Title:
Predicting Multi-View Stereo Reconstruction Quality using Neural Radiance Fields

Authors:
Satoshi Kanai, kanai-satoshi@ist.hokudai.ac.jp, Hokkaido University
Kutai Ito, sora1010bluesky@gmail.com, Hokkaido University
Hiroaki Date, hdate@ist.hokudai.ac.jp, Hokkaido University
Yasuhito Niina, ysh.niina@ajiko.co.jp, Asia Air Survey Co.,Ltd.
Ryohei Honma, ryh.honma@ajiko.co.jp, Asia Air Survey Co.,Ltd.
Kazuo Oda, kz.oda@ajiko.co.jp, Asia Air Survey Co.,Ltd.

Keywords:
Neural Radiance Fields, Multi-View Stereo, Structure-from-Motion, 3D Reconstruction, Photogrammetry

DOI: 10.14733/cadconfP.2026.140-145

Introduction:
In recent years, 3D photogrammetry techniques such as Structure from Motion (SfM) and Multi-View Stereo (MVS) have become standard tools for modeling the as-is state of medium- to large-scale objects [1]. As shown in Fig. 1(a), the reconstruction process begins with SfM, which estimates camera poses from overlapping images and extracts tie points, followed by MVS to generate a dense 3D mesh model via multi-view stereo matching.

The quality of the resulting dense mesh strongly depends on surface coverage, image quality, and shooting geometry, where surface coverage describes how completely the target object surface is captured by the images, and shooting geometry denotes the cameras' relative poses with respect to the surface. Insufficient surface coverage or improper shooting geometry can result in low-quality areas, such as holes or distortions, within the dense mesh. However, MVS requires substantial computation time, often ranging from tens of minutes to several hours, making it difficult to promptly identify low-quality areas or determine additional shooting positions after image capture. Therefore, a faster approach is needed to predict dense mesh quality without performing MVS processing.

Several studies have estimated the approximate target object surface from input images without MVS processing to estimate surface coverage and shooting geometry. For example, a method has been proposed that uses prior knowledge of the target object shape, represented as CAD or 2D map data [8];

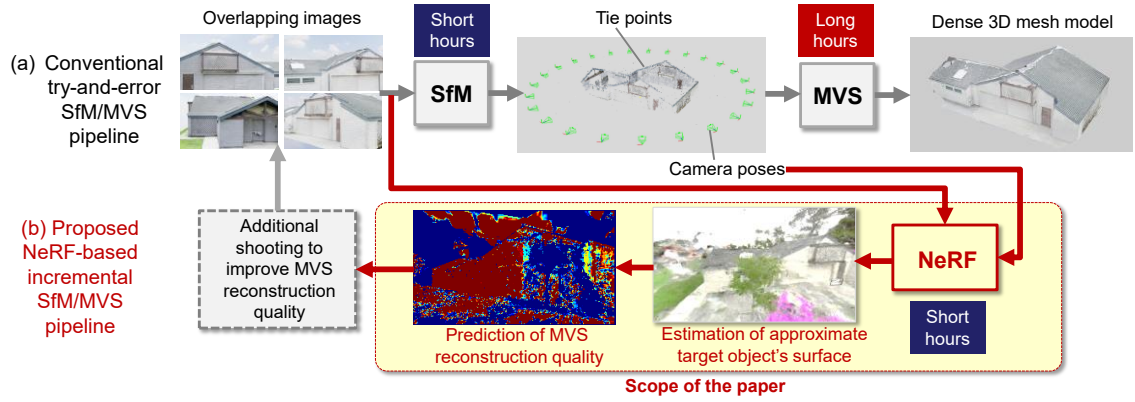

Fig. 1: NeRF-based prediction of MVS reconstruction quality.

however, challenges remain regarding the cost and accuracy of generating the knowledge. Another method has been proposed that applies Delaunay triangulation to tie points to estimate the approximate object surface [4]. However, the method risks underestimating the object's surface domain due to the sparse spatial distribution of tie points.

To address the issue, this paper uses Neural Radiance Fields (NeRF), which enable rapid image synthesis from novel viewpoints by learning radiance fields from captured images. As shown in Fig. 1(b), our objective is to develop a method that, using a NeRF, estimates the approximate target object surface more quickly than MVS and predicts surface coverage and MVS reconstruction quality based on the relative pose of the approximate target object surface with respect to the capturing camera.

