DEPTH MAP ESTIMATION FROM MULTI-VIEW IMAGES WITH NERF-BASED REFINEMENT

Shintaro Ito, Kanta Miura, Koichi Ito, and Takafumi Aoki

Graduate School of Information Sciences, Tohoku University,
6-6-05, Aramaki Aza Aoba, Sendai, 9808579, Japan.
shintaro@aoki.ecei.tohoku.ac.jp

ABSTRACT

In this paper, we propose a method to refine depth maps estimated by Multi-View Stereo (MVS) with Neural Radiance Field (NeRF) optimization to estimate depth maps from multi-view images with high accuracy. MVS estimates the depths on object surfaces with high accuracy, and NeRF estimates the depths at object boundaries with high accuracy. The key ideas of the proposed method are (i) to combine MVS and NeRF to utilize the advantages of both in depth map estimation, (ii) not to require any training process, therefore no training dataset and ground truth are required, and (iii) to use NeRF for depth map refinement. Through a set of experiments using the Redwood-3dscan dataset, we demonstrate the effectiveness of the proposed method compared to conventional depth map estimation methods.

***Index Terms*—** multi-view stereo, neural radiance fields, depth map estimation

1. INTRODUCTION

Multi-View Stereo (MVS) reconstructs the 3D shape of a target object from multiple images taken from different viewpoints [1]. MVS estimates a depth map for each viewpoint that represents the distance from the camera to the object using the correspondence between images and the camera parameters for each image. A 3D point cloud is then reconstructed by integrating the depth map for each viewpoint based on the camera position. Accurate estimation of the depth map is necessary to obtain a highly accurate 3D point cloud.

A major MVS method based on image matching is COLMAP [2, 3], which is an integrated method consisting of Structure from Motion (SfM) [1] for camera parameter estimation and sparse 3D reconstruction and MVS for dense 3D reconstruction. The accuracy of depth estimation is high on object surfaces, while that is low at object boundaries and in poor-texture regions. With the rapid development of deep learning, learning-based MVS has been proposed, such as CasMVSNet [4]. Depth maps are estimated for each viewpoint based on features extracted by Convolutional Neural Network (CNN) [5]. There are some problems that the camera parameters are known in advance and that a large amount of training data is required.

Recently, depth map estimation methods using Neural Radiance Fields (NeRF) [6] have been proposed. NeRF estimates the radiance field, which consists of lines connecting the camera and the object, from multiple images taken from different viewpoints and generates arbitrary viewpoint images using the estimated radiance field. Depth maps can also be estimated in the process of NeRF estimation. NeRF can estimate the depths of object boundaries with high accuracy. On the other hand, to generate high-quality arbitrary viewpoint images and estimate accurate depth maps using NeRF, several hundred images with close viewpoints and their accurate camera parameters are indispensable. Therefore, there are several methods to reduce the above limitations of NeRF by using a depth map and/or 3D point clouds as initial values [7, 8, 9, 10]. Deng et al. proposed Depth-Supervised NeRF (DS-NeRF) [10], which can generate arbitrary viewpoint images from a small number of images by training NeRF using camera parameters and sparse 3D point clouds obtained by SfM. The accuracy of the depth map obtained by DS-NeRF is not always high since the loss function is defined based on a sparse 3D point cloud. There is a method to estimate depth maps with high accuracy by combining MVS and NeRF. RC-MVSNet [11] combines CasMVSNet and NeRF to train CasMVSNet by unsupervised learning. Although unsupervised learning reduces the limitation on the number of training data, the depth map cannot always be estimated with high accuracy since NeRF is estimated based on the depth map generated by CasMVSNet.

In this paper, we propose a method to refine depth maps estimated by MVS with NeRF optimization to estimate depth maps from multi-view images with high accuracy. MVS estimates the depths on object surfaces with high accuracy, and NeRF refines them to estimate the depths at object boundaries with high accuracy. The proposed method estimates the depth map with COLMAP and refines it based on DS-NeRF, making it applicable in an unsupervised process. Through a set of experiments using the Redwood-3dscan dataset [12], we demonstrate the effectiveness of the proposed method compared to conventional depth map estimation methods.

