NOVA: NOvel View Augmentation for Neural Composition of Dynamic Objects

Dakshit Agrawal*[1] Jiajie Xu*[1] Siva Karthik Mustikovela [2]
Ioannis Gkioulekas [1] Ashish Shrivastava [2] Yuning Chai [2]

[1] Carnegie Mellon University [2] Cruise LLC

Abstract

We propose a novel-view augmentation (NOVA) strategy to train NeRFs for photo-realistic 3D composition of dynamic objects in a static scene. Compared to prior work, our framework significantly reduces blending artifacts when inserting multiple dynamic objects into a 3D scene at novel views and times; achieves comparable PSNR without the need for additional ground truth modalities like optical flow; and overall provides ease, flexibility, and scalability in neural composition. Our codebase is on GitHub.

1. Introduction

Photo-realistic composition of objects in a 3D scene has significant applications, one of which is creating realistic content and experiences inside the Metaverse. Despite recent advances in neural radiance fields (NeRFs) [9], photo-realistic composition from dynamic monocular videos remains a challenging problem. This is primarily due to the ill-posed nature of this task—multiple scene configurations can lead to identical observed image sequences, a problem we refer to as the 3D structure ambiguity.

Current approaches for this task [2, 7] build implicit representations of the static scene and dynamic objects separately by predicting a per-point blending factor along with color and density. To deal with structure ambiguity, these methods also predict modalities such as 3D scene flow and depth to regularize the prediction within each frame and between neighboring frames. This requires ground truth data for these modalities, thus limiting applicability. These approaches also suffer from blending mask prediction errors when rendering a novel view, causing blending artifacts at the boundaries of the image that are not present in the reference frustum. This effect is amplified when inserting multiple objects into the scene and dramatically degrades the rendering quality (see Fig. 1).

We introduce a framework, NOVA, that helps mitigate these issues. NOVA reduces blending artifacts by augmenting NeRF with losses for different views during training and requiring the network to predict consistent masks and colors across novel views. NOVA additionally extends prior works to facilitate learning different dynamic objects of the scene using separate implicit representations and controlling their movement by manipulating these representations. NOVA does not require 3D scene flow regularization, thus removing the need for a scene flow predictor during data preparation and reducing training time without impacting PSNR. In summary, our contributions are three-fold:

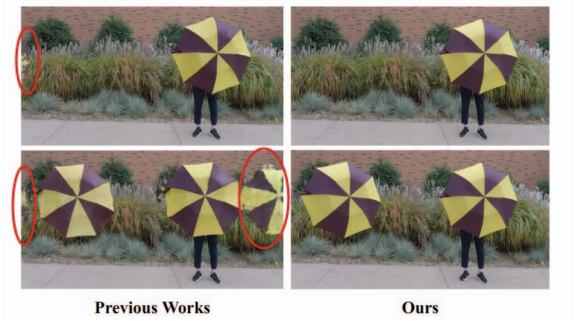

Figure 1: Prior works (left column) have blending artifacts that are amplified when multiple objects are inserted at different points in the same scene. Our method (right column) reduces these blending artifacts significantly.

1. a flexible NeRF composition framework to add an arbitrary number of dynamic objects into a static 3D scene;

2. a novel-view augmentation strategy for learning better per-point blending factors;

3. corresponding novel-view losses for high rendered image fidelity.

2. Related Work

Object composition via inverse-rendering. Inserting objects into a scene requires properties like lighting, depth, geometry, and material. [3, 14, 8, 24] estimate these properties for an indoor scene from a single image. For outdoor scenes, a high dynamic range light field is necessary to rep-

