Artistic Style Novel View Synthesis Based on A Single Image

Kuan-Wei Tseng[1,2] Yao-Chih Lee[1] Chu-Song Chen[1]

[1]National Taiwan University, Taipei, Taiwan [2]Tokyo Institute of Technology, Tokyo, Japan

http://kuan-wei-tseng.github.io/ArtNV

Abstract

Recent progress in 3D display technologies has raised the demand for stylized 3D digital content. Previous approaches either perform style transfer on stereoscopic image pairs or reconstruct 3D environment with multiple view images. In this paper, we propose a novel view stylization framework that can convert a single 2D image into multiple stylized views. It is a two-stage solution that contains view synthesis and neural style transfer. We estimate dense optical flow between the source and novel views so that the style transfer model can produce consistent results. Experimental results show that our method significantly improves the consistency among views compared to the baseline method.

1. Introduction

Artistic style transfer can modulate the existing content and enrich the content diversity. With the advancement in display technologies, 3D digital content for Augmented Reality (AR), Virtual Reality (VR), and naked-eye 3D displays have attracted much interest. How to produce multiple views with consistent artistic styles in 3D thus becomes an interesting issue. It enables various applications, including artwork with a VR display, style enhancement in AR glasses, and artistic animation in 3D televisions.

Novel view synthesis aims to generate unseen views from source view(s), while artistic style transfer extracts and migrates the style of an example image to target images. To integrate them, we should maintain the consistency of style for every individual point between the views. A naive approach is to directly cascade the novel view synthesis and neural style transfer models. However, such a method neglects spatial consistency. Despite great quality for every single view, users can easily perceive discrepancies when they navigate into the scene. Another way is to first modulates the style of a source view and then synthesize multiple novel views. Although it seems to successfully circumvent the consistency problem, the visual quality tends to be poor because 3D reconstruction information on which view synthesis relies is destroyed after style transfer.

In this work, we propose ArtNV, a novel view stylization framework that can synthesize novel views with designated artistic styles from only a single image. Our approach simply requires a source image together with an example style image as input; then multiple views can be generated at designated camera poses with designated artistic styles. Our approach leverages view synthesis techniques and can then convert both the input and novel views using neural style transfer. To ensure spatial consistency, we estimate dense optical flow to find the correspondence between source and novel views. The experimental results reveal that the proposed method is simple yet effective.

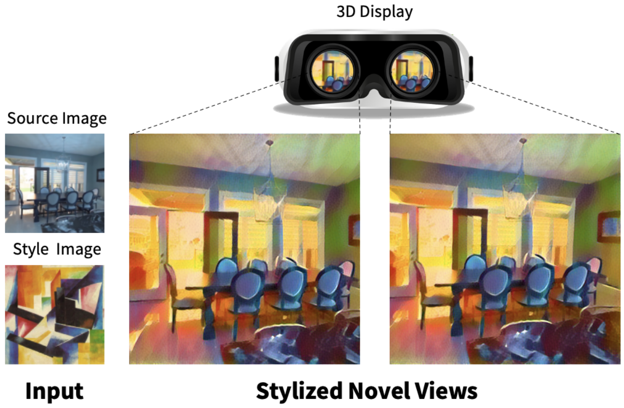

Figure 1. Result Demonstration

2. Related Works

We briefly review the relevant works including the novel view synthesis, neural style transfer, and their integration.

2.1. Novel View Synthesis

Novel View Synthesis is to generate unseen novel views using existing views. Learning-based view synthesis either utilizes implicit [3, 5, 16, 24] or explicit [9, 17, 21, 23, 26] representations to generate novel views. Neural Radiance Fields (NeRF) [16, 24] learn the 3D scenes with multilayer perceptrons (MLP). Given a 3D position and view direction, it renders novel views using the color and volume density regressed by MLPs, Notwithstanding, training NeRF requires multi-view images with camera poses, and the rendering process is notoriously slow. On the other hand, meth-