2. PROPOSED METHOD

The proposed method consists of depth map estimation and depth map refinement as shown in Fig. 1. Depth map estimation consists of camera parameter estimation by SfM [2] and depth map estimation by MVS [3] using COLMAP. Depth map refinement refines the depth map estimated by COLMAP using a module based on DS-NeRF [10]. The key ideas of the proposed method are (i) to combine MVS and NeRF to utilize the advantages of both in depth map estimation, (ii) not to require any training process, therefore no training dataset or ground truth is required, and (iii) to use NeRF for depth map refinement. The following describes the details of each process.

2.1. Depth Map Estimation Using COLMAP

COLMAP is a 3D reconstruction pipeline consisting of SfM and MVS processes, and it is used as a standard method in MVS. COLMAP SfM [2] uses SIFT [13] to find the correspondence be-

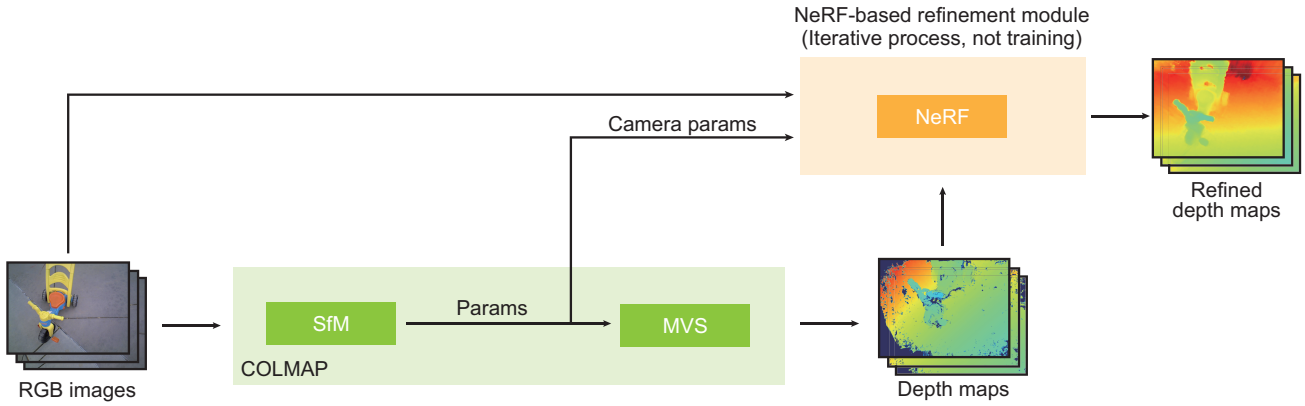

Fig. 1. Overview of the proposed method consisting of depth map estimation by COLMAP and depth map refinement by NeRF.

tween multi-view images, and then calculates the camera parameters and sparse 3D point clouds for each viewpoint based on the correspondence. COLMAP MVS [3] estimates a depth map for each viewpoint by multi-view image matching using PatchMatch [14], taking the camera parameters and sparse 3D point clouds estimated by SfM. Then, a dense 3D point cloud is obtained by integrating and optimizing the depth maps. The camera parameters and depth maps estimated by COLMAP are inputted into the subsequent refinement module to refine the depth maps.

2.2. Depth Map Refinement Using NeRF

The depth map refinement module proposed in this paper is designed based on DS-NeRF [10]. DS-NeRF takes multi-view images, camera parameters, and sparse 3D point clouds estimated by SfM as input and estimates the radiance field to generate arbitrary viewpoint images using the estimated radiance field. Since DS-NeRF is estimated based on a sparse 3D point cloud, we cannot always obtain a highly accurate depth map. The proposed refinement module improves the accuracy of the depth map by inputting the depth map estimated by COLMAP. In addition, the proposed refinement module does not require training to obtain the NeRF through iterative optimization, while DS-NeRF requires training.