*Equal contribution

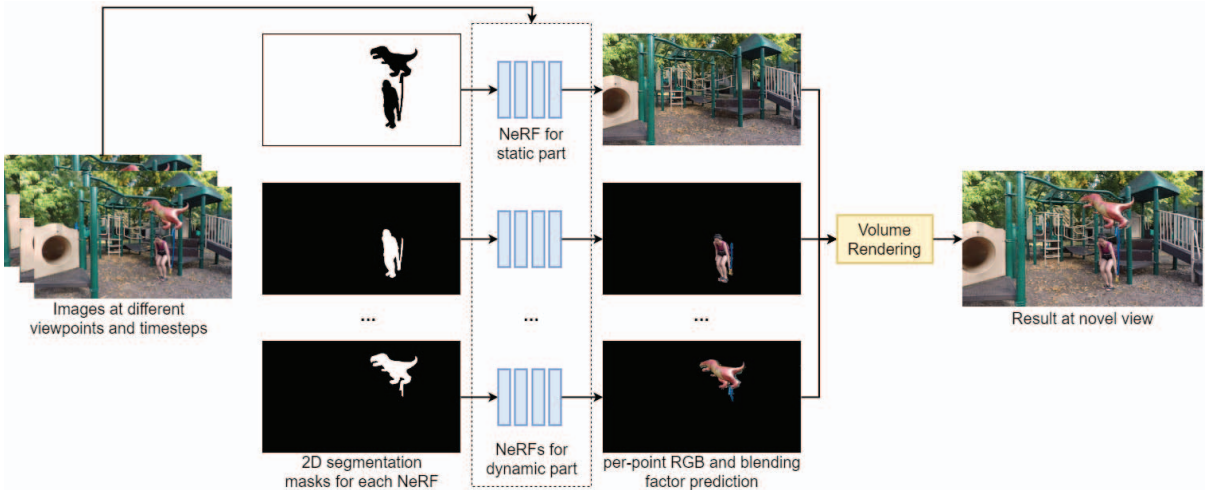

Figure 2: Overview of our training framework. Based on the 2D segmentation masks, separate NeRFs are initialized. These NeRFs predict per-point RGB color and blending factors, which are passed through a differentiable volume renderer to generate the final composed image from a novel viewpoint.

resent sun and sky [5, 20], and adversarial methods are commonly used to train photo-realistic results [20, 6, 19, 10].

Composing dynamic objects using NeRFs. NeRFs [9] achieve impressive novel-view synthesis results with a simple formulation for static scenes, encouraging research to compose multiple NeRFs. Guo *et al*. [4] proposed training per-object scattering functions for proper lighting effects during composition. Yang *et al*. [22] separated the scene into background and object branches, using 2D segmentation as supervision. To allow for 3D pose control, Ost *et al*. [11] proposed a learnable scene graph to decompose dynamic objects into nodes encoding transformation and radiance. Tancik *et al*. [15] proposed a framework to tune and compose individually trained NeRFs into city-scale scenes.

Novel-view synthesis for dynamic videos. Current works either learn a static canonical radiance field, with a second per-time-step field to apply deformation [17, 12, 13], or learn a dynamic radiance field directly conditioned on time [7, 21, 1, 2, 16]. For the latter direction, it is common to learn a scene flow field [18] concurrently and constrain adjacent frames for pixel consistency. Besides scene flow, Li *et al*. [7] also applied geometric consistency and depth as prior; Gao *et al*. [2] introduced additional auxiliary losses. Tian *et al*. [16] propose a flow-based feature aggregation module to incorporate spatial and temporal features.

3. Method

Our framework is inspired by Gao *et al*. [2], which jointly trains two NeRFs that separately handle the time-invariant static and time-varying dynamic parts of a monocular video. The static NeRF predicts the per-point color and density (c, σ) given the point's position and viewing direction (x, y, z, θ, ϕ). The dynamic NeRF predicts the per-point color, density, scene flow, and blending factor $(c, \sigma, s_f, s_b, \beta)$ given the point's position, viewing direction, and time $(x, y, z, \theta, \phi, t)$. Ground-truth optical flow is used to learn the scene flow, and several regularizing losses are applied to scene flow and depth to resolve the 3D structure ambiguity when learning from a monocular view. The NeRF composition is done in an unsupervised manner using the per-point blending factors β.

This approach works well for scene reconstruction but produces blending artifacts when manipulating the scene (see Fig. 1). We introduce a framework with three novel modules to alleviate these issues, described in detail in the subsequent sections. We also remove the losses based on ground-truth optical flow in Gao *et al*. [2] from our framework to reduce the amount of supervision.

