Geometry Transfer for Stylizing Radiance Fields

Supplementary Material

Please refer to our project page for video results.

A. Additional Implementation Details

We utilized TensoRF-VM-48 [2] as our 3D scene representation. The number of vector-matrix components is set to $R_{x,1} = R_{x,2} = 4, R_{x,2} = 16$ (where $x \in c, \sigma, \Delta$), ensuring all three grids have the same number of components. For modeling appearance and deformation fields, we employed two-layer small MLPs as feature decoding functions for each component. The intermediate channel dimension is set to 128 for colors, consistent with the original TensoRF, and 32 for deformation fields.

During the style transfer, we also implemented deferred back-propagation [9] for memory-efficient training. The loss function comprises the proposed style loss (Eq. 4 in the main paper), along with the content loss (i.e., $\mathcal{L}_{content} = l_2(F_{render}, F_{content})$) to control the level of stylization. Here, F_{render} represents the extracted VGG feature from the rendered RGB image, and $F_{content}$ is that from the original training image. We maintained the gradient flow from the content loss to the volume density to enable indirect content preservation in both appearance and shape. The scale factor applied to the content loss is set to $5e-3$, while that for the style loss is set to 1. Optimization was performed over 500 iterations with a batch size of 1, and training can be completed in 2 to 3 hours on a single V100 GPU with 32GB of memory.

For the style images, we downloaded them online and estimated their depth maps using a zero-shot depth estimation network [1].

To stylize Panoptic Lifting [6] on the ScanNet dataset, we utilized the authors' pre-trained model and incorporated deformation fields to apply our proposed methods.

Patch-wise optimization. As mentioned in the main paper, we utilized the `conv2` and `conv3` blocks of VGG-16 for computing the style loss. In the original nearest neighbor loss [3], the sizes of different blocks were reduced by bilinear downsampling to match them with the spatial resolution of `conv3`, allowing for nearest matching to be performed independently for each block. However, we identified two minor issues with this approach: 1) The downsampling operation tends to lose detailed local shape information due to the sampling of sparse locations; 2) Performing independent matching for multiple layers can result in the generation of inaccurate and overlapping patterns.

To address the first problem, we applied different strides for the patchwise scheme. Our patch-wise optimization can be straightforwardly implemented by the `unfold` operation in PyTorch [5]. For instance, given that the spatial resolution of the $\texttt{conv2} \in \mathbb{R}^{\frac{h}{2} \times \frac{w}{2} \times c}$ block is twice that of the $\texttt{conv3} \in \mathbb{R}^{\frac{h}{4} \times \frac{w}{4} \times c'}$, we unfold them with a stride of 4 for the former and 2 for the latter. This adjustment matches their spatial resolutions to $\frac{h}{8} \times \frac{w}{8}$. Here, the channel dimension becomes $c \times k^2$ and $c' \times k'^2$, where k denotes the size of each patch. Through this scheme, we are able to more effectively align each patch in the content feature with the style feature, preserving intermediate information without loss.

For the second problem, we do not perform matching independently. Instead, we first identify the closest pair using the output of `conv3` and then apply the identical pixel locations of these pairs for optimizing `conv2`. This approach ensures that the losses from multiple blocks coordinate to transfer the same parts of the style image to the content scene. Such coordination prevents patterns from overlapping and contributes to the formation of complete and clear shapes. While concatenating the features from both blocks is an alternative since they share the same resolutions, this method significantly increases the computational load for nearest matching without yielding a meaningful difference in results.

Perspective style augmentation. The human visual system perceives depth through various cues present in images [7, 8]. We have developed perspective augmentation based on well-defined design principles derived from these cues. While rendering a video of the stylized scene easily conveys depth via motion parallax, achieving a natural and structured stylization in still images requires the incorporation of several additional effects.

1. Diminishing scale. As we described in the main paper, the major effect of the proposed augmentation is to vary the size of style patterns depending on their distance, so that the sizes diminish as the surfaces become farther away.

2. Atmospheric perspective. To maximize the perception of depth, objects that are closer should appear clearer than those that are farther away. In the context of our stylization scheme, this implies that closer objects should exhibit higher intensity/contrast, vivid colors, and complete and accurate patterns and shapes. However, this principle conflicts with our proposed patch-wise optimization, which primarily focuses on transferring the clear and complete patterns of style images. In other words, while the patch-wise optimization allows for the transfer of clean and precise patterns from style images, this does not always equate to an “aesthetically better” outcome.

To improve this technique, we vary the patch size based on depth. The pixels assigned to the closest bin, B_1, are stylized using a loss computed with a larger patch defined by dilation, while the pixels in the rest of the bins are stylized with a smaller patch that does not use dilation.

Simultaneously, the background colors should contain less intensity and less contrast. In terms of style transfer, a shallower block tends to produce colors with high intensity and contrast, whereas a deeper block yields the opposite effect. Therefore, during the loss computation, the pixels in the closest bin, B_1, are stylized using features from both `conv2` and `conv3` blocks. In contrast, the pixels in the remaining bins are stylized only with features from the `conv3` block to achieve these effects.

