Artistic Radiance Fields for Architectural Style Transfer

Chuhan Chen
Carnegie Mellon University

Qiyu Chen
Carnegie Mellon University

Ziyi Zhou
Carnegie Mellon University

Abstract

This study introduces a novel technique for generating stylized views of buildings using NeRF-based 3D representations, which allows for highly detailed and realistic stylized scenes with intricate surface and lighting effects. The proposed approach offers a cost-effective and time-efficient way to explore different architectural styles without relying on physical mockups or extensive photo editing. To induce geometry warping while preserving the background during the stylization process, the study proposes several strategies, such as patch-based loss, content loss weighting, Plenoxel density optimization, and auxiliary NNFM loss and CLIP loss from complementary style images or text descriptions. The results demonstrate the effectiveness of the proposed method in generating stylized views of buildings with user-specified architectural styles. Our code is available at: `https://github.com/Jameschen7/ARF-for-architecture.git`

1 Introduction

3D reconstruction is the process of creating a three-dimensional representation of a real-world object or scene from a collection of 2D images. This can be achieved using various techniques, such as photogrammetry or structured light scanning. NeRF (Neural Radiance Fields)[7] is a recent method for 3D reconstruction that uses deep neural networks to learn a volumetric representation of a scene's appearance and geometry from a set of 2D images. Unlike traditional 3D reconstruction techniques, NeRF does not require explicit correspondences between images or geometric information. Instead, it models the scene as a continuous function, allowing for highly detailed and realistic reconstructions that capture complex lighting and surface effects.

NeRF has recently gained attention in the field of computer graphics and art due to its ability to produce high-fidelity 3D reconstructions of real-world scenes.[15] This technique offers a novel way to create digital art and can be used in various applications, such as video game design, movie production, and virtual reality experiences. Additionally, NeRF has been used in the creation of digital replicas of cultural heritage sites and objects, enabling the preservation and dissemination of important historical artifacts.

In our project, we propose a novel approach to architectural stylization using NeRF-based 3D building representations. While previous work in stylization using NeRF has focused on artistic style changes on the surface[12], we wanted to explore the potential of using this method for architectural and make-up style transfer. By leveraging the volumetric representation learned by NeRF, we can provide a more efficient and effective way for designers and artists to experiment with new decoration styles on building surfaces or new make-up styles on faces.

Our proposed approach generates new views of a building with a user-specified architectural style, which involves using multi-view image input and a pre-trained Plenoxels radiance field. The aim is to produce consistent free-viewpoint rendering that matches the desired architectural style exemplar. A significant challenge in achieving this is inducing geometry warping during the stylization process to match the global characteristic geometry of the target style. We introduces several possible

approaches, such as utilizing patch-based loss, adjusting the weight on the content loss, optimizing only the density part of Plenoxels, and incorporating auxiliary loss from complementary style images or text descriptions.

Our approach to architectural style transfer using NeRF-based 3D building representations has several advantages over traditional methods. For example, it allows for the creation of highly detailed and realistic stylized scenes, including intricate surface and lighting effects that are difficult to achieve with other methods. Additionally, our approach can be used to explore different architectural styles in a more efficient and cost-effective manner, as it eliminates the need for time-consuming physical mockups or extensive photo editing.

We highlight the contribution of this project as follows.

- We introduce a novel approach to architectural stylization using NeRF-based 3D building representations, which can be used to generate highly detailed and realistic stylized scenes.
- We propose several approaches to induce geometry warping during the architectural style optimization process while preserving the background appearance, including patch-based NNFM loss and auxiliary CLIP loss from text descriptions.
- We demonstrate the effectiveness of our approach to accurately capture the desired stylization effects, through experimental results.
- Our work highlights the potential of NeRF for a wide range of creative applications, extending beyond architecture to make-up stylization.

2 Related Work

2.1 3D Reconstruction

3D reconstruction is a key problem in computer vision that has been tackled by a variety of methods. Traditional approaches, such as Structure from Motion (SfM) and Multi-View Stereo (MVS), estimate the geometry and appearance of a scene from multiple images or depth maps. Recently, learning-based methods like DeepSFM [13] and COLMAP [10] which leverage deep neural networks, have shown improved accuracy and robustness.

Neural implicit surface representations (NISR) have emerged as a promising technique for representing complex 3D shapes with fewer parameters. Occupancy Networks [5] and DeepSDF [8] are examples of NISR-based 3D reconstruction methods. They use deep neural networks to predict the signed distance function (SDF) of a 3D shape given a 3D coordinate.