3.1. Multiple NeRFs

Our framework uses separate NeRFs to learn different parts of the scene. Each NeRF is provided a segmentaion mask of the scene and is either static or dynamic based on the dynamicity of the scene parts it models (see Fig. 2). The static and dynamic NeRF architectures are similar to that of Gao *et al*. [2]. The final RGB image is produced from a novel viewpoint by combining the outputs of all the NeRFs as follows:

$$\mathbf{C}_P^{full}(\mathbf{r}) = \sum_{k=1}^{K} T_k^{full} \left(\sum_{n=1}^{num\_NeRFs} \alpha_k^n \beta_k^n \mathbf{c}_k^n \right) \quad (1)$$

where K is the number of samples along the ray $\mathbf{r}$, T_k^{full} is the transmittance at the k^{th} sample along the ray after accounting for rays from all the NeRFs, and α_k^n, β_k^n, and $\mathbf{c}_k^n$

are the alpha, blending factor, and color respectively predicted by the n^{th} NeRF for the k^{th} sample along the ray.

3.2. Novel-View Augmentation

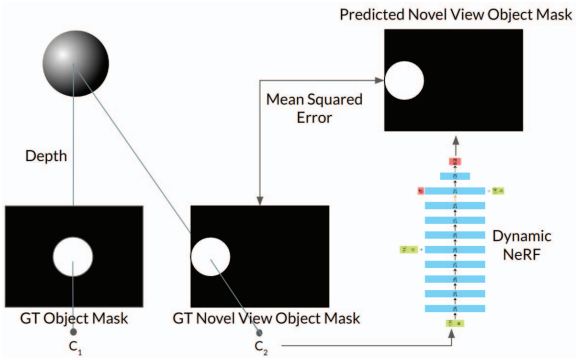

Figure 3: Novel-view augmentation training strategy

Our novel-view augmentation training strategy reduces blending artifacts when manipulating multiple dynamic objects and composing them into the scene. During training, we shift the camera responsible for the dynamic object to a novel view (see Fig. 3). Given the camera's relative transformation, we calculate the ground truth segmentation mask at the novel view using stereo geometry. Points are sampled along the rays of the camera at the novel viewpoint C_2 and passed through the corresponding NeRF. We render the predicted segmentation mask $\mathbf{M}_P^n$ for the n^{th} NeRF as follows:

$$\mathbf{M}_P^n(\mathbf{r}) = \sum_{k=1}^{K} T_k^{full} \alpha_k^{full} \beta_k^n \tag{2}$$

where T_k^{full} is the transmittance and α_k^{full} is the alpha at the k^{th} sample along the ray after accounting for rays from all the NeRFs, and β_k^n is the blending factor predicted by the n^{th} NeRF at the k^{th} sample along the ray. This augmentation strategy can be applied to other ground truths available for training like RGB images.

3.3. Novel-View Losses

We introduce a few losses to ensure high image fidelity when placing objects at novel points in the scene.

Novel-View Mask Loss. We take the squared error loss between the predicted and ground-truth masks for the novel viewpoint:

$$\mathcal{L}_{nvm} = \sum_{n=1}^{num\_NeRFs} \sum_{ij} \|\mathbf{M}_{GT}^n(\mathbf{r}_{ij}) - \mathbf{M}_P^n(\mathbf{r}_{ij})\|_2 \tag{3}$$

Per-Camera Novel-View RGB Loss. We render the RGB image of each NeRF as follows:

$$\mathbf{C}_P^n(\mathbf{r}) = \sum_{k=1}^{K} T_k^n \alpha_k^n \beta_k^n \mathbf{c}_k^n \tag{4}$$

We take the squared error loss between the predicted and the ground-truth RGB image from the novel viewpoint of only the pixels for which the NeRF is responsible for:

$$\mathcal{L}_{nvcn} = \sum_{n=1}^{num\_NeRFs} \sum_{ij} \mathbf{M}_{GT}^n(\mathbf{r}_{ij}) \|\mathbf{C}_{GT}(\mathbf{r}_{ij}) - \mathbf{C}_P^n(\mathbf{r}_{ij})\|_2 \tag{5}$$

Full Novel-View RGB Loss. After rendering the final RGB image using Eq. 1, we take the squared error loss with the ground truth full RGB image as follows:

$$\mathcal{L}_{nvcf} = \sum_{ij} \|\mathbf{C}_{GT}(\mathbf{r}_{ij}) - \mathbf{C}_P^{full}(\mathbf{r}_{ij})\|_2 \tag{6}$$