3. Overlapping. Object boundaries play a crucial role in providing cues for overlapping. When objects are partially overlapped by others, the relative distance between them becomes discernible. However, as stylization methods strive to transfer accurate patterns (via patch-wise optimization), there can be significant and strong alterations to the original shape. This tends to distort the original content and overlay it with style patterns. Such effects can obscure object boundaries that are defined by color differences, even though shape boundaries may still be present. To counter this, reducing the patch size for farther surfaces also helps in maintaining detailed structure and object boundaries, thereby enhancing depth cues provided by overlapping objects.

As a result of this combined approach, clearly improved results are observable, as demonstrated in Fig. 8 of the main paper. Detailed ablation studies and analysis are provided in the following section.

B. Experiments

B.1. Additional Ablations

In this section, we conduct more detailed ablation studies of our contributions, including the specifics mentioned in Sec. A.

Depth map as a style guide. In Fig. 1, we compare stylization results with and without the depth style guide $\mathcal{S}_D$ to demonstrate the benefits of stylizing geometry and using a depth map as a style guide.

In Fig. 1 (a), we freeze the geometry update and only update the appearance based on the RGB style image, $\mathcal{S}_{rgb}$. Without updating the geometry, the scene cannot be fully stylized to transfer the exact patterns from the style image. It replicates some patterns on the large surfaces, but the detailed structures, such as leaves, maintain their original form, which does not accurately represent the given style.

In Fig. 1 (b-1) and (b-2), we enable geometry updates. Although these examples use a style guide as an RGB image, $\mathcal{S}_{rgb}$, the gradient flow from the style loss is connected and capable of updating volume density. Compared to Fig. 1 (a), the details on the leaves better reflect the style patterns as the geometry updates, and the edges on the depth maps slightly become more block-shaped. This demonstrates that updating the geometry enhances the expressiveness of stylization. However, the geometrical guidance from $\mathcal{S}_{rgb}$ is limited, so it’s not sufficient to fully stylize the geometry. As shown in the depth maps, most of the surfaces do not provide any cues from the style image, indicating that the patterns in the rendered RGB are merely hallucinations drawn by changing appearance. Additionally, the right parts of the leaves are removed and disconnected, sometimes failing to maintain the original content when updating geometry without a proper geometric guide. This is one of the reasons why previous works [9, 10] disable geometry updates and focus solely on stylizing appearance.

In Fig. 1 (b-1), we disable our patch-wise optimization but enable it in Fig. 1 (b-2). We observe that with better and more accurate shape of patterns, our patch-wise strategy proves effective in producing accurate shapes even when exclusively using an RGB image as a style guide.

In Fig. 1 (c), we enable geometry updates and add the style guide from the depth map, $\mathcal{S}_D$. Compared to the previous examples, the shape more accurately reflects the style patterns, overall diversity increases, and the content structure is maintained without any areas being removed.

Patch-wise optimization. In Fig. 2, we present a more detailed comparison of our patch-wise optimization, as discussed in Sec. A. It’s important to note that Fig. 2 (a) corresponds to the top figure in Fig. 7 of the main paper, while Fig. 2 (d) aligns with the bottom figure of Fig. 7 (main paper).

In Fig. 2 (a), without the implementation of patch-wise optimization, the patterns are found to be imperfect and unclear. This is primarily because many surfaces fail to properly exhibit patterns from the style image, merely changing colors without forming distinct patterns, attributable to their limited receptive fields.

In Fig. 2 (b), we apply our proposed patch-wise optimization and perform nearest matching of two specific VGG blocks, namely `conv2` and `conv3`, independently. Each block processes different parts of the style image and independently stylizes colors and shapes based on their closest locations. As a result, despite the application of patch-wise optimization, the overall patterns still appear incomplete and overlapping.

In Fig. 2 (c), we ensure the alignment of matched locations across each block. As explained in Sec. A, the process begins with the nearest matching of `conv3`, where we identify the closest pairings between content features and style features. Subsequently, rather than performing matching with conv2, we employ the closest pairs from conv3 to optimize feature distances for conv2. This coordinated approach allows both blocks to stylize the content feature using identical locations from the style image, thereby yielding clearer and more complete patterns.

Finally, in Fig. 2 (d), we achieve an expansion of the receptive fields by defining local patches with dilation. This expansion allows each feature to fully capture the patterns from the style image, facilitating the creation of clearer and more complete pattern shapes, avoiding the issue of vague and flat surfaces.

Perspective style augmentation. In Fig. 3, we present a more detailed comparison of our perspective style augmentation, including verifications of each depth cue mentioned in Sec. A. Please note that Fig. 3 (a) corresponds to the top left figure in Fig. 8 of the main paper, and Fig. 3 (c) aligns with the top right figure in Fig. 8 (main paper). As shown in Fig. 3 (a), without perspective style augmentation, the overall patterns are created with similar size, color intensity, and level of color contrast. Moreover, as the overall stylization focuses on creating complete and accurate shapes of style patterns, the content structure, especially the fine-detailed geometry, tends to be lost or washed out.

In Fig. 3 (b), we apply perspective augmentation by using a reduced size of the style image for the farther surfaces, while the closer surfaces are stylized with a larger size of the style image. This approach successfully maps smaller patterns to the background while keeping the original size of patterns for the foreground objects, providing an increased sense of depth. However, it lacks a naturally structured feeling due to the absence of atmospheric perspective, where the distant region should appear less clear and less intense than the closer area.