Multi-Layer Perceptron (MLP) [5] is iteratively optimized to estimate the RGB values and density of the 3D points from the coordinates and line-of-sight vectors of the 3D points. The parameters for coordinates and line-of-sight vectors of each 3D point on the ray are obtained from the camera parameters estimated by SfM. Using the estimated RGB values and densities, the image's pixel values from the same viewpoint as the input image are obtained by volume rendering as described below. For a camera image I, the ray $\boldsymbol{r}_i(t)$ passing through the camera center $\boldsymbol{o}$, pixel i $(\in I)$ in the camera image, and the 3D point (x_i, y_i, z_i) is defined by

$$\boldsymbol{r}_i(t) = \boldsymbol{o} + t\boldsymbol{d}_i, \tag{1}$$

where t indicates the parameter of position on the line, and $\boldsymbol{d}_i$ is the line-of-sight vector represented by θ_i and ϕ_i. Using the RGB value $\boldsymbol{c}(\boldsymbol{r}_i(t), \boldsymbol{d}_i)$ of a 3D point on the ray and the density $\sigma(\boldsymbol{r}_i(t))$ representing opacity, the RGB value of a pixel i, $\boldsymbol{C}_i$, is reconstructed by

$$\boldsymbol{C}_i = \int_{t_{\mathrm{near}}}^{t_{\mathrm{far}}} T_i(t)\sigma(\boldsymbol{r}_i(t))\boldsymbol{c}(\boldsymbol{r}_i(t), \boldsymbol{d}_i)\, dt, \tag{2}$$

where t_{near} and t_{far} indicate the range of volume rendering. $T_i(t)$ is the accumulated transmittance, which is calculated by

$$T_i(t) = \exp\left(\int_{t_{\mathrm{near}}}^{t} \sigma(\boldsymbol{r}_i(s))\, ds\right). \tag{3}$$

The depth D_i at a pixel i can be calculated using $T_i(t)$ and $\sigma(\boldsymbol{r}_i(t))$ by

$$D_i = \int_{t_{\mathrm{near}}}^{t_{\mathrm{far}}} T_i(t)\sigma(\boldsymbol{r}_i(t))t\, dt. \tag{4}$$

The following two loss functions, i.e., the objective function of optimization, are used to optimize NeRF in the proposed refinement module. The first is the reconstruction loss L_{Color} of pixel values. L_{Color} is defined as the mean squared error between the pixel values in the camera image and those obtained by volume rendering from NeRF, which is calculated by

$$L_{\mathrm{Color}} = \sum_{j \in J} ||C_j - C_j^{\mathrm{gt}}||^2, \tag{5}$$

where J indicates a set of pixels in the input image, C_j indicates the pixel value at pixel j reconstructed by Eq. (2), C_j^{gt} indicates the pixel value of pixel j in the camera image, and $||\cdot||$ indicates the L2 norm. The second is the depth loss L_{Depth}. L_{Depth} is defined as the mean squared error between the depth obtained by COLMAP and the depth obtained by NeRF, which iscalculated by

$$L_{\mathrm{Depth}} = \sum_{k \in K} ||D_k - D_k^{\mathrm{gt}}||^2, \tag{6}$$

where K indicates the set of pixels at which the depth D_k^{gt} obtained by COLMAP is available and D_k indicates the depth at pixel k reconstructed by Eq. (4). Note that the depth loss is calculated only for rays passing through the pixel where the depth obtained by COLMAP is available. The total loss function L is calculated by

$$L = L_{\mathrm{Color}} + \lambda_d L_{\mathrm{Depth}}, \tag{7}$$

where λ_d indicates a weight parameter that balances between reconstruction and depth loss. We employ $\lambda_d = 0.1$ in this paper. MLP is iteratively optimized to minimize loss L, and the depth map with the smallest L after a given number of iterations is output as the refined depth map. In this paper, the iterations are set to 20,000.

Table 1. Accuracy of depth map estimation for each method, where each result is the mean value of 14 scenes, and only pixels with an error less than 30 cm are evaluated, except for SciMSE. In the upper rows, the accuracy is evaluated for a set of pixels where the depth exists in both the depth map estimated by COLMAP and the ground truth (GT). In the lower rows, the accuracy is evaluated for a set of pixels where the depth exists in both the depth map estimated by each method and GT.