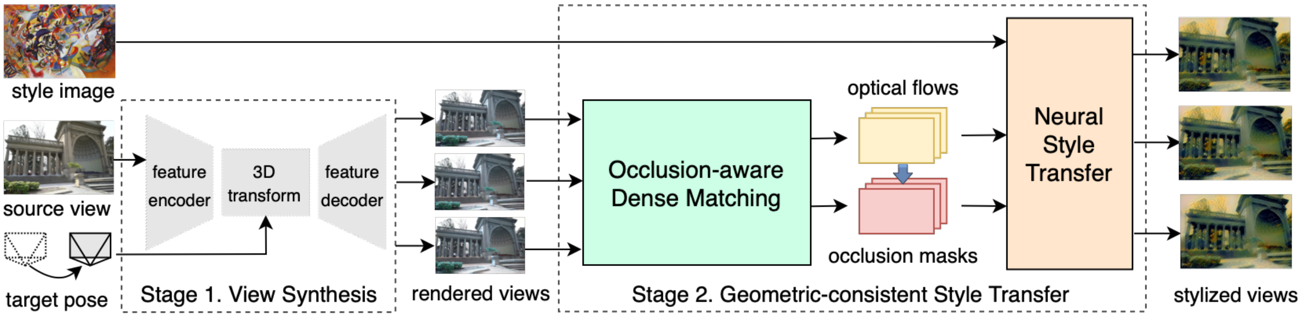

Figure 2. Pipeline of the proposed method. It is a two-stage method that contains a view synthesis stage and a style transfer stage. Before neural style transfer, we perform occlusion aware dense matching using optical flow to ensure spatial consistency.

ods using explicit representation generally estimate depth or multiplane images (MPI) to support view synthesis. SynSin [23] and PixelSynth [17] project the extracted CNN features of the input image into 3D space then reproject them on the novel image plane with differentiable rendering. Instead of a single depth image, MPI-based methods [21, 26] produce layered RGBA images with equaled space between layers. Each layer is warped and composited to synthesize novel views. One of the advantages of these methods is that they can produce novel views with only one source view.

2.2. Artistic Neural Style Transfer

Artistic Neural Style Transfer [7, 11, 13–15, 18] performs content-style separation and substitutes the style of template images for that of target images using deep neural networks. Image-optimization-based methods [7, 13, 14], also known as online methods, iteratively optimize the input images to make their style features approach to that of template image while preserving their content features. Building upon pre-trained VGG networks, these methods do not require any model training, but they suffer from slow test time optimization. Model-optimization-based methods [11, 15, 18], also known as offline methods, utilize an encoder-decoder architecture. Content and style features are extracted by the encoder and modulated at feature level before being decoded by the decoders. Though requiring a training domain, offline methods can operate in real-time during test time.

2.3. 3D Image Stylization

Stereoscopic Neural Style Transfer [2, 8] literally perform style transfer on stereo image pairs. They take a stereo pair as input and estimate disparity and occlusion maps with subnetworks to warp feature maps of stereo images. With the feature map of the left and right views, the decoder can produce two stylized images. However, a limitation is that these methods require a stereo image pair as input. Huang et al. [10] propose an approach that integrates the novel view synthesis and style transfer based on image sequences. It utilizes traditional Structure from Motion (SfM) tool to reconstruct the 3D point cloud of the scene and then performs style transfer directly on the point cloud before rendering on the novel image planes. Although it achieves globally consistent visual quality, it relies on delicate and time-consuming SfM that needs multiple views as inputs. Chiang et al. [4] introduce a 3D scene stylization using implicit representation. It adopts a hypernetwork to update the MLP weights that controls the color in NeRF, making it possible to generate stylized novel views. However, these approaches leverage multi-images or image sequences as inputs. We propose an approach that can produce consistent stylized images in 3D based on only a single image in this paper, which enforces more flexibility for the applications.

3. Proposed Method

3.1. Overview

The pipeline of our ArtNV is shown in Fig.2. It consists of two stages, *view synthesis* and *style transfer*. Our method takes a source image I_0 and a style image I_s as input, and produces stylized source view I'_0 and N stylized novel views $\{I'_i\}_{i=1:N}$. At the first stage, we utilize SynSin [23], an explicit multi-view synthesis method to produce novel views. Note that this module is interchangeable with other view synthesis methods such as [9, 17].

The second stage is neural style transfer with spatial consistency. View synthesis methods such as SynSin [23] provides dense depth information which can be used to compute the disparity. However, we found that the provided depth maps are blurry and inaccurate, which produce erroneous image correspondence. It could be because that common neural scene-synthesis methods (such as SynSin [23]) tends to find only sub-optimal depths since their objective focuses on the net effect of rendering the novel view via both depths and pose-related texture information.