In Fig. 3 (c), for a more enhanced effect, we also reduced the patch size for computing style loss for the distant surfaces to decrease the pattern accuracy in these areas. Additionally, the loss for the distant areas is computed using only the conv3 block, to achieve lower color intensity and contrast. Consequently, the overall stylization focuses on transferring complete shapes and clear colors of larger patterns for the foreground areas, while making the distant surfaces less clear. This approach provides a more enhanced feeling of depth through the size of shapes and color differences depending on surface distance. Moreover, the detailed structure is better maintained, providing improved differentiation between surface boundaries due to overlapping.

B.2. Interpolation of Deformation Fields

By utilizing deformation fields and combining them with the original grid, we can implement a smooth interpolation from the original 3D scene to the stylized scene. This is achieved by multiplying a scale factor s, which can range between $[0, 1]$, with $\Delta\mathbf{x}$, allowing us to smoothly interpolate the geometry between the original and the stylized shapes.

For the appearance, we can further utilize the original color grid, denoted as $\mathcal{G}_c$, in conjunction with $\mathcal{G}_{c'}$, which is the stylized color grid. When a sample point $\mathbf{x}_i$ is specified, the color at this point is defined by the following expression:

$$c_i = (1 - s) \cdot \mathcal{G}_c(\mathbf{x}_i) + s \cdot \mathcal{G}_{c'}(\mathbf{x}_i + s \cdot \Delta\mathbf{x}_i) \tag{1}$$

When $s = 0$, the resulting render is identical to the original scene. Conversely, when $s = 1$, the rendered scene is fully stylized. By gradually changing s from 0 to 1, we are able to render a smooth interpolation between these two states.

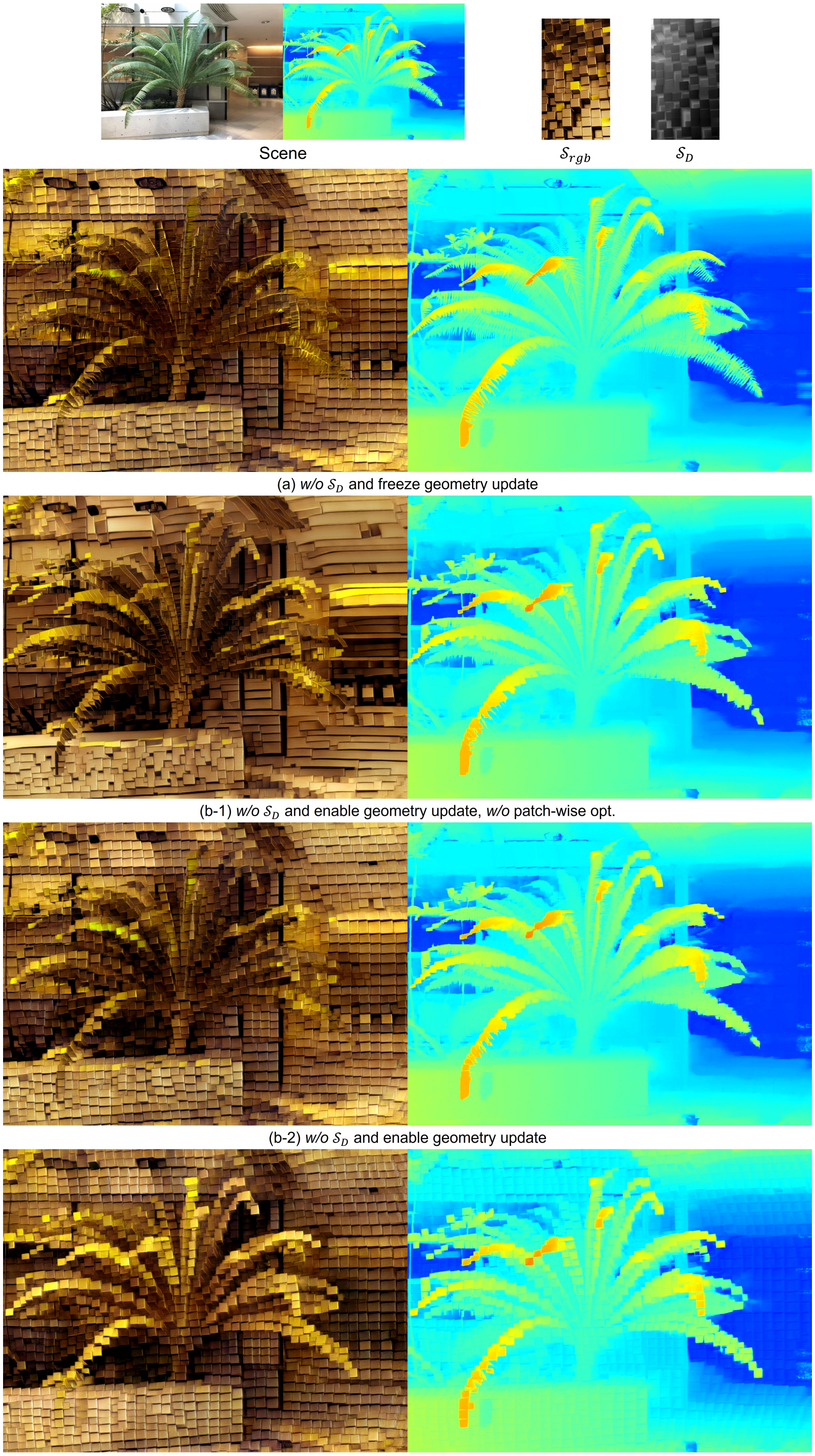
The results are available at our project page.