Evaluation area	Method	SciMSE	Abs Rel	Sq Rel	RMSE (linear)	RMSE (log)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
COLMAP ∩ GT	COLMAP [3]	0.041	0.030	0.288	5.953	0.050	93.899	96.227	97.490
	DS-NeRF [10]	0.009	**0.029**	0.418	7.402	0.058	93.206	97.764	99.169
	Proposed	**0.006**	0.030	**0.266**	**5.864**	**0.043**	**96.701**	**98.652**	**99.393**
Estimated ∩ GT	CasMVSNet [4]	0.026	0.245	6.272	24.62	0.228	22.92	95.09	98.25
	Proposed	**0.006**	**0.081**	**0.899**	**10.66**	**0.084**	**95.59**	**99.44**	**99.94**

3. EXPERIMENTS AND DISCUSSION

We describe the experiments for evaluating the performance of the proposed method using the public multi-view image dataset. We use the Redwood-3dscan dataset (Redwood) [12] in the experiments. The dataset consists of video images of various objects and their 3D mesh models captured by an RGB-D camera. The video images are captured at 30 fps, and the image size is 640 × 480 pixels. This dataset contains many video images that are difficult to reconstruct because of a large number of poor texture regions and small image sizes. In this experiment, we use 14 video images as the target and 11 frames extracted from each video as the input images. We compare the accuracy of depth map estimation by COLMAP [3], DS-NeRF [10], CasMVSNet [4], and the proposed method. Note that RC-MVSNet [11] is not included in the comparison since the depth map cannot be estimated from video images in Redwood.

The accuracy of the depth map estimation is evaluated using the following accuracy evaluation metrics [15], taking the depths in millimeters provided by Redwood [12] as the ground truth. In the following, y_i denotes the depth of the pixel i in the estimated depth map, y_i^* denotes the depth of pixel i in the ground-truth depth map, and T denotes a set of pixels for evaluation.

(i) $\mathbf{SciMSE} = \frac{1}{2\|T\|}\sum_{i\in T}\left(\log\frac{y_i}{y_i^*} + \frac{1}{\|T\|}\sum_{i\in T}\log\frac{y_i^*}{y_i}\right)^2$
This is the scale-invariant evaluation metric, and the lower the value, the higher the accuracy of the estimation.

(ii) $\mathbf{AbsRel} = \frac{1}{\|T\|}\sum_{i\in T}\|y_i - y_i^*\|/y^*$
This is the mean absolute error between the estimated and ground-truth depth maps, and the lower the value, the higher the accuracy of the estimation.

(iii) $\mathbf{SqRel} = \frac{1}{\|T\|}\sum_{i\in T}\|y_i - y_i^*\|^2/y_i^*$
This is the mean squared error between the estimated and ground-truth depth maps, and the lower the value, the higher the accuracy of the estimation.

(iv) $\mathbf{RMSE\ (linear)} = \sqrt{\frac{1}{\|T\|}\sum_{i\in T}\|y_i - y_i^*\|^2}$
This is the root mean squared error between the estimated and ground-truth depth maps, and the lower the value, the higher the accuracy of the estimation.

(v) $\mathbf{RMSE\ (log)} = \sqrt{\frac{1}{\|T\|}\sum_{i\in T}\|\log y_i - \log y_i^*\|^2}$
This is the root mean squared error between the logarithm of the estimated and ground-truth depth maps, and the lower the value, the higher the accuracy of the estimation.

(vi) $\delta_i = \max_i(y_i/y_i^*, y_i^*/y_i)$
This is the ratio of the ground-truth value to the estimated value or the ratio of the estimated value to the ground-truth value among the pixels used for evaluation. The higher the ratio of pixels below the threshold, the higher the estimation accuracy.

Scales are adjusted in millimeters for the evaluation metrics other than SciMSE since the scales of the depth maps estimated by each method are different. The methods other than CasMVSNet use COLMAP in the process of depth map estimation, and therefore the scale is adjusted based on the sparse 3D point cloud obtained by COLMAP SfM. The accuracy is evaluated for a set of pixels where the depth exists in both the depth map estimated by COLMAP and the ground truth. Only pixels with an estimation error less than 30cm are used for evaluation metrics other than SciMSE, in order to eliminate the effect of outliers, which can be easily removed by filtering. In comparison to CasMVSNet, a scale is adjusted based on the depth map estimated by CasMVSNet and the ground truth. The accuracy is evaluated for a set of pixels where the depth exists in both the depth map estimated by CasMVSNet and the proposed method and that of the ground truth.