In our method, we estimate the dense optical flow using RAFT [20] between the source view and novel views instead. To address the occlusion and noisy flows, we perform forward-backward consistency check to ensure the reliability of optical flow. With the estimated correspondence among views, we add an Occlusion-aware Consistency Loss

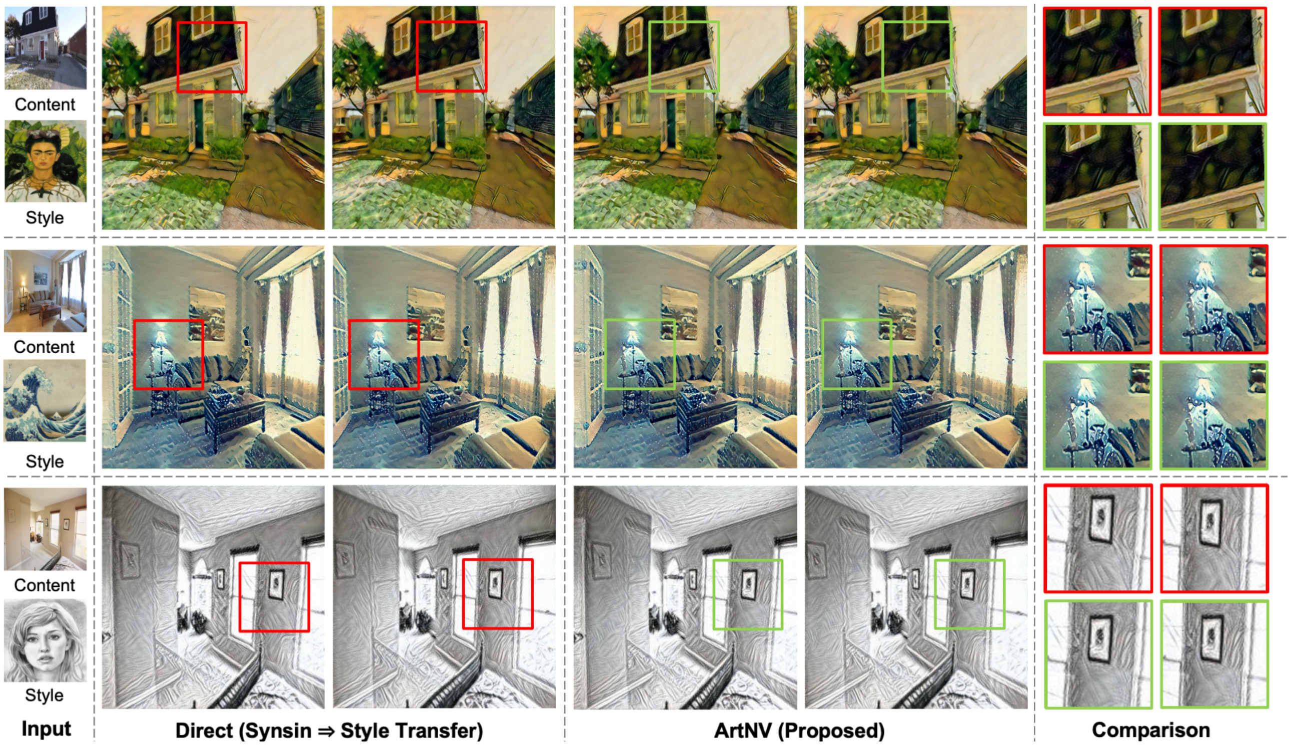

Figure 3. Qualitative results and the comparison between direct and the proposed method with occlusion-aware dense matching.

to the optimization objectives in addition to the vanilla image optimization-based style transfer. Without loss of generality, our current approach uses the online method [7] for neural style transfer. Offline methods (*e.g.* [11]) can be used as well by fine-tuning the pre-trained decoder with our Occlusion-aware Consistency Loss for spatial consistency.

3.2. Optimization Objectives

We conduct three losses, style loss $\mathcal{L}_{style}$, content loss $\mathcal{L}_{content}$ and occlusion-aware consistency loss $\mathcal{L}_{corr}$.