B.3. Partial Stylization

As mentioned in the main paper, it is possible to partially stylize a 3D scene, targeting specific classes as well as individual objects. In the main paper, due to space limitations, we demonstrated stylization focused only on specific classes. In Fig. 4 and Fig. 5, presented here, we provide additional examples of stylizing target objects.

C. Additional Qualitative Results

In Fig. 6, Fig. 7 and Fig. 8 we provide more stylized results of trex, fern, and horns scenes in the LLFF dataset [4].



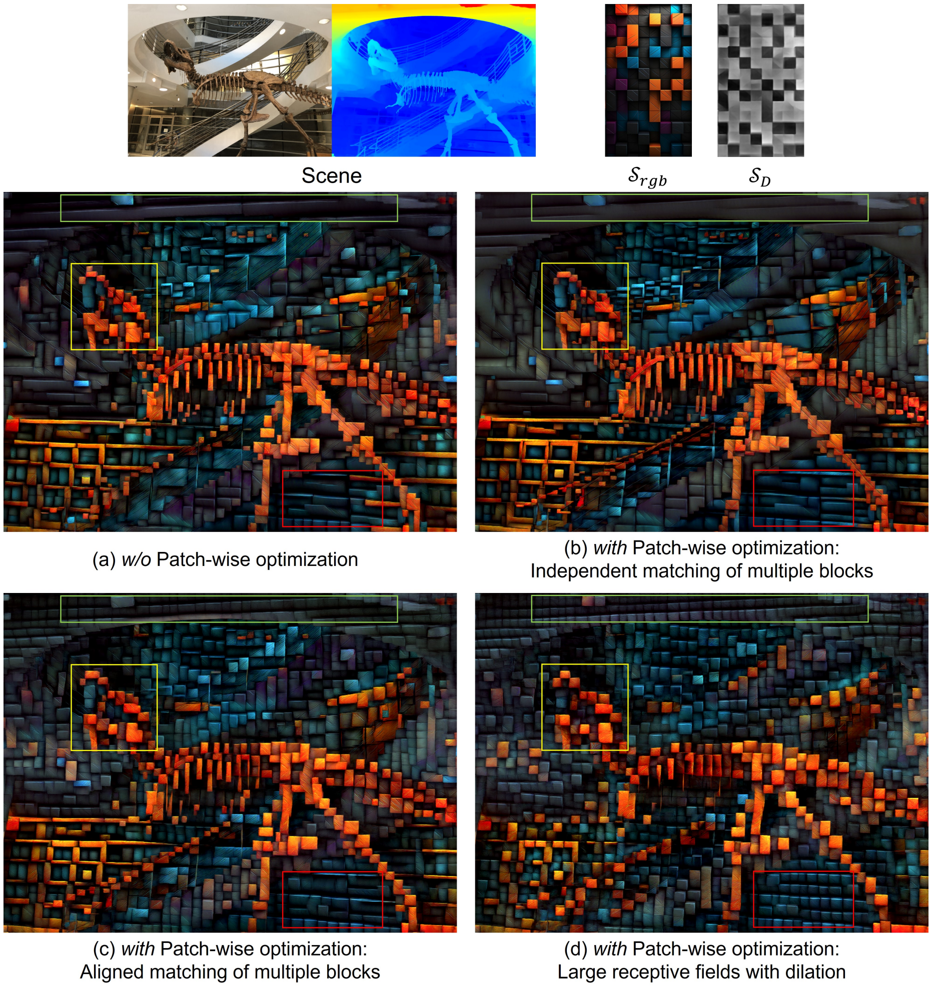
(a) *w/o* $\mathcal{S}_D$ and freeze geometry update

(b-1) *w/o* $\mathcal{S}_D$ and enable geometry update, *w/o* patch-wise opt.

(b-2) *w/o* $\mathcal{S}_D$ and enable geometry update

(c) *with* $\mathcal{S}_D$ and enable geometry update

Figure 1. Detailed ablation studies on using a depth map as a style guide.