Neural Radiance Fields (NeRF) [7] is a recent approach that uses deep neural networks to represent the radiance field of a 3D scene. NeRF is trained on a set of input images and their corresponding camera poses to estimate the radiance field at any given 3D coordinate. NeRF has been shown to produce high-quality photorealistic images and is capable of generating novel views of a scene. Hybrid approaches that combine traditional methods and neural networks have also been proposed, such as the Depth-supervised NeRF method [2].

In summary, traditional 3D reconstruction methods have limitations in capturing complex scenes. Neural Implicit Surface Representations and Neural Radiance Fields have shown promising results in capturing complex 3D shapes and generating photorealistic images. Hybrid approaches that combine traditional methods and neural networks are also being explored, which can leverage the strengths of both approaches. These techniques have potential applications in various domains such as virtual reality and autonomous driving.

2.2 NeRF stylization

In addition to its original application, NeRF has also inspired several related works that leverage the representation power of radiance fields for 3D stylization and manipulation.

ARF (Artistic Radiance Fields) [15] is a recent 3D stylization technique that utilizes a pre-reconstructed radiance field of a real scene and converts it into an artistic radiance field by matching feature activations extracted from an input 2D style image. This approach results in high-quality stylized novel view synthesis. In contrast to other style transfer methods that operate in 2D image

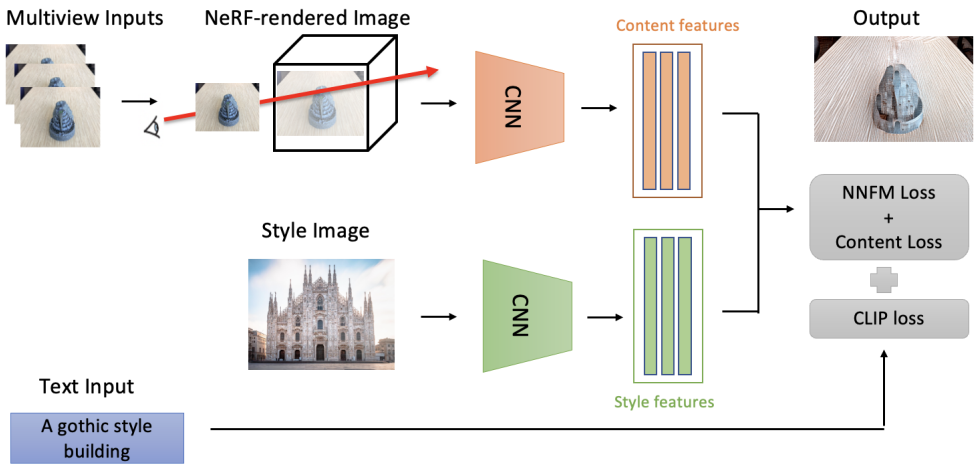


Figure 1: Our method combines image based feature prompt with language prompt for stylization

space, ARF operates directly on the 3D radiance field, leading to more consistent and realistic stylization results.

NeRF-Art [12] is another NeRF stylization approach that manipulates the style of a pre-trained NeRF model with a simple text prompt. It utilizes a global-local contrastive learning strategy, combined with the directional constraint to simultaneously control both the trajectory and the strength of the target style. By leveraging the text input, NeRF-Art enables the user to specify arbitrary styles for 3D scene synthesis, leading to a wide range of creative applications.

CLIP-NeRF [11] is another multi-modal 3D object manipulation method for NeRF that enables user-friendly manipulation of NeRF models, using either a short text prompt or an exemplar image, by leveraging the Contrastive Language-Image Pre-Training (CLIP) model[9]. This method allows for easy modification of 3D objects and scenes by simply providing a natural language description or a visual example.

These related works demonstrate the potential of using NeRF-based techniques for stylization and manipulation of 3D objects and scenes, and provide inspiration for further development of similar approaches in the future.

3 Methods

Our idea for building stylization will be based on NeRF [7]–a NN-based 3D scene representation structure for novel view synthesis–and its following variants [14] for improved rendering efficiency. The reason is that previous works such as ARF[15], NeRF-Art[12] and CLIP-NeRF[11] have already demonstrated the possibility to achieve image-driven or text-driven neural radiance field stylization using some nearest neighbor matching loss or contrastive loss as the style guide, and performed well on objects like chairs, faces, and indoor scenes.

