addition, ERF introduces the EMA module and the CLIP module to enhance global style migration and efficiently inject the overall style into the 3D scene using text and images. Experimentally, ERF is able to generate novel viewpoint stylization results with geometric and semantic consistency while maintaining content recognition, demonstrating higher stylization quality and detail expressiveness compared to existing methods.

Acknowledgment

This work was supported by the National Key Research and Development Project of China (2021ZD0110505), the Zhejiang Provincial Key Research and Development Project (2023C01043), and Academy Of Social Governance Zhejiang University.

References

[1] Yu A, Ye V, Tancik M, et al. pixelnerf: Neural radiance fields from one or few images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4578-4587.

[2] Mu, F., Wang, J., Wu, Y., Li, Y.: 3d photo stylization: Learning to generate stylized novel views from a single image. arXiv preprint arXiv:2112.00169 (2021)

[3] Fridovich-Keil S, Yu A, Tancik M, et al. Plenoxels: Radiance fields without neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5501-5510.

[4] Kolkin N, Salavon J, Shakhnarovich G. Style transfer by relaxed optimal transport and self-similarity[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 10051-10060.

[5] Chen, Y., Yuan, Q., Li, Z., Liu, Y., Wang, W., Xie, C., Wen, X., Yu, Q.: Upst-nerf: Universal photorealistic style transfer of neural radiance fields for 3d scene. arXiv preprint arXiv:2208.07059 (2022)

[6] Miao X, Bai Y, Duan H, et al. ConRF: Zero-shot Stylization of 3D Scenes with Conditioned Radiation Fields[J]. arXiv preprint arXiv:2402.01950, 2024.

[7] Ouyang D, He S, Zhang G, et al. Efficient multi-scale attention module with cross-spatial learning[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[8] Zhang Y, He Z, Xing J, et al. Ref-npr: Reference-based non-photorealistic radiance fields for controllable scene stylization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 4242-4251.

[9] Zhang K, Kolkin N, Bi S, et al. Arf: Artistic radiance fields[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 717-733.

[10] Xu S, Li L, Shen L, et al. Desrf: Deformable stylized radiance field[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 709-718.

[11] Liu K, Zhan F, Chen Y, et al. Stylerf: Zero-shot 3d style transfer of neural radiance fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 8338-8348.

[12] Nguyen-Phuoc T, Liu F, Xiao L. Snerf: stylized neural implicit representations for 3d scenes[J]. arXiv preprint arXiv:2207.02363, 2022.

[13] Huang Y H, He Y, Yuan Y J, et al. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 18342-18352.

[14] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.

[15] Wang C, Chai M, He M, et al. Clip-nerf: Text-and-image driven manipulation of neural radiance fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3835-3844.

[16] Hyung J, Hwang S, Kim D, et al. Local 3d editing via 3d distillation of clip knowledge[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 12674-12684.

[17] Gordon O, Avrahami O, Lischinski D. Blended-nerf: Zero-shot object generation and blending in existing neural radiance fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 2941-2951.

[18] Song H, Choi S, Do H, et al. Blending-NeRF: Text-Driven Localized Editing in Neural Radiance Fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 14383-14393.

[19] Kerr J, Kim C M, Goldberg K, et al. Lerf: Language embedded radiance fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 19729-19739.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[21] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.

[22] Li C, Li S, Zhao Y, et al. RT-NeRF: Real-time on-device neural radiance fields towards immersive AR/VR rendering[C]//Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design. 2022: 1-9.

[23] Xu H, Alldieck T, Sminchisescu C. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion[J]. Advances in Neural Information Processing Systems, 2021, 34: 14955-14966.

[24] Jing Y, Yang Y, Feng Z, et al. Neural style transfer: A review[J]. IEEE transactions on visualization and computer graphics, 2019, 26(11): 3365-3385.

[25] Höllein L, Johnson J, Nießner M. Stylemesh: Style transfer for indoor 3d scene reconstructions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 6198-6208.