Scene

$\mathcal{S}_{rgb}$ $\mathcal{S}_D$

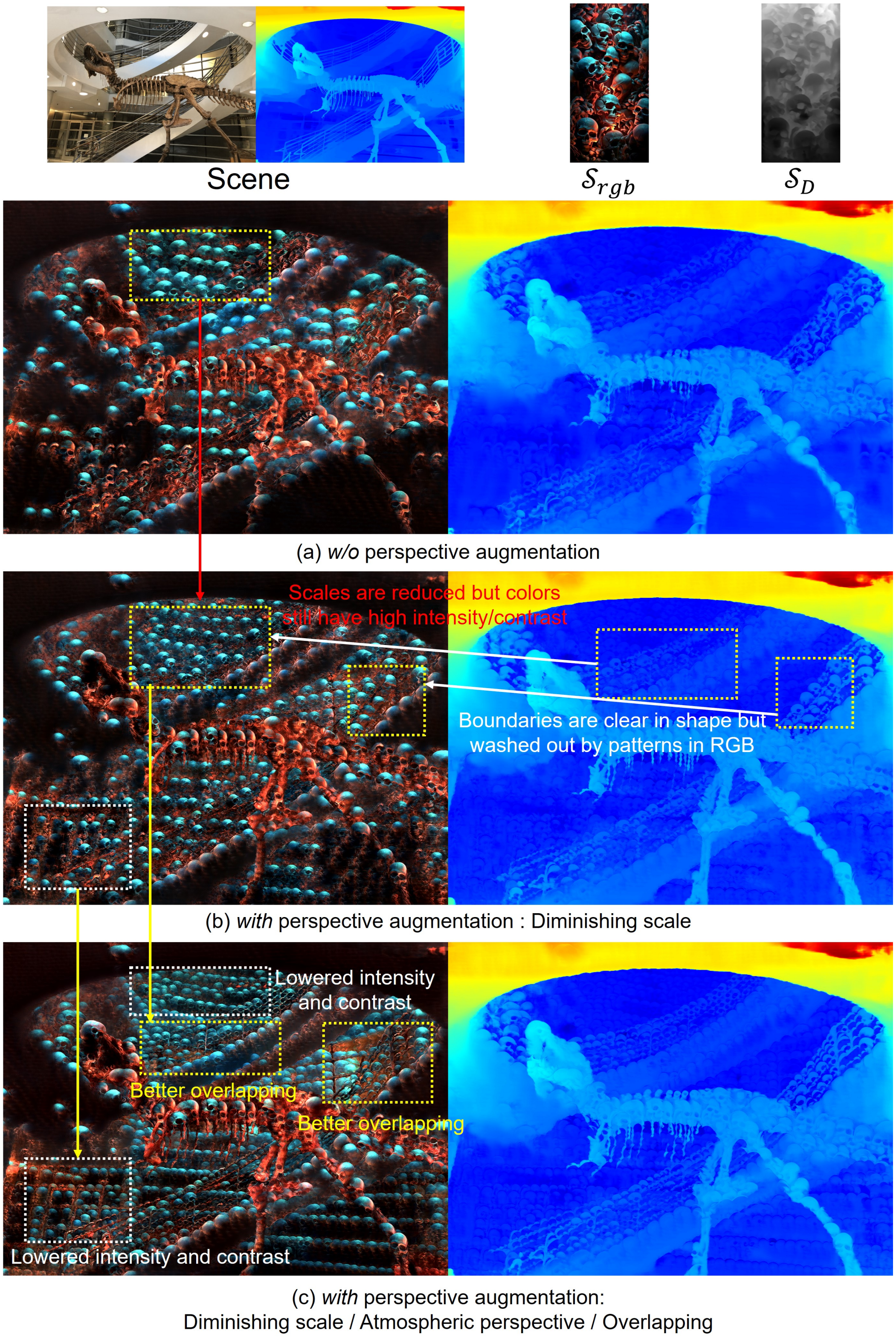
(a) *w/o* Patch-wise optimization

(b) *with* Patch-wise optimization: Independent matching of multiple blocks

(c) *with* Patch-wise optimization: Aligned matching of multiple blocks

(d) *with* Patch-wise optimization: Large receptive fields with dilation

Figure 2. Detailed ablation studies of patch-wise optimization.


Scene
$\mathcal{S}_{rgb}$
$\mathcal{S}_{D}$
Scales are reduced but colors still have high intensity/contrast
Boundaries are clear in shape but washed out by patterns in RGB
Lowered intensity and contrast
Better overlapping
Better overlapping
Lowered intensity and contrast

(a) *w/o* perspective augmentation

(b) *with* perspective augmentation : Diminishing scale

(c) *with* perspective augmentation:
Diminishing scale / Atmospheric perspective / Overlapping

Figure 3. Detailed ablation studies and analysis of perspective style augmentation.

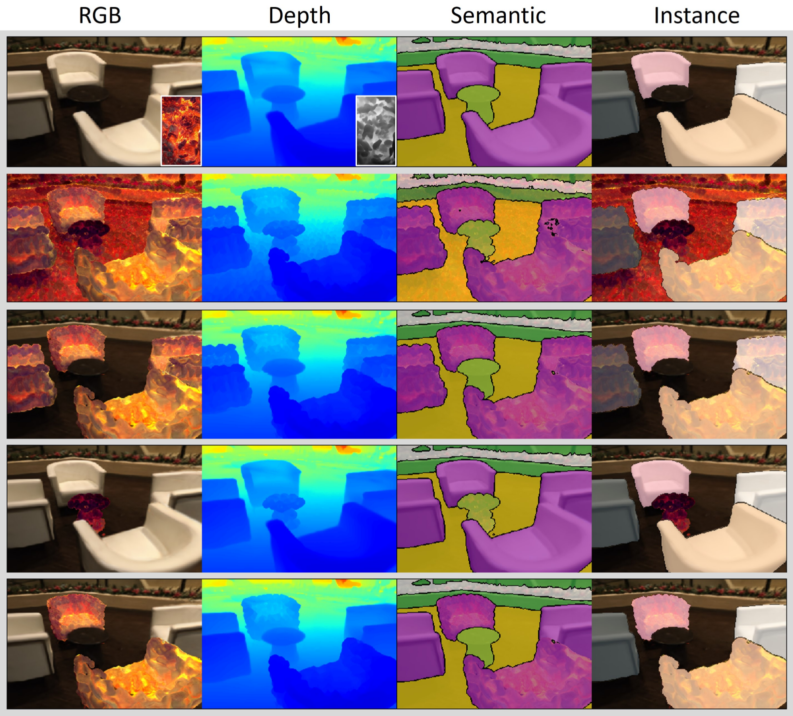


Figure 4. Qualitative results demonstrating partial stylization of the scene based on target classes or specific individual objects.