Compared to previous application targets, our project is targeted at outdoor building architectural style transfer via 3D neural radiance field stylization and mainly builds on ARF [15]. Given multiple views of a building (e.g. a temple or a playground) and a target architectural style image (e.g. Notre-Dame de Paris representing French Gothic) plus a set of complementary negative target style images and textual descriptions of architectural styles, we want to synthesize novel view images of the same building with the user-specified architectural style. Similar to how ARF [15] primarily makes use of the Plenoxels [14] as the fundamental 3D scene representation, we also first reconstruct a photo-realistic Plenoxels radiance field from multi-view image input of the building, and then optimize the radiance field to match the desired architectural style exemplar. At test time, we use volumetric rendering to generate stylized novel views of the building.

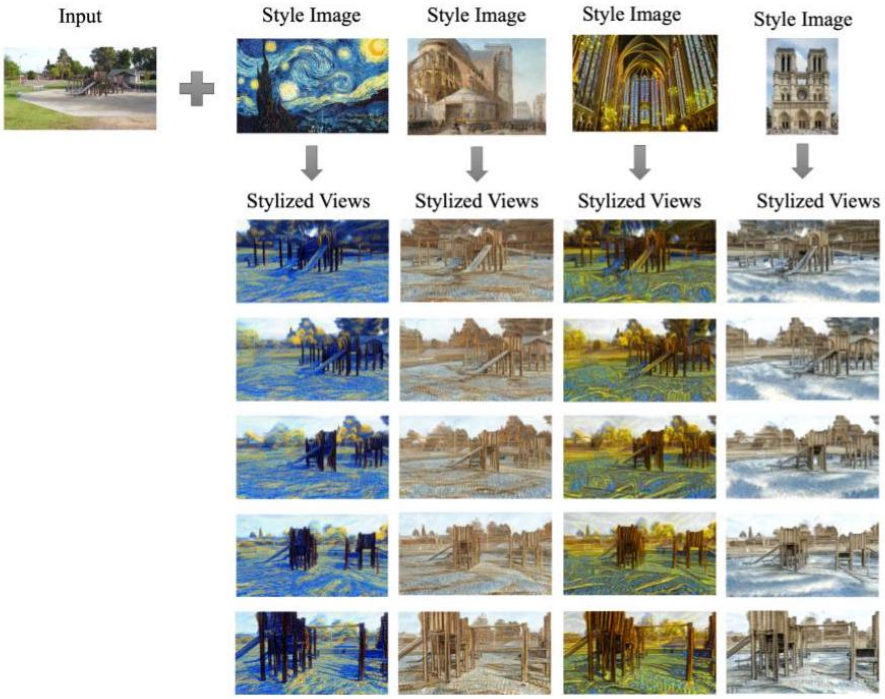


Figure 2: Preliminary results on ARF [15]: style transfer of Playground scene using artistic style image, artistic architectural image, and fully realistic architectural style image. Note how in each case the stylized playground scene only learns a few arches from different Gothic style images and both the foreground building and the background are stylized.

Architectural style transfer and the general artistic style transfer for 3D scene are similar in a way that both of them need color match between the rendering and the style image and require accurate 3D geometry to accomplish consistent free-viewpoint rendering. However, one significant difference between them is that the latter emphasizes at the transfer of local style feature in terms of brushstrokes and local painting patterns (see Fig. 2), whereas the former anticipates the target architectural style only on the foreground building in terms of certain characteristic geometry. For instance, Byzantine architecture is featured by its round arch; Gothic architecture is characterized by its pointed arch and its thin and tall structure.

Therefore, we need some way to modify the content to encourage additional warping towards corresponding geometric shapes during the stylization process while preserving the background appearance.

3.1 Method Details

We propose several possible approaches to induce geometry warping during the architectural style optimization process of the pre-trained Plenoxels radiance field:

1. In Artistic Nerf [15], NNFM (Nearest Neighbor Feature Matching) loss is used to find closest features between style and rendered pre-trained Nerf image and force the feature to be close to the style feature accordingly. This method also changes the background into the desired style, which can become distracting and affect the overall effect of the visual quality. We modify the input to this loss function by concatenating the style image with images that are similar to the background of the pre-trained Nerf scene, so that when looking for nearest neighbors, background of the rendered image would be matched with features from the background image instead of the style image, making the background unchanged.
2. We tried different weight on the content loss so that not only the style but also geometry can be altered based on given image.