Table 1 summarizes the accuracy of the depth map estimation for each method. Compared to COLMAP and DS-NeRF, the proposed method exhibits higher accuracy in the evaluation metrics, except for AbsRel. The reason is that, in addition to the image reconstruction loss, NeRF is optimized based on the depth map estimated by COLMAP MVS to generate a highly accurate depth map in the volume rendering by NeRF. The proposed method has a low SciMSE, resulting in a smoother estimation of the depth map than the other methods. $\delta < 1.25$ indicates the percentage of pixels for which the error ratio between the estimated value and the ground truth is smaller than 1.25. Since the proposed method has higher values than the other methods, the estimated depths have fewer outliers and the depth map is refined. The proposed method exhibits higher accuracy in all the evaluation metrics compared to CasMVSNet, which is a training-based MVS. In particular, the proposed method achieves significantly higher accuracy for SqRel, RMSE (linear), and $\delta < 1.25$. These results show that the iterative optimization of NeRF can compensate for the outliers in the depth map estimated by COLMAP.

Fig. 2 shows the camera image, the depth map of the ground truth, and the depth map obtained by each method. The proposed method can smoothly estimate the depth in the regions where COLMAP cannot estimate the depth, for example, in "sculpture #06287" and "telephone #06133". The reason is that the color reconstruction loss considered in the iterative optimization of NeRF makes it possible to reconstruct the depths by volume rendering even in regions where the depths are missing. In "amp #05668" and "childseat #04134", the depth maps obtained by DS-NeRF contain blurred depths near object boundaries and noise. On the other hand, the depth map obtained by the proposed method has less noise and blurring at the object boundaries, and the boundaries between the object and the background are clear. Furthermore, in "radio #09655" and "chair #05119", the proposed method smoothly estimates both