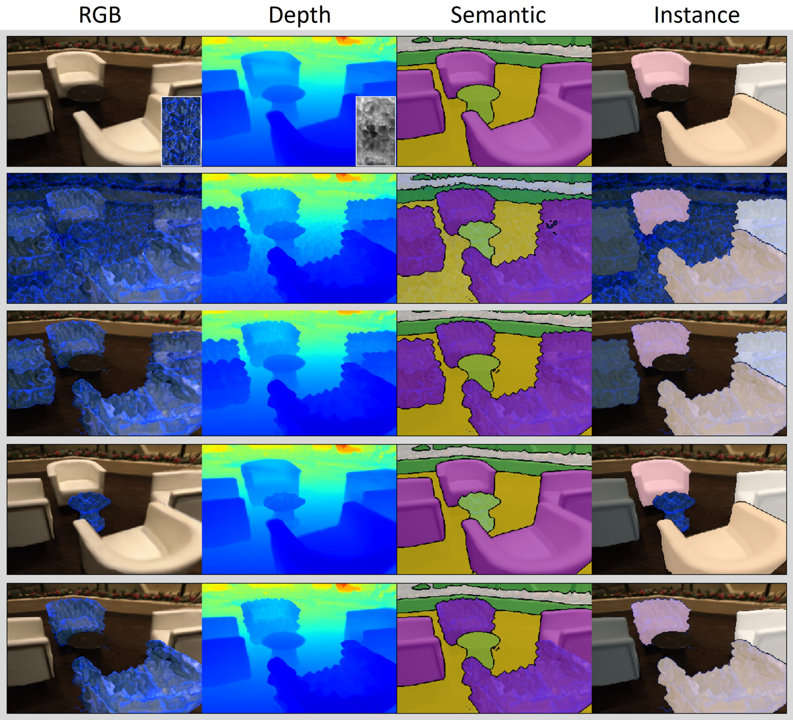


Figure 5. Qualitative results demonstrating partial stylization of the scene based on target classes or specific individual objects.