3. We examined the effect of optimizing the density part of plenoxels in addition to the color part during optimization to allow the loss focus on altering the geometry as well.
4. In Nerf-Art[12] and Clip-Nerf[11], language based Clip loss has been proved to be effective in changing the structure of pretrained Nerf given a language prompt. Specifically, in Nerf-Art, the relative directional loss was proposed to reduce the cosine difference between 2 images and 2 language prompts for better diversity of generated result. In our architecture stylization case, we use the 2 images as rendered view of the original pre-trained Nerf and styled Nerf that is being optimized, and phrase describing the pre-trained Nerf scene and the target architecture style. Additionally, a global contrastive loss that uses embedding of a group of negative prompts as negative samples and the target prompt as positive samples to compare the target image embedding with is also used to control the strength of stylization. In our case, we use generic architecture style prompts as negative samples and target style prompt as positive sample, and rendered image of stylized Nerf as target image.

4 Experiments

4.1 Experimental Setup

We first pre-train the Playground scene from the Tanks and Temples dataset [4] and the fortress scene from LLFF dataset [6] using implementation of radience field from Plenoxels [14]. Then, we fine-tune the pre-trained plenoxel model to perform style transfer using different architecture style images by encouraging both content and style change towards the target architecture with the NNFM loss from [15] and the clip loss from [12], and discouraging background stylization via our negative style NNFM loss (not style loss in the negative direction but style not to be changed). We performed the experiments based on ARF code-base.

4.2 Comparison with ARF

Fig. 3 shows the comparison of our method with vanilla ARF. While in ARF, the artistic texture is transferred to the pre-trained scenes, the geometry is not so well transferred. For instance, on the first level of the fortress, ARF result preserves most of the random protrusion from the original scene, but our method generates flat building surface with some regular spikes left. More importantly, ARF indiscriminately stylizes the table surface in the background and fills it with random arch shapes; in contrast, our method utilizes the nature of nearest neighbor matching to preserve the background via the negative style images.

4.3 More Result

Fig. 4 demonstrates the effect of our architectural style radiance field method using different architecture images. In the second column, the stylized building learns the characteristic arches from the Gothic style image; in the third column, the plain, flat, and sharp geometry from brutalism is successfully transferred to the pre-trained fortress scene; and in the last column, the Japanese eaves are added to the top of every level of the fortress. These examples show the success of our method in learning the architectural style from the target image without much disturbance to the background.

4.4 Naive Ablation

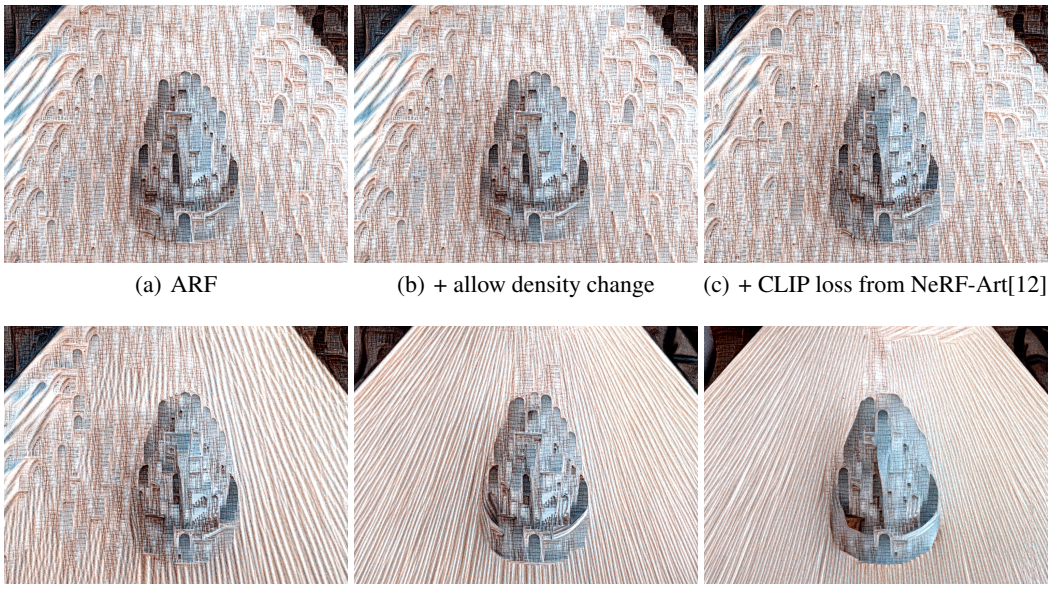
Next, we naively show the effect of each module in our approach by gradually adding them to the vanilla ARF method (Fig. 5 (a)). First, to increase the amount to content/density change, we raise the learning rate of sigma in ARF from 0 to 3e3 to allow density learning along with color learning (Fig. 5 (b)). Nevertheless, the mount of content change is insignificant, so we add the CLIP loss to stimulate additional geometry change (Fig. 5 (c)). At this step, the synthesized image can have more detailed shape compared to the original ARF result. However, the background is also mistakenly stylized. To reduce the extent to which the background is changed, we introduce the negative style NNFM (Fig. 5 (d)), which successfully removes a large proportion of random arches. To further improve the data efficiency, we add data augmentation for the negative style images using flipping and 90-degree-rotation to create eight negative targets from each image. This change in turn preserves more texture from background in the original scene (Fig. 5 (e)). Finally, in order to remove the noise