Blending Loss. To ensure the contributions of all the NeRFs for a particular point sum to one, we introduce a blending loss:

$$\mathcal{L}_{nvb} = \sum_{ijk} \left| \left(\sum_{n=1}^{num\_NeRFs} \beta_{ijk}^n \right) - 1 \right| \tag{7}$$

Alpha Loss. We force the NeRFs to not predict anything outside the masks they are responsible for by explicitly adding a loss for alphas to be 0 outside the camera mask:

$$\mathcal{L}_{nva} = \sum_{n=1}^{num\_NeRFs} \sum_{ij} (1 - \mathbf{M}_{GT}^n(\mathbf{r}_{ij})) \cdot \left(\sum_{k} |\alpha_{ijk}^n| \right) \tag{8}$$

4. Experimental Results

4.1. Dataset

We use the preprocessed Dynamic Scene Dataset [23] provided by Gao *et al*. [2], which contains video sequences for seven scenes, each consisting of a static background and moving objects. Each sequence has 12 images captured at different time steps and camera poses, which make them effectively monocular.

4.2. Evaluation

4.2.1 Quantitative Evaluation

We evaluate the image fidelity quantitatively by assessing the PSNR between the synthesized image and the corresponding ground truth image at a fixed viewpoint but changing time. Our framework performs comparably to other methods without the need for additional modalities of ground truth data like optical flow (see Tab. 1).

4.2.2 Qualitative Evaluation

We compare our novel-view renderings with Gao *et al*. [2] in Fig. 4. Our framework reduces blending artifacts, as visible clearly from our predicted object masks and generated final images, with the improvement being significant when composing multiple dynamic objects.

Method	Balloon1	Balloon2	Jumping	Playground	Skating	Truck	Umbrella	Average
NeRF + time	17.32	19.66	16.72	13.79	19.23	15.46	17.17	17.05
Yoon *et al*. [23]	18.74	19.88	20.15	15.08	21.75	21.53	20.35	19.64
Li *et al*. [7]	21.35	24.02	24.10	20.85	28.88	23.33	22.56	23.58
Gao *et al*. [2]	21.43	26.59	23.57	23.74	31.92	25.50	22.68	25.06
Ours	21.52	25.08	20.27	22.31	27.73	23.31	23.08	23.33

Table 1: We compare PSNR of our method against other methods that report their PSNR on Dynamic Scene Dataset [23]. The best results are highlighted in red while the second best are in blue. Our model performs comparably to other methods despite not using ground-truth optical flow supervision.