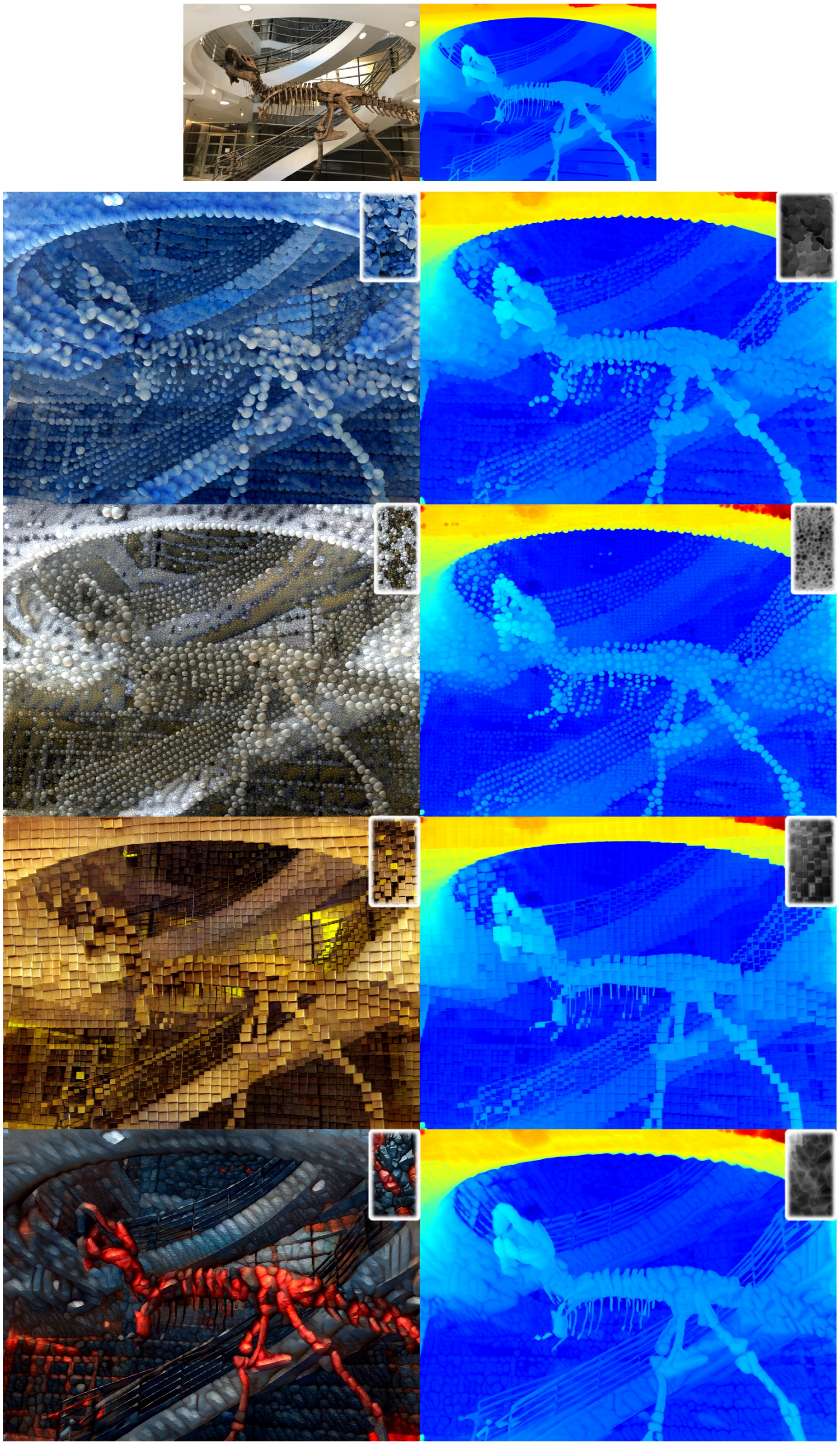


Figure 6. Qualitative results of the `trex` scene from the LLFF dataset.

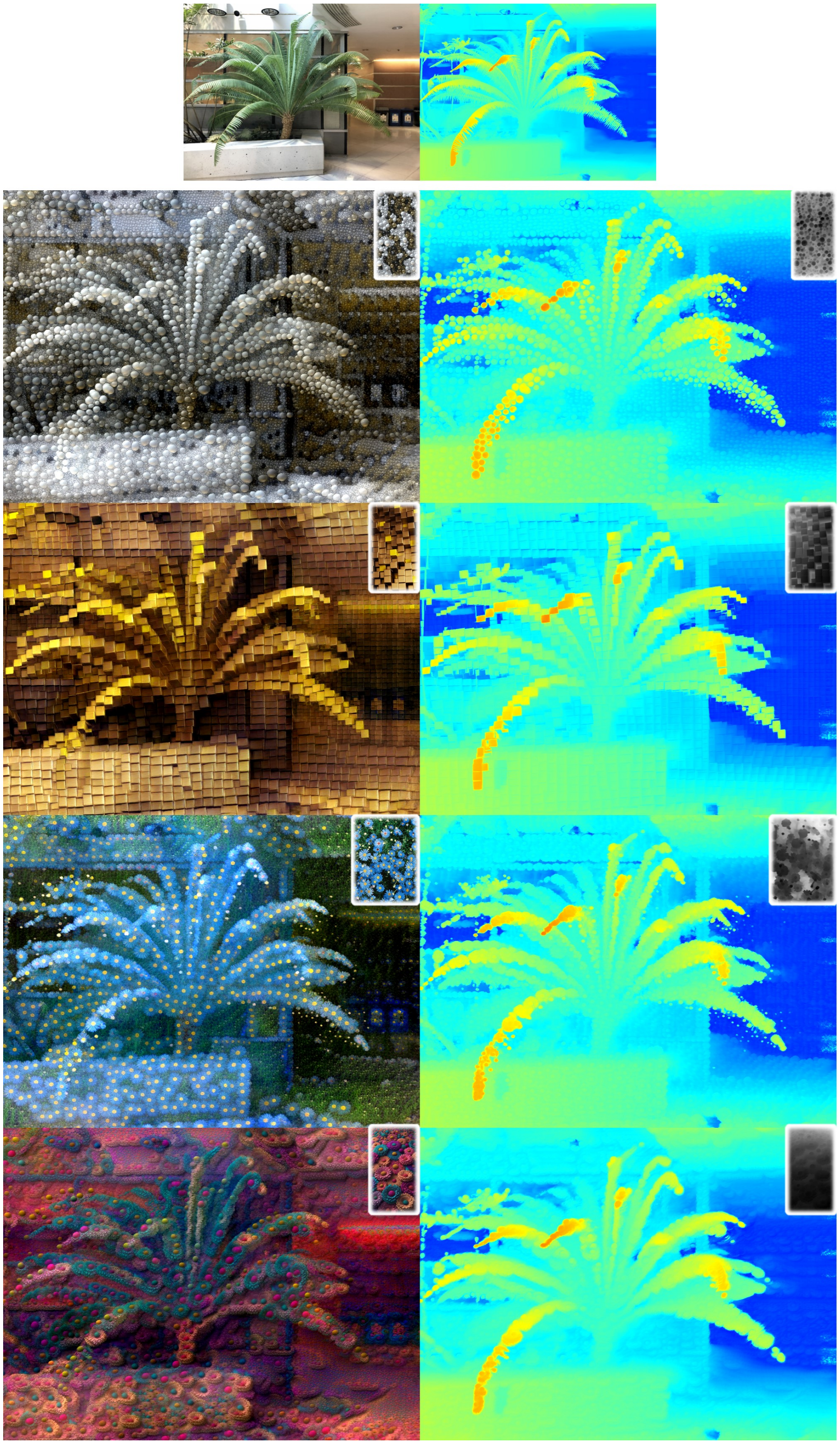


Figure 7. Qualitative results of the `fern` scene from the LLFF dataset.