(a) Content (b) Architectural style

(c) ARF (d) Ours

Figure 3: Comparison of our method with vanilla ARF [15] result.

Figure 4: Additional visual result of our method. The top raw shows the corresponding architectural style, and the first image in the following row is the original content.



(a) ARF (b) + allow density change (c) + CLIP loss from NeRF-Art[12]

(d) + negative style NNFM (e) + negative style augmentation (f) + finetune

Figure 5: Visual effect of each module used in our method.

(a) Content: playground (b) Architectural style: brutalism (c) Synthesized

Figure 6: Visual effect of each module used in our method.

and smooth building surface, we finetuned the weight for different loss and in particular raise image total variation loss weight to 5 (Fig. 5 (f)).

4.5 Limitation

Although our method adds both NNFM loss for content and CLIP loss, the degree to which the pre-trained radiance field can change is still sometimes limited depending on the similarity between the original scene and the target architecture. An example is shown in Fig. 6, whose large and wide surface creates difficulty for the thin playground architecture to learn.

Besides, though we proposes an innovative way to preserve the background pattern without relying on other information like segmentation, the resultant preservation may often fail to achieve the same level of quality.

5 Conclusion

5.1 Summary

In summary, the proposed project aims to use NeRF-based 3D reconstruction and stylization techniques to transfer architectural styles onto buildings. This approach allows for the creation of highly detailed and realistic stylized scenes with intricate surface and lighting effects, which are difficult to achieve with traditional methods. The project builds on previous works such as ARF[15], NeRF-Art[12], and CLIP-NeRF[11] to optimize the pre-trained Plenoxels radiance field to match the desired architectural style exemplar. The proposed methods include adjusting the weight on the density part of Plenoxels, and incorporating auxiliary CLIP loss[9] from text descriptions and negative style NNFM loss[15] from complementary style images to induce geometry warping during the stylization process while preserving the background appearance. Overall, the proposed approach offers an efficient and cost-effective way for designers and artists to experiment with new decoration styles on building surfaces.

5.2 Future Direction

In the future, there are two potential directions for the work done in this paper.

The first direction involves extending the tool to new and unique makeup styles. As technology advances and makeup trends evolve, there will be a growing demand for makeup style transfer tools that can accurately replicate new styles. Moreover, one similarity between makeup style transfer and architectural style transfer is that they both require some changes to the content like extending the eyebrow length or appending more ornaments. By continuing to develop the technology, researchers and engineers can enable makeup style transfer tools to replicate even the most intricate and complex makeup styles, opening up new possibilities for creative expression.

The second direction for the future of style transfer technology involves taking inspiration from recent advances in natural language processing, such as GPT[1] and SAM[3], to create a more generalized style transferring tool. By incorporating text-prompt features, makeup style transfer tools could become even more sophisticated and versatile, enabling users to transfer styles across a wide range of different images and contexts. This would allow makeup artists and enthusiasts to experiment with architecture, makeup, etc. styles in new and exciting ways, without being limited by the existing capabilities of the technology. By combining these two directions, the future of style transfer technology is likely to be a highly advanced and versatile tool that enables users to express themselves in new and exciting ways.

References

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[2] Keliang Deng, Yichao Shi, Shuangming Chen, Xian Cao, and Jiebo Luo. Depth-supervised nerf: Fewer views and faster training for free. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[4] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[5] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019.

[6] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[7] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[8] 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 Conference on Computer Vision and Pattern Recognition*, pages 5544–5553, 2019.

[9] 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. *arXiv preprint arXiv:2103.00020*, 2021.

[10] Johannes L Sch"onberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113. IEEE, 2016.

[11] 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.

[12] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization. *arXiv preprint arXiv:2212.08070*, 2022.

[13] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu1, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. *ECCV*, pages 230–247, 2020.

[14] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021.

[15] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields, 2022.