Content Loss preserves the content information during the stylization by enforcing the similarity of high-level features obtained from a pre-trained neural network model.

$$\mathcal{L}_{content} = \sum_{i\in[0,N]} \sum_{l\in\{l_c\}} \left\|\mathcal{F}^l\left(I_i\right) - \mathcal{F}^l(I_i')\right\|^2, \tag{1}$$

where l_c are the layers for computing the content loss, and $\mathcal{F}^l$ are the feature maps.

Style Loss migrates the style information from style image I_s to target images I_i' using the correlations between features maps. We use the pre-trained VGG-19 [19] to extract the feature maps since its layer-wised structure without skip connections suits better for artistic style transfer. As explained in [22], advanced object classification models such as ResNet are inappropriate for style transfer as the residual connection would diminish the entropy of the feature map, resulting in failure of high level style patterns.

$$\mathcal{L}_{style} = \sum_{i\in[0,N]} \sum_{l\in\{l_s\}} \left\|G\left(\mathcal{F}^l\left(I_s\right)\right) - G\left(\mathcal{F}^l(I_i')\right)\right\|^2, \tag{2}$$

where l_s are the layers for computing the style loss, and G is the Gram Matrix for feature correlation.

Occlusion-aware Consistency Loss ensures the spatial consistency between the stylized views. It utilizes RAFT [20], a state-of-the-art optical flow to find the inter-view correspondence. Since there might be occlusion or mismatch flows, we estimate bidirectional optical flows, i.e., the flow $F^{s\to t}$ from the source view I_0 to the target views I_i and the flows $F^{t\to s}$ from the target view I_i to the source view I_0.

$$\begin{aligned}\mathcal{L}^v_{\text{corr}} = \sum_{i=1}^{N} M_i^{t\to s} \odot \left\|I_0' - \mathcal{W}\left(I_i', F_i^{t\to s}\right)\right\|^2 + \\ M_i^{s\to t} \odot \left\|I_i' - \mathcal{W}\left(I_0', F_i^{s\to t}\right)\right\|^2,\end{aligned} \tag{3}$$

where $\mathcal{W}(I,F)$ is the warping operator that warps the image I via optical flow F. $M_i^{t\to s}$ is the occlusion mask obtained by the forward-backward consistency check.

Total Loss is the combination of the content loss, style loss, and occlusion-aware consistency loss.

$$\mathcal{L} = \lambda_1\mathcal{L}_{content} + \lambda_2\mathcal{L}_{style} + \lambda_3\mathcal{L}_{corr}, \tag{4}$$

Table 1. Experimental comparison between the Direct and the proposed (ArtNV) methods. Left-hand side shows the quantitative results on the LPSIS metric, which reveal that our ArtNV outperforms the Direct method consistently for all scenes. The remaining part shows the results of user preference study. Note that our method is also more favorable on all criteria with statistical significance (p value < 0.05).

	Tanks and Temples (Quantitative Study; warped LPIPs metric)						RealEstate 10K (User Study; 7 point Likert scale)			
Method	Ballroom	Barn	Church	Courtroom	Museum	Temple	Colors	Strokes	Points	Flickering
Direct	0.3519	0.3517	0.3240	0.3441	0.3387	0.3605	5.4938	5.1750	5.3125	2.9750
ArtNV	**0.3300**	**0.3198**	**0.2972**	**0.3190**	**0.3183**	**0.3346**	**5.9000**	**5.7500**	**5.6625**	**5.6375**

with $\lambda_1, \lambda_2, \lambda_3$ the hyperparameters for combination.

3.3. Implementation Details

We use SynSin model trained on RealEstate10K [26] datasets for the view synthesis stage; it covers common indoor and outdoor scenes. The source image is resized into 256×256 before being fed into the view synthesis network. The dense matching is achieved by RAFT [20] and we adopt the default settings. For the style transfer stage, we optimize the source and novel views resized into 512×512 with the content and style loss computed with pre-trained VGG-19 network [19]. Following the convention of online methods, we use Limited-memory BFGS (LBFGS) as our optimizer. We optimize each image for 300 iterations. The hyperparameters are $\lambda_1 = 1e6$, $\lambda_2 = 1$, and $\lambda_3 = 3000$.