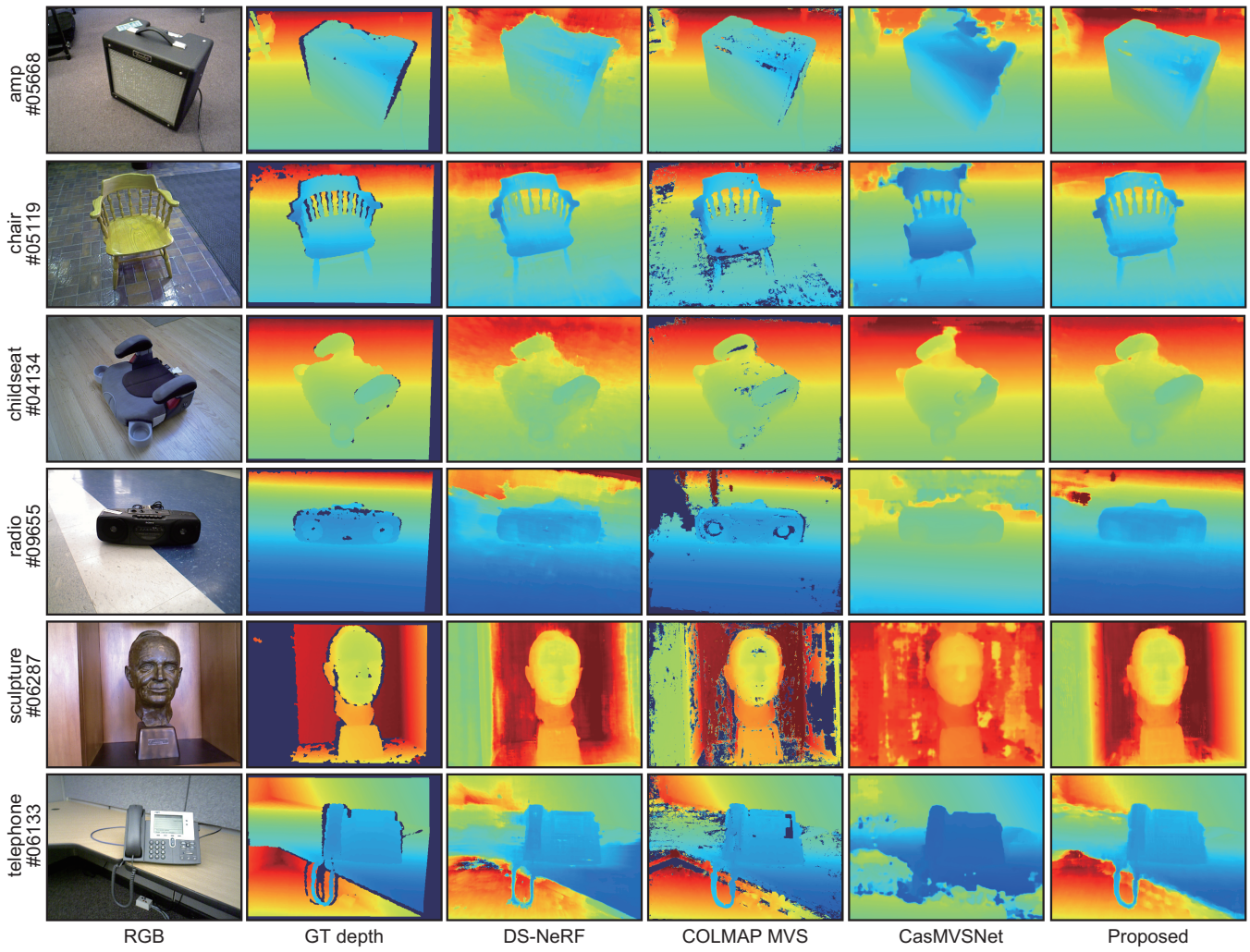

Fig. 2. Examples of the estimated depth map using each method, where depth maps are visualized as a color map with the minimum depth value of 0 and the maximum depth value of GT.

object and floor shapes, although DS-NeRF does not estimate floor shapes near the object. This is because the proposed method uses the dense depth map estimated by MVS for optimizing NeRF, while DS-NeRF uses only the depth of the sparse 3D point cloud for training, and thus can estimate the depth of the whole image with high accuracy. As described above, we confirmed that the proposed method can improve the accuracy of the depth map estimated by MVS.

4. CONCLUSION

We proposed a method to refine depth maps estimated by MVS with NeRF optimization to obtain highly accurate depth maps from multi-view images. The key idea is to combine the advantage of MVS which can estimate the depths on object surfaces with high accuracy and NeRF which can estimate the depths at object boundaries with high accuracy. Through a set of experiments using the Redwood-3dscan dataset [12], we demonstrated the effectiveness of the proposed method compared to COLMAP [3], DS-NeRF [10], and CasMVSNet [4]. In the future, we consider combining learning-based MVS with NeRF for highly accurate depth map estimation and 3D reconstruction methods.

5. REFERENCES

[1] R. Szeliski, *Computer Vision: Algorithms and Applications*, Springer-Verlag New York Inc., 2010.

[2] J. L. Schönberger and J. Frahm, "Structure-from-Motion revisited," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4104–4113, Oct. 2016.

[3] J. L. Schönberger, E. Zheng, M. Pollefeys, and J. Frahm, "Pixelwise view selection for unstructured multi-view stereo," *Proc. European Conf. Computer Vision*, pp. 501–518, 2016.

[4] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," pp. 2495–2504, June 2020.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[6] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Respresenting scenes as neural radiance fields for view synthesis," *Computer Vision – ECCV 2020*, pp. 405–421, Nov. 2020.

[7] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "PixelNeRF: Neural radiance fields from one or few images," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 4578–4587, June 2021.

[8] Y. Wei, S. Liu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo," *Proc. IEEE/CVF International Conf. Computer Vision*, pp. 5610–5619, Oct.

[9] B. Roessle, J. T. Barron, B. Mildenhall, P. P. Srinivasan, and M. Nießner, "Dense depth priors for neural radiance fields from sparse input views," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 12892–12901, June 2022.

[10] K. Deng, A. Liu, J. Y. Zhu, and D. Ramanan, "Depth-supervised NeRF: Fewer views and faster training for free," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 12882–12891, June 2022.

[11] D. Chang, A. Božič, T. Zhang, Q. Yan, Y. Chen, S. Süsstrunk, and M. Nießner, "RC-MVSNet: Unsupervised multi-view stereo with neural rendering," *Proc. European Conf. Computer Vision*, pp. 665—680, Oct. 2022.

[12] S. Choi, Q. Zhou, S. Miller, and V. Koltun, "A large dataset of object scans," *CoRR*, vol. abs/1602.02481, pp. 1–7, 2016.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[14] M. Bleyer, C. Rhemann, and C. Rother, "PatchMatch stereo-stereo matching with slanted support windows," *Proc. British Machine Vision Conference*, pp. 1–11, Aug. 2011.

[15] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Proc. Conf. Neural Information Processing Systems*, vol. 27, pp. 1–9, Dec. 2014.