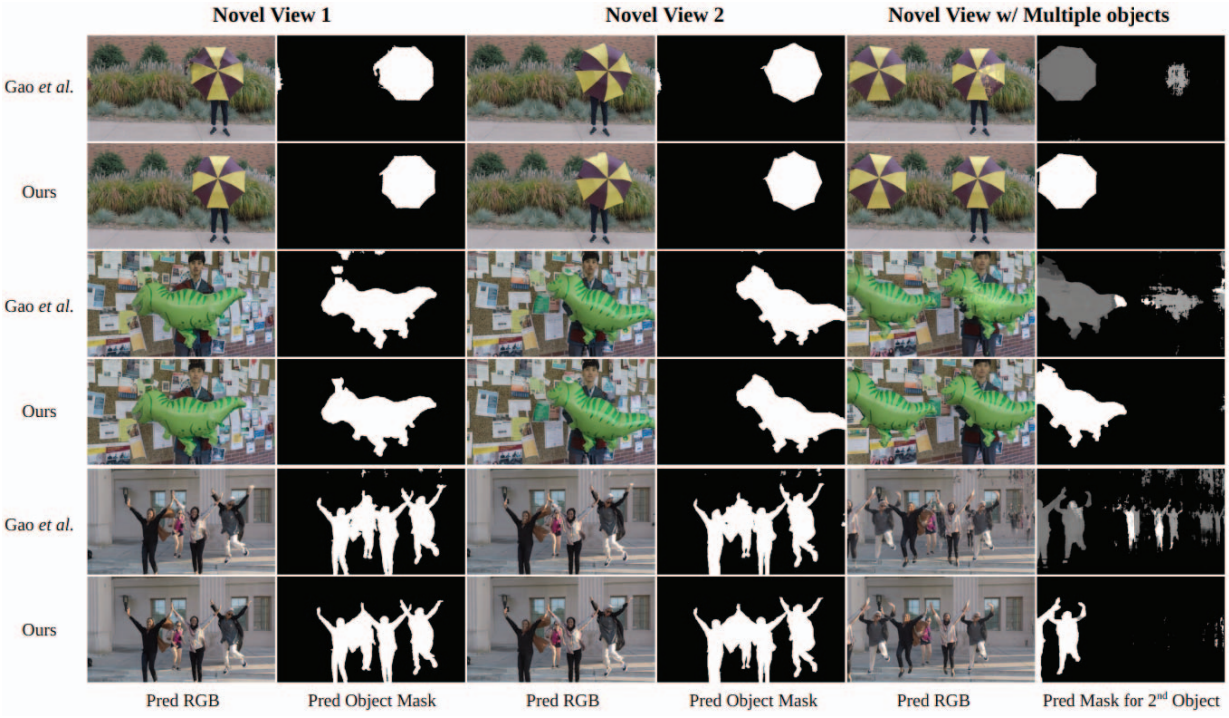

Figure 4: Qualitative results of our model on Umbrella, Balloon1, and Jumping scenes. Compared with Gao *et al*. [2], our novel-view augmentation training significantly reduces artifacts in the novel-view mask prediction, and produces images with higher fidelity, especially when composing multiple objects in a scene.

4.2.3 Ablation Study

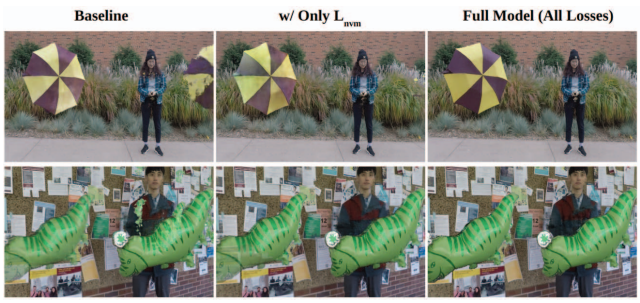

Figure 5: Ablation study on $\mathcal{L}_{nvm}$ and novel-view RGB losses.

We study the impact of each of our losses on the quality of the final image. As seen in Fig. 5, using just $\mathcal{L}_{nvm}$ can remove the blending artifacts, but RGB losses are necessary to ensure the inserted objects have proper color.

5. Conclusion

We have introduced a framework, NOVA, for the neural composition of dynamic scenes using NeRFs. Our major contributions are three modules: multiple NeRFs, novel view augmentation, and novel view losses. Using monocular dynamic video, object segmentation masks, and depth information, our results demonstrate our framework's reliability, ease, flexibility, and scalability of inserting multiple dynamic objects into a scene photo-realistically.

References

[1] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[2] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 1, 2, 3, 4

[3] DUAN GAO, Xiao Li, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images. *ACM Transactions on Graphics*, 38(4):1–15, jul 2019. 1

[4] Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. Object-centric neural scene rendering, 2022. 2

[5] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-Francois Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019. 2

[6] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. DriveGAN: Towards a controllable high-quality neural simulation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 2

[7] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 1, 2, 4

[8] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 1

[9] 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. 1, 2

[10] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[11] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2856–2865, June 2021. 2

[12] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 2

[13] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[14] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1

[15] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. *arXiv*, 2022. 2

[16] Fengrui Tian, Shaoyi Du, and Yueqi Duan. Mononerf: Learning a generalizable dynamic radiance field from monocular videos, 2022. 2

[17] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 2

[18] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*. IEEE, 1999. 2

[19] Zian Wang, David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Object instance annotation with deep extreme level set evolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2019. 2

[20] Zian Wang, Wenzheng Chen, David Acuna, Jan Kautz, and Sanja Fidler. Neural light field estimation for street scenes with differentiable virtual object insertion. In *Lecture Notes in Computer Science*, pages 380–397. Springer Nature Switzerland, 2022. 2

[21] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9421–9431. Computer Vision Foundation / IEEE, 2021. 2

[22] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *International Conference on Computer Vision (ICCV)*, October 2021. 2

[23] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5336–5345, 2020. 3, 4

[24] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. IRISformer: Dense vision transformers for single-image inverse rendering in indoor scenes. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022. 1