4. Experimental Results

As there is no previous study that requires only a single view for 3D-consistent artistic style transfer, we compare our method with the baseline approach that cascades the view synthesis and style transfer models directly (called the **Direct Method**). We conduct quantitative, qualitative, and user preference studies to evaluate our **ArtNV method**.

4.1. Quantitative Results

We perform stylized view synthesis using content images from Tanks and Temples [12] dataset and 21 style images from [6]. Tanks and Temples are video sequences taken by handheld camera at different scenes. We test our methods on six different scenes, including *Ballroom, Barn, Church, Courtroom, Museum*, and *Temple*, with a 2410 images in total. For each image in the sequence, we generate 12 surrounding views and randomly select one style image to perform style transfer. The comparison between the baseline Direct and proposed ArtNV methods are shown in the left part of Table 1. We use the warped LPIPS metric [25] to evaluate the spatial consistency between novel views,

$$E_{\text{warp}}\left(\mathbf{I}_i, \mathbf{I}'_j\right) = \text{LPIPS}\left(\mathbf{I}_i, \mathcal{W}\left(\mathbf{I}'_j, F_j^{t \to s}\right), \mathbf{M}_j^{t \to s}\right). \tag{5}$$

The results show that our method outperforms the baseline method significantly in terms of perceptual similarity.

4.2. Qualitative Results

We also report results on RealEstate10K [26] with style images sampled from the WikiArt [1]. RealEstate10K collects indoor and outdoor images from YouTube videos. The visual comparison between the baseline Direct and the proposed ArtNV methods are shown in Fig. 3. The proposed method achieves better spatial consistency in terms of strokes, dots, and colors in each image pair. The number of dots, color of patches, and the position of strokes are inconsistent without using the proposed multi-view constraints.

4.3. User Preference Study

We recruit 16 users (10 males and 6 females) to evaluate our method. We show 10 stereo image pairs sequentially on a large screen and ask users to grade their consistency in terms of colors, strokes, and points of each pair. We also show 10 videos consisting of 12 novel views and let users judge the flickering effect. We adopt 7 Point Likert Scale as our grading metrics, higher is better. The results are shown in the right part of Table 1. The one-way ANOVA test shows that most users believe the proposed methods outperform the Direct method with statistical significance.

5. Conclusions and Future Works

We propose a flexible framework for stylized novel view synthesis that requires only a single content image. Our ArtNV method first performs view synthesis, which is then followed by a spatially consistent style transfer. It leverages the dense optical flow and occlusion map to achieve spatial consistency between stylized views. Experimental analyses reveal that the proposed method can synthesize more consistent stylized novel views. Our method is potentially extendable to a general solution handling both individual images and videos. As for future work, we would like to integrate the stylized view synthesis method with 3D display and create interactive applications.

Acknowledgement

This work was partially supported by Ministry of Science and Technology in Taiwan (MOST 110-2634-F-002-050 and MOST 110-2218-E-002-033-MBK).

References

[1] Visual art encyclopedia. available at https://www.wikiart.org/.

[2] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stereoscopic neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] Wen-Cheng Chen, Min-Chun Hu, and Chu-Song Chen. Str-gqn: Scene representation and rendering for unknown cameras based on spatial transformation routing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5966–5975, October 2021.

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

[5] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

[6] Wei Gao, Yijun Li, Yihang Yin, and Ming-Hsuan Yang. Fast video multi-style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[7] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Xinyu Gong, Haozhi Huang, Lin Ma, Fumin Shen, Wei Liu, and Tong Zhang. Neural stereoscopic image style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[9] Ronghang Hu, Nikhila Ravi, Alexander C. Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12528–12537, October 2021.

[10] Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, and Ming-Hsuan Yang. Learning to stylize novel views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] 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.

[13] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] Shaohua Li, Xinxing Xu, Liqiang Nie, and Tat-Seng Chua. Laplacian-steered neural style transfer. In *Proceedings of the 25th ACM international conference on Multimedia (MM)*, pages 1716–1724, 2017.

[15] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6649–6658, October 2021.

[16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020.

[17] Chris Rockwell, David F. Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14104–14113, October 2021.

[18] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2018.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[20] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419. Springer, 2020.

[21] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[22] Pei Wang, Yijun Li, and Nuno Vasconcelos. Rethinking and improving the robustness of image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 124–133, June 2021.

[23] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

[25] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[26] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4), 2018.