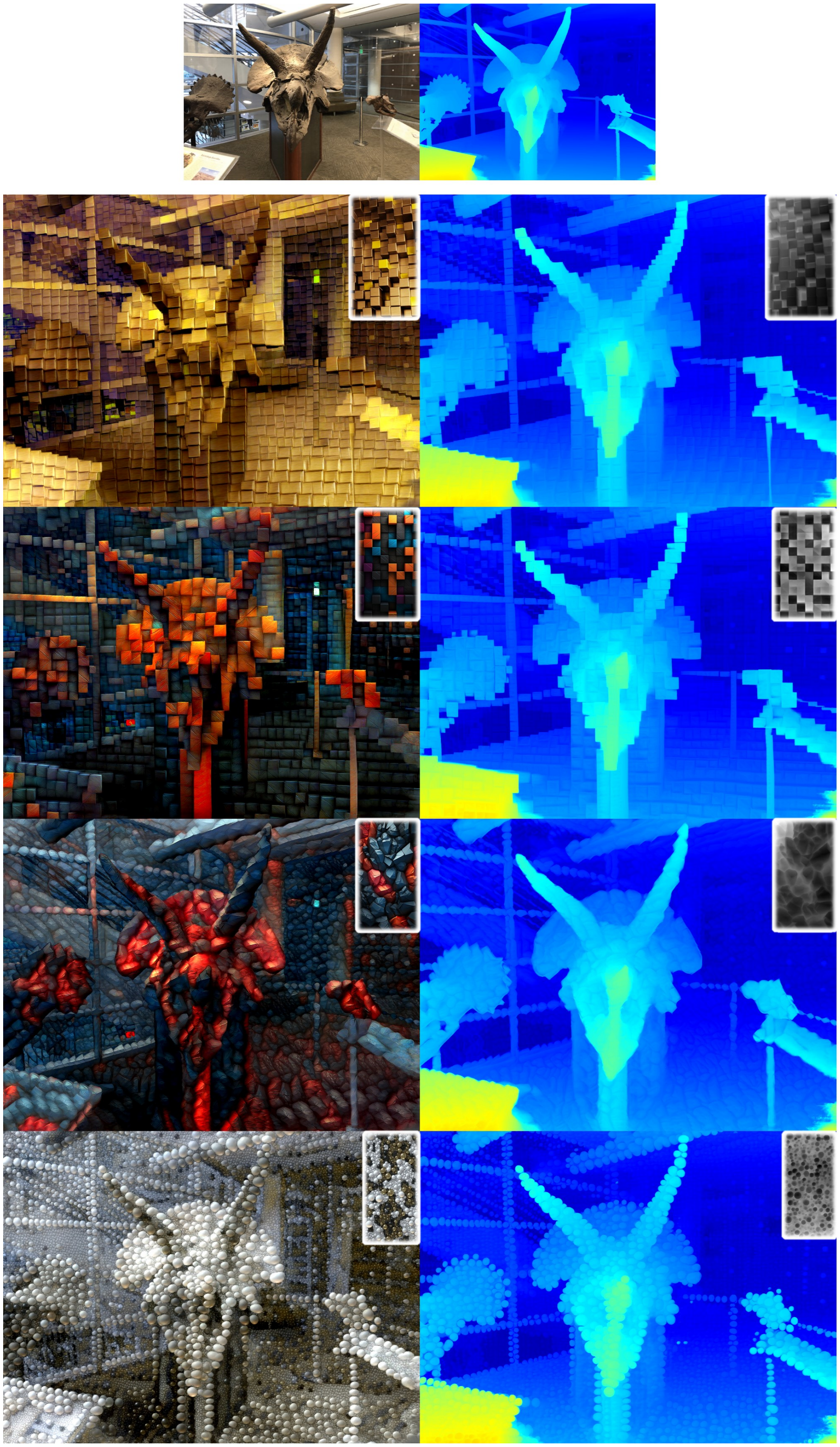


Figure 8. Qualitative results of the `horns` scene from the LLFF dataset.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1

[2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*. Springer, 2022. 1

[3] Nicholas Kolkin, Michal Kucera, Sylvain Paris, Daniel Sykora, Eli Shechtman, and Greg Shakhnarovich. Neural neighbor style transfer. *arXiv preprint arXiv:2203.13215*, 2022. 1

[4] 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, 2019. 3

[5] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[6] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *CVPR*, 2023. 1

[7] Marko Teittinen. Depth cues in the human visual system. `https://www.hitl.washington.edu/projects/knowledge_base/virtual-worlds/EVE/III.A.1.c.DepthCues.html`. 1

[8] Bruce Walters. Space: Indicators of depth on a flat surface. `https://augustana.net/users/arwalters/design/depth.htm`. 1

[9] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *ECCV*, 2022. 1, 2

[10] Yuechen Zhang, Zexin He, Jinbo Xing, Xufeng Yao, and Jiaya Jia. Ref-NPR: Reference-based non-photorealistic radiance fields for controllable scene stylization. In *CVPR*, 2023. 2