Neural Radiance Fields (NeRF):

Neural Radiance Fields (NeRF) is a neural rendering and view synthesis technique that represents scenes as implicit radiance fields rather than explicit meshes and textures, enabling photorealistic novel-viewpoint image synthesis from multi-view images [3]. NeRF is trained using input images with camera poses estimated by SfM. A neural network models a 3D radiance field by taking a 3D position and viewing direction as input and outputting the emitted color and volume density. The images from novel viewpoints are generated by sampling color and density along the rays passing pixels and computing the pixel values via volume rendering of the field.

NeRF reconstructs scenes using a different principle from MVS and offers several advantages. In particular, the processing time from learning to novel-view synthesis is significantly faster than MVS, and dense surface coverage with the input images is not required. As a result, NeRF can often reconstruct scenes from sparse image sets, whereas MVS often fails to do so. In addition, depth images can be rendered from arbitrary viewpoints by estimating expected depth values from the volumetric density distribution, which provides an approximate surface representation of the target object. Our proposed method exploits this depth estimation capability of NeRF.

Prediction Pipeline of MVS Reconstruction Quality utilizing NeRF:

The proposed method takes as input a set of captured images $\mathcal{I}^{input} = \{I_k^{input} \mid k \in [1, K^{input}]\}$ and an evaluation camera c_q placed at an arbitrary viewpoint $\boldsymbol{q}$ specified by the user, where K^{input} denotes the number of captured images. It then renders an RGB image $I^{NeRF}(c_q)$ and depth image $D^{NeRF}(c_q)$ when observed from the camera c_q using NeRF and predicts reconstruction quality score image $Q(c_q)$ of the dense mesh M^{dense} that will be reconstructed from the input images $\mathcal{I}^{input}$ and observed from c_q. The detailed processing steps are presented in the followings.

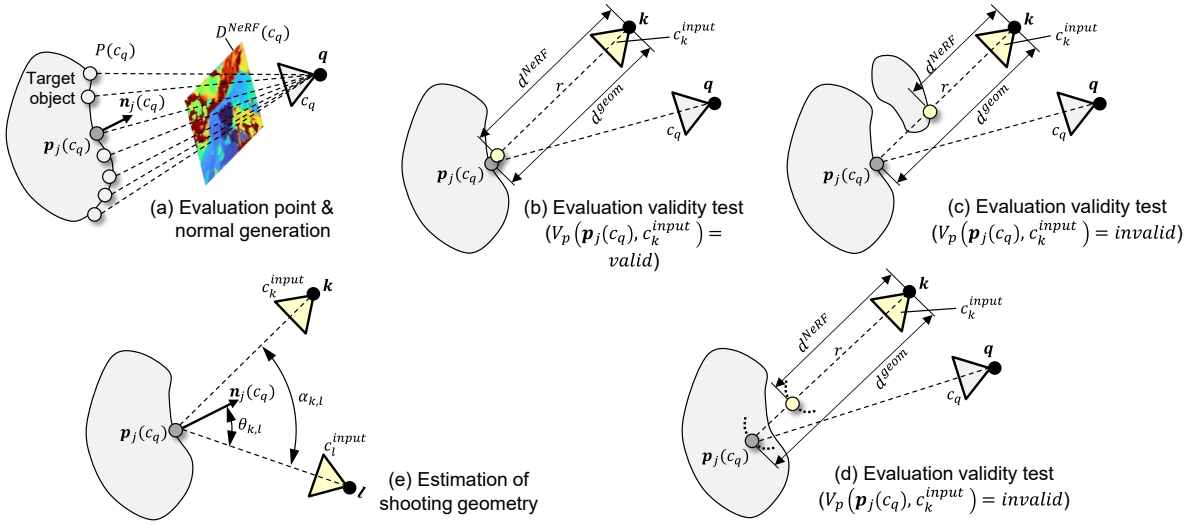

Fig. 2: Principles of MVS reconstruction quality prediction using NeRF.

Camera Pose Estimation and NeRF Model Training
The standard SfM process using COLMAP [5] is applied to the captured image set $\mathcal{I}^{input}$ to estimate the poses of the camera c_k^{input} for each image $I_k^{input} \in \mathcal{I}^{input}$. Then a NeRF model is trained using the image set $\mathcal{I}^{input} = \{I_k^{input}\}$ and camera set $C^{input} = \{c_k^{input}\}$.

Evaluation Point Generation
As shown in Fig. 2(a), a depth image $D^{NeRF}(c_q)$ is rendered for an evaluation camera c_q placed at a user-specified viewpoint $\boldsymbol{q}$ using the radiance intensity of NeRF model. Then the inverse projection of the depth image $D^{NeRF}(c_q)$ is performed to generate a set of 3D evaluation points $P(c_q) = \{\boldsymbol{p}_j(c_q) | j \in J^{NeRF}\}$ corresponding to a pixel j in $D^{NeRF}(c_q)$, where J^{NeRF} denotes a set of pixels in D^{NeRF}. Furthermore, a set of the unit normal vector of the object surface facing toward c_q is estimated at j as $N(c_q) = \{\boldsymbol{n}_j(c_q) | j \in J^{NeRF}\}$ from the depth difference between adjacent pixels in $D^{NeRF}(c_q)$. The evaluation points $P(c_q)$ and their normal vectors $N(c_q)$ work as a discrete approximation of a target object surface.

Evaluation Validity Test
Based on the evaluation points $P(c_q)$, their normal vectors $N(c_q)$ and the NeRF model, we test whether each evaluation point $\boldsymbol{p}_j(c_q)$ is suitable for assessing the quality of the dense mesh that would be reconstructed from the image set $\mathcal{I}^{input}$. It examines whether the following two conditions hold:
(1) Is the evaluation point $\boldsymbol{p}_j(c_q)$, reconstructed from the input image set $\mathcal{I}^{input}$ and NeRF, reliable?
(2) Is the evaluation point $\boldsymbol{p}_j(c_q)$ is visible from the camera c_k^{input}?
Condition (1) holds when the number and orientation of input images are sufficient for training the NeRF model, and a consistent radiance field is generated around the point $\boldsymbol{p}_j(c_q)$. On the other hand, condition (2) holds when no other object exists between the input camera c_k^{input} and the point $\boldsymbol{p}_j(c_q)$.

To test these conditions, as shown in Fig. 2(b)-(d), we first generate a ray $r\left(\boldsymbol{k}, \boldsymbol{p}_j(c_q)\right)$ from the viewpoint $\boldsymbol{k}$ of the camera $c_k^{input} \in C^{input}$ toward the evaluation point $\boldsymbol{p}_j(c_q)$ within the field of view of c_k^{input}, then estimate the depth $d^{NeRF}\left(r\left(\boldsymbol{k}, \boldsymbol{p}_j(c_q)\right)\right)$ from the NeRF model along that ray $r\left(\boldsymbol{k}, \boldsymbol{p}_j(c_q)\right)$. On the other hand, we calculate the simple geometric distance $d^{geom}\left(\boldsymbol{k}, \boldsymbol{p}_j(c_q)\right)$ between the viewpoint $\boldsymbol{k}$ and $\boldsymbol{p}_j(c_q)$. Finally, if the relative error of the depth d^{NeRF} with respect to the distance d^{geom} does not exceed a predetermined threshold (currently 10%) as shown in Fig. 2(b), it is determined that conditions (1) and (2) are satisfied, the evaluation validity V_p is set to $V_p(\boldsymbol{p}_j(c_q), c_k^{input}) = valid$ for the pair of evaluation point $\boldsymbol{p}_j(c_q)$ and input camera c_k^{input}. If the relative error exceeds the threshold as shown in Fig. 2(c)-(d), conditions (1) or (2) do not hold, and we set $V_p(\boldsymbol{p}_j(c_q), c_k^{input}) = invalid$.

Estimations of Shooting Geometry and Reconstruction Quality Score
Finally, MVS reconstruction quality score $S\left(\boldsymbol{p}_j(c_q)\right)$ for each evaluation point $\boldsymbol{p}_j(c_q)$ is estimated from the shooting geometry between $\boldsymbol{p}_j(c_q)$ and the input camera set $C^{input} = \{c_k^{input}\}$. A previous study [5] empirically demonstrated that MVS reconstruction quality is influenced by the triangulation angle, incident angle, and the distance from the camera to the point. As shown in Fig. 2(e), the triangulation angle $\alpha_{k,l}$ is the angle between the viewing rays from a camera pair c_k^{input} and c_l^{input} to the same point $\boldsymbol{p}_j$, while the incident angle $\theta_{k,l}$ is the shallower of the ones between the surface normal and two viewing rays.

To estimate reconstruction quality from $\alpha_{k,l}$ and $\theta_{k,l}$, for an evaluation point $\boldsymbol{p}_j$ and camera pair $(c_k^{input}, c_l^{input})$, we calculate $w_1(\boldsymbol{p}_j)$, $w_2(\boldsymbol{p}_j)$ and $w_3(\boldsymbol{p}_j)$ using the following Eqns. (1) to (3):

$$w_1(\boldsymbol{p}_j) = \left\{1 + \exp\left(-k_1(\alpha_{k,l} - \alpha_1)\right)\right\}^{-1} \quad (1)$$

$$w_2(\boldsymbol{p}_j) = 1 - \left\{1 + \exp\left(-k_2(\alpha_{k,l} - \alpha_2)\right)\right\}^{-1} \quad (2)$$

$$w_3(\boldsymbol{p}_j) = \begin{cases} \cos\theta_{k,l} & (\cos\theta_{k,l} \geq 0.5) \\ 0 & (\cos\theta_{k,l} < 0.5) \end{cases} \quad (3)$$

Eqns. (1) and (2) imply that penalties are applied when the triangulation angle is either too small or too large, while Eqn. (3) implies that a large penalty is imposed if the incident angle is too large. As the parameters k_1, k_2, α_1 and α_2, we used $\alpha_1 = \pi/16$, $\alpha_2 = \pi/4$, $k_1 = 32$, $k_2 = 8$ in accordance with the recommendation in [6]. Then the pair reconstruction quality score $s_p(\boldsymbol{p}_j, (c_k^{input}, c_l^{input}))$ for an evaluation point $\boldsymbol{p}_j(c_q)$ w.r.t. a camera pair $(c_k^{input}, c_l^{input})$ is evaluated using Eqn. (4).

$$s_p(\boldsymbol{p}_j, (c_k^{input}, c_l^{input})) = w_1(\boldsymbol{p}_j) w_2(\boldsymbol{p}_j) w_3(\boldsymbol{p}_j) \quad (4)$$

Considering that the reconstruction quality does not improve significantly even with three or more excess camera pairs in MVS, finally, the reconstruction quality score $S\left(\boldsymbol{p}_j(c_q)\right)$ for an evaluation point $\boldsymbol{p}_j(c_q)$ is obtained from Eqn.(5).

$$S\left(\boldsymbol{p}_j(c_q)\right) = max\left[\ \sum_{(c_k,c_l)\in C_{valid}} s_p(\boldsymbol{p}_j(c_q), (c_k, c_l)),\ \ 3\right] \quad (5)$$
$$C_{valid} = \left\{\ (c_k, c_l)\ \middle|\ c_k, c_l \in C^{input}, c_k \neq c_l,\ V_p(\boldsymbol{p}_j(c_q), c_k) = valid,\ \ V_p(\boldsymbol{p}_j(c_q), c_l) = valid\ \right\}$$

By evaluating the reconstruction quality score $S\left(\boldsymbol{p}_j(c_q)\right)$ for each pixel in the image rendered from the camera c_q, the reconstruction quality score image $Q(c_q)$ can be rendered as an image format.

Experiments:

Setting

We conducted experiments using *Barn* and *Truck* image sets ($|\mathcal{I}^{all}| = 410$ for *Barn*, and 230 for *Truck*) from a publicly available image benchmark [2]. COLMAP [5] was used for SfM and MVS reconstruction, while a custom implementation based on nerfacto [7] was used for NeRF training, rendering, and reconstruction quality prediction. The experiment aimed to verify whether the proposed method can predict the degradation of a dense mesh reconstructed from MVS with the reduced input images.

Fig. 3 shows the experimental flow. First, we created an image subset $\mathcal{I}^{sub}$ by extracting small number of images at biased positions from the complete image set $\mathcal{I}^{all}$. We then estimated camera poses for the images of $\mathcal{I}^{sub}$ ($|\mathcal{I}^{sub}| = 131$ for *Barn*, and 83 for *Truck*) and for $\mathcal{I}^{all}$ using SfM. Then, the ground truth dense mesh M_{GT} was reconstructed from $\mathcal{I}^{all}$ using MVS, and the dense mesh for quality evaluation M_{mvs} was reconstructed from $\mathