

CSCI-677 Project Final Report:

L-3DS: Language-driven 3D Stylization

Pranav Budhwant
Univ. of Southern California
budhwant@usc.edu

Jingmin Wei
Univ. of Southern California
jingminw@usc.edu

Xianshi Ma
Univ. of Southern California
xianshim@usc.edu

Abstract

3D Style transfer is a pivotal aspect in the field of computer vision. Traditional 3D stylization approaches have predominantly focused on stylizing entire scenes, often overlooking the nuanced details of individual objects. While there have been strides in object-specific style transfer, they are primarily instance-based and suffer from artifacts. Addressing these limitations, we introduce L-3DS, Language-driven 3D Stylization, a framework that leverages open-ended language queries for precise object selection and stylization within a 3D scene. Our approach allows for the language input to specifically choose the objects to be stylized, thereby enabling a more focused and detailed stylization. Experiments show that our method effectively achieves language-driven object-level targets selection and completes 3D stylization tailored to the specific targets. Some experiment results in video format are available at our [Google Drive](#).

1. Introduction

As 3D reconstruction technology become widespread, the demand for editing and manipulating these scenes is also increasing. The ability to edit 3D scenes has empowered creators to break through the boundaries of imagination and precision. A prominent application in this area is 3D style transfer, which aims to transfer artistic features from a single 2D image to a real-world 3D scene. Several significant works [2, 3, 9, 15, 17] have already achieved success in this domain. However, these methods predominantly focused on applying a singular style image to stylize the entire scene. While there are stride [7] in object-specific 3D stylization, it is primarily instance-based and still suffer from artifacts. While an intriguing study known as Sine [1], offered a degree of flexibility for user-edit based stylization, it necessitates that the user edits a single image, and these edits are limited by semantic constraints and geometric priors.

To enhance the level of target control in the stylization

of 3D scenes, we introduce L-3DS, Language-driven 3D Stylization, which leverages open-ended language queries to precisely select and stylize objects within 3D scenes. Unlike previous approaches, our method allows language inputs to directly target objects for 3D stylization with any chosen style, resulting in more detailed and focused outcomes.

Specifically, L-3DS facilitates language-driven 3D stylization of specific objects in a radiance field using calibrated scene images, a natural language query for object identification, and a style image. The framework includes three main phases: i) Generating a photo-realistic radiance field using Plenoxels for 3D representation; ii) Detecting objects within the radiance field using GroundedSAM, a combination of GroundingDINO [8] and SAM [6], for accurate segmentation based on text prompts; iii) Applying stylization to the identified object, employing a VGG-based Masked Nearest-Neighbor Feature Matching (NNFM) Loss for style alignment while maintaining the object's geometry. This process integrates precise language-guided segmentation, effective style transfer, and robust optimization for targeted 3D stylization.

The qualitative experiments showcase advancements in 3D scene stylization, demonstrating that our method, L-3DS, enables precise language-guided object selection and accurate 3D stylization. Our approach highlights the potential of merging 3D visualization with language modality, leading to more intuitive and adaptable 3D stylization, design and editing.

2. Related Work

3D style transfer: In the burgeoning field of scene style transfer, several key studies stand out. Kai Zhang *et al.* [17] revolutionized 3D scene stylization by integrating 2D artistic elements. Yi-Hua Huang *et al.* [4] and Hsin-Ping Huang *et al.* [3] focused on consistent style transfer across 2D and 3D domains. Fangzhou Mu *et al.* [12] explored generating stylized views from a single image. Kangxue Yin *et al.* [15] and Oscar Michel *et al.* [9] introduced innovative methods

for creating and transforming 3D shapes and meshes with style variations. Lastly, Lukas Hollein *et al.* [2] applied style transfer to indoor 3D reconstructions, enhancing the aesthetic appeal of interior models. Together, these studies exemplify the dynamic interplay between artistic creativity and technological innovation in 3D environments.

Most stylization approaches above focused on the stylization of entire scenes and use either image- or language-based style transfer techniques. Unlike previous studies, our approach focuses on developing a novel 3D stylization method that enables object-level selection, achieving style transfer tailored to the specific targets.

ARF: Artistic Radiance Fields [17] marks a significant advancement in 3D stylization by introducing a novel radiance field-based approach that enables high-quality, consistent transfer of artistic styles from 2D images to 3D scenes. It incorporates the Nearest Neighbor Feature Matching (NNFM) Loss and a new deferred back-propagation method, improving detail preservation and reducing GPU memory demands. Extensive evaluations and user studies demonstrate the method’s superior ability to handle complex scenes and a wide range of artistic styles, resulting in detailed and immersive stylized 3D renderings.

LERF: Language Embedded Radiance Fields [5] is a technology that integrates natural language processing with 3D scene understanding. It combines language embeddings, similar to those from CLIP [11], with Neural Radiance Fields (NeRFs) [10] to create a dense, multi-scale 3D environment that can be interactively queried in real-time using natural language. The key feature of this technology is its ability to support a wide range of long-tail, open-vocabulary queries, demonstrating strong capabilities in scene localization and understanding of abstract concepts across different scales of a volume. LERF’s relevancy maps generated by language queries can be leveraged to harvest activation maps and accurately segment the objects targeted for stylization. However, this method is computationally slow due to real-time rendering.

Grounding DINO: This method [8] innovatively combines the Transformer-based DINO detector with grounded pre-training for open-set object detection, significantly improving performance by fusing vision-language modalities. It extends evaluations to Referring Expression Comprehension (REC) and shows exceptional results on various datasets.

Segment Anything: This work [6] presents the Segment Anything Model (SAM) and the unprecedented SA-1B dataset, which is the most extensive segmentation dataset currently available, featuring over 1 billion masks. This study is notable for SAM’s remarkable zero-shot capabilities, enabling it to adapt and perform efficiently across a diverse range of tasks and image distributions. Additionally, the work addresses responsible AI practices and contributes

significantly to the field by making both SAM and the SA-1B dataset available for further research, thereby setting a new benchmark in the domain of computer vision.

3. Method

Our method enables language-driven 3D stylization of specific objects within a radiance field. Specifically, the method utilizes a set of calibrated scene images, an open-ended natural language query identifying an object, and a style image. Taking these, our framework trains a photo-realistic radiance field in which the identified object is stylized according to the chosen style.

The framework L-3DS, depicted in Fig. 1, consists of three key components:

1. **Generation of Radiance Fields:** The initial phase involves training a photo-realistic radiance field using the scene’s calibrated images. Building upon ARF [17], our method is flexible regarding the representation of radiance fields. In this work, we employ Plenoxels [16] which utilize a sparse voxel grid to efficiently represent the 3D scene. Each voxel stores a scalar opacity and a vector of spherical harmonic coefficients for each color channel. The radiance field is defined through trilinear interpolation over this sparse voxel grid, employing differentiable volume rendering as described in NeRF [10]. We employ the losses and training methodology as outlined in the original Plenoxel method.
2. **Object Detection within Radiance Fields:** Initial attempts to use LERF [5] for joint training of a language field with the radiance field yield suboptimal segmentation masks due to noisy relevancy maps (Fig. 2). To address this, we adopt GroundedSAM (GroundingDINO + SAM) for precise segmentation mask generation of user-specified objects, detailed in Section 3.1.
3. **Stylization of the Object within Radiance Fields:** The final stage involves fine-tuning the radiance field to apply stylization to the object. This is achieved by employing a VGG-based Masked Nearest-Neighbor Feature Matching Loss [17], ensuring that the style closely mirrors the reference image while retaining the object’s inherent geometric characteristics. Detailed elaboration is provided in Section 3.2.

3.1. GroundedSAM

In our L-3DS framework, we incorporate the GroundedSAM to enhance segmentation accuracy. GroundedSAM represents a fusion of GroundingDINO [8] and Segment

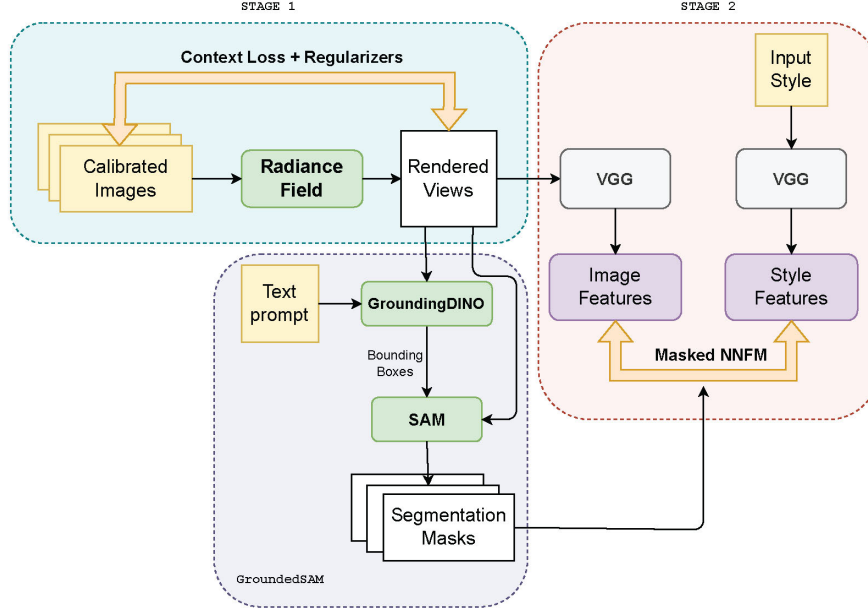


Figure 1. Pipeline of L-3DS. It consists of 3 components: a Radiance Fields for view rendering (green block); GroundedSAM for language-driven object selection (purple block); a Radiance Fields for 3D stylization based on NNFM loss (pink block).

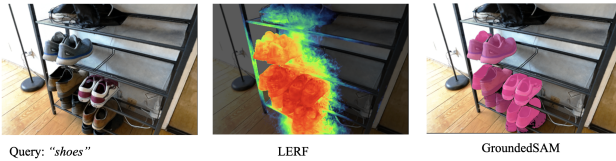


Figure 2. Segmentation Masks obtained from LERF vs GroundedSAM

Anything [6]. This integration aims to detect and segment objects in images based on text inputs. In GroundedSAM pipeline, given any text prompts, GroundingDINO initially identifies targeted boxes. Subsequently, SAM utilizes these identified boxes as prior knowledge to perform precise segmentation masks. This process forms a critical component of our pipeline, enabling enhanced object-level target selection based on textual descriptions. The accurate mask can then be applied for targeted 3D stylization, thus contributing significantly to the overall efficacy of our approach.

We present the process from a calibrated 2D view to segmentation mask by GroundedSAM in Fig. 3. When given the text prompt 'desk', the model successfully generates an accurate segmentation mask of the television within the scene. Additionally, we have optimized the response mech-

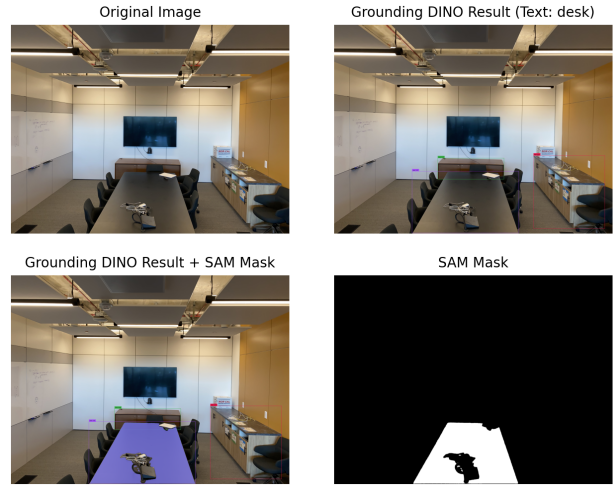


Figure 3. A accurate segmentation mask returned by GroundedSAM, given text prompt='desk' under room scene.

anism of SAM for instances where certain objects are undetectable or when masks cannot be generated, thereby preventing errors and enhancing the robustness of our system.

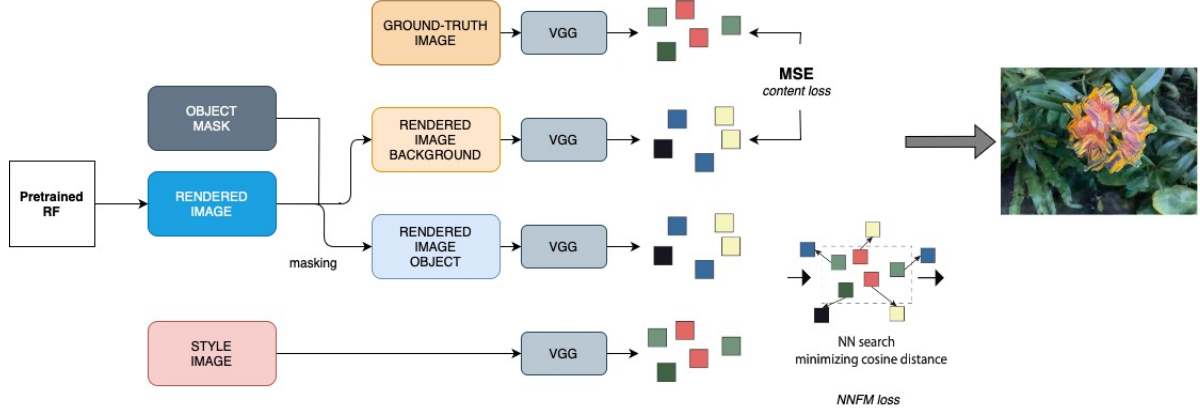


Figure 4. Masked NNFM. It minimizes the cosine distance between features in the rendered views’ feature map and their nearest counterparts in the style images’ feature map.

3.2. Masked NNFM and Stylization

Utilizing the photo-realistic radiance field reconstructed as per Section 3, along with masks generated via the approach in Section 3.1, our framework fine-tunes the radiance field for object-specific stylization. This is accomplished by applying Nearest Neighbor Feature Matching (NNFM) loss [17] to the object. The NNFM loss minimizes the cosine distance between features in the rendered image’s feature map and their nearest counterparts in the style image’s feature map. This loss is selectively applied to the object using masks from Section 3.1, effectively confining the style transfer to targeted objects. For background preservation, a mean squared error (MSE) content loss is applied to the background pixels, as illustrated in Fig. 4.

These image losses, both content and style, are applied using the deferred backpropagation method introduced in [17]. The process involves initially rendering the image for a given viewing direction, from the radiance field with auto-differentiation disabled. Subsequently, image-level losses are computed using the rendered image, ground truth, and mask, giving a cached gradient image. Finally, auto-differentiation is re-enabled and in a patch-wise manner the image is re-rendered and the cached gradients are back-propagated to the scene parameters for accumulation.

4. Experiments

Due to space constraints in this paper, we primarily showcase our results through static images. However, for a more comprehensive view, including dynamic visualizations, we invite readers to explore additional results available in video format on our [Google Drive](#).

4.1. Experiments Settings

Similar to previous work [17], in our study, we trained L-3DS over a spectrum of 8 different 3D scenes, includ-

ing Fern, Flower, Fortress, Horns, Leaves, Orchids, Room and Trex. We build L-3DS through the PyTorch [13] framework. All models are trained with 1 RTX-3090 GPU.

For the style images, we primarily use 3 different style images as defined in Fig. 6.

We pretrain the radiance field for 10 epochs, and fine-tune for style transfer for another 10 epochs.

4.2. Qualitative Results

Initially, we focus on experiments with different text prompts under the same scene. In most of the 3D scenes, the targeted object constitutes a major portion of the scene’s space, leading us to select the room scene for this experiments. When presented with different text prompts such as ‘tv’ in Fig. 7a and ‘desk’ in Fig. 7b, GroundedSAM proficiently returns precise segmentation masks for each. The capability of L-3DS to distinguish and act on varied object-specific prompts demonstrates its proficiency in achieving object-level selection for 3D stylization in complex scenes.

We then shifted our focus to experiments featuring consistent text prompts paired with varying style inputs under the same 3D scene. Fig. 5 illustrates this concept: for the prompt ‘fern’, the top row displays a series of different 3D stylization, each corresponding to a unique style input as categorized in Fig. 6. A similar pattern is observed in the bottom row for the prompt ‘flower’. These results collectively demonstrate the adaptability and precision of the L-3DS framework. It not only maintains accurate object-level stylization across diverse style inputs but also seamlessly adjusts to these varying inputs to execute effective 3D style transfer.

What if we change the text prompt and the style input together? Does the L-3DS framework maintain its robustness? Fig. 8 demonstrates this scenario within the same 3D fortress scene. In the left Fig. 8a, we depict the 3D stylization achieved using the text prompt ‘castle’ combined with



Figure 5. Different Styles, Same Text Prompt under Different Scene. Given 3 different style images input: from left to right, **style 1, 2, 3** as we defined in Fig. 6), the first rows show the 3D stylization given **text prompt='fern'** and the second row shows the 3D stylization given **text prompt='flower'**.



Figure 6. Style Inputs. We named these three styles, from left to right, as style 1, style 2, and style 3.

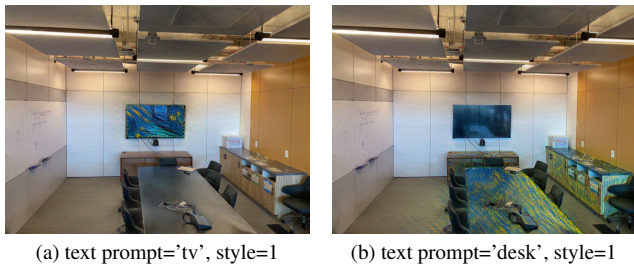


Figure 7. Different Text Prompts under Same Scene 'Room'. The left one shows the 3D stylization given **text prompt='tv'**, while the right one shows the result given **text prompt='desk'**.

style 1. Conversely, in the right Fig. 8b, we present the styl-

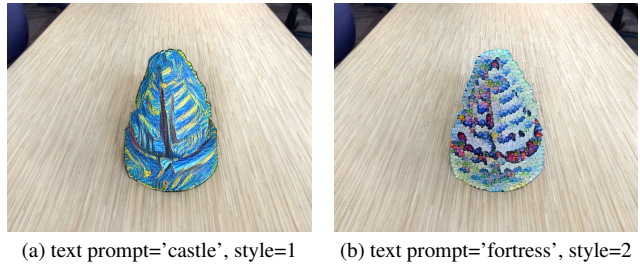


Figure 8. Different Text Prompts & Different Styles under Same Scene. Change the text prompt and the style input together.

ization resulting from the prompt 'fortress' combined with style 2. The experiment indicates that L-3DS adeptly handles simultaneous changes in both text prompts and style inputs, successfully executing object-level 3D stylization.

4.3. Ablation Study

Our ablation study explores various content weights and VGG block selections for computing the NNFM loss. Fig. 10 compares different content weights, revealing that excessively high content weight introduces artifacts, while too low a weight impairs background reconstruction due to insufficient training signal for background pixels. An optimal balance is achieved with a content weight of $1e-3$, as

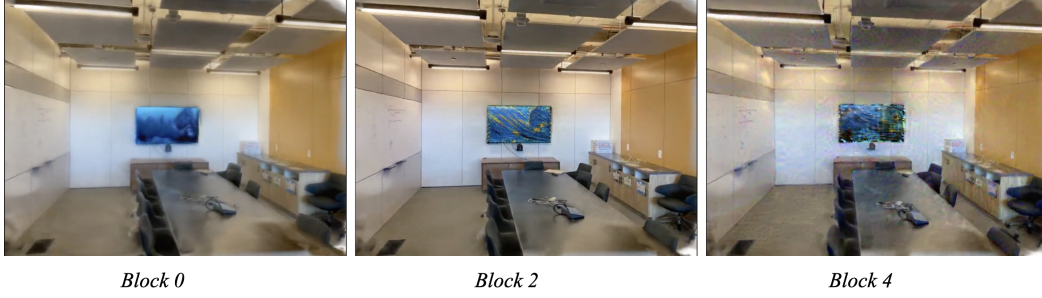


Figure 9. VGG Block Ablation

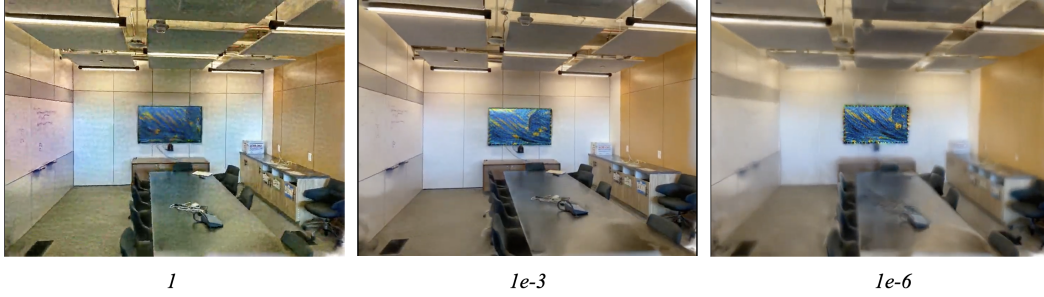


Figure 10. Content Weight Ablation

empirically determined for the best qualitative outcomes.

Additionally, Fig. 9 contrasts style transfer results using different VGG blocks. Early blocks result in blurry reconstructions due to the lack of detailed features, whereas the final block introduces artifacts. Thus, we empirically choose block 2 features for NNFM, as it provides a balanced outcome.

5. Conclusions and Future Work

5.1. Conclusions

In this work, we present L-3DS, a novel framework termed **Language-driven 3D Stylization**. This approach harnesses the power of open-ended language queries to precisely select and stylize objects within a 3D scenes. Our method stands out by allowing language inputs to directly target objects for 3D stylization, leading to more detailed and focused results. Experiments and qualitative results demonstrated that L-3DS not only facilitates language-guided selection of objects under different camera views but also successfully accomplishes the accurate 3D stylization tailored to the chosen targets. The outcomes of our research highlight the potential of integrating 3D visualization with language modality.

5.2. Limitations and Future Work

Our L-3DS is subject to the limitations inherent in GroundingDINO and SAM. Notably, SAM only cares

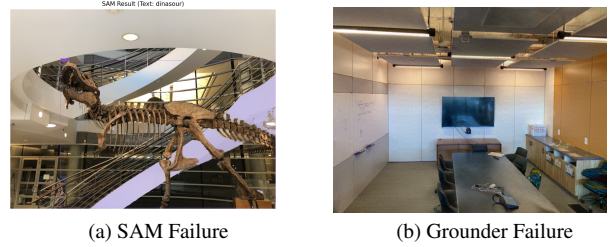


Figure 11. Limitations

about segmenting foreground and background instead of given text or labels, thus its segmentation is influenced by GroundingDINO bounding boxes, rather than the input query directly. A typical failure example is shown in Fig. 11a. This dependency can result in variability in the segmentation quality due to differing bounding box scores that GroundingDINO output across different 3D views, potentially leading to inconsistent mask generation and, thus, inaccurate 3D style transfer.

Besides, under some complex scenes, GroundingDINO’s output boxes vary as camera views change, reflecting different scores for bounding boxes under various viewpoints. This variation allows for multi-object stylization to a certain degree, contingent on GroundingDINO’s output. An example of this can be seen in our experiments with room scenes and the prompt “chair” shown in Fig. 11b. However, this implementation, while functional, does not align with our

envisioned approach of true multi-object style transfer. Our goal for GroundedSAM is to attain the capability to simultaneously segment dense targets, enabling dense and accurate multi-object style transfer. This is in contrast to relying on the variability of scores from the grounder to randomly generate different masks for style transfer.

To improve the L-3DS, our future endeavors will focus on extending the work in several key areas:

1. **Multi-Object Stylization:** Exploring the stylization of multiple objects within a single scene. By enhancing GroundedSAM’s precision in segmentation, we aim to move beyond the limitations of score-dependent mask generation and achieve a more controlled multi-object stylization.
2. **Individualized Multi-Instance Stylization:** Enabling the application of distinct styles to different instances of the same object. We aim to achieve the stylization of the same target (text prompt) within a single scene with different styles. This method necessitates the linking of segmentation masks and style images, thereby facilitating the style transfer of distinct styles to the different instances of the same object.
3. **”Stylize Anything”:** Achieving language-driven style specification, combined with the current language-driven object specification. This involves integrating text-conditioned style image generation using a generative model like Stable Diffusion [14]. By combining GroundedSAM’s object-focused capabilities with style images generated through Stable Diffusion, we aim to realize a comprehensive ”stylize anything” model. The goal is to interpret text inputs such as ”stylize [object] with [style]” and execute the 3D stylization accordingly.

References

- [1] Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Sine: Semantic-driven nerf editing with prior-guided editing field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20919–20929, 2023. 1
- [2] Lukas Höllein, Justin Johnson, and Matthias Nießner. Stylemesh: Style transfer for indoor 3d scene reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6198–6208, 2022. 1, 2
- [3] Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maanesh Kumar Singh, and Ming-Hsuan Yang. Learning to stylize novel views. *CoRR*, abs/2105.13509, 2021. 1
- [4] 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. 1
- [5] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lurf: Language embedded radiance fields. *arXiv preprint arXiv:2303.09553*, 2023. 2
- [6] 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. 1, 2, 3
- [7] Dishani Lahiri, Neeraj Panse, and Moneish Kumar. S2rf: Semantically stylized radiance fields. *arXiv preprint arXiv:2309.01252*, 2023. 1
- [8] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2
- [9] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022. 1
- [10] 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. 2
- [11] Shailesh Mishra and Jonathan Granskog. Clip-based neural neighbor style transfer for 3d assets. *arXiv preprint arXiv:2208.04370*, 2022. 2
- [12] Fangzhou Mu, Jian Wang, Yicheng Wu, and Yin Li. 3d photo stylization: Learning to generate stylized novel views from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16273–16282, 2022. 1
- [13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zem-

- ing Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, pages 8026–8037, 2019. [4](#)
- [14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. [7](#)
- [15] 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. [1](#)
- [16] 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*, 2(3):6, 2021. [2](#)
- [17] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *European Conference on Computer Vision*, pages 717–733. Springer, 2022. [1](#), [2](#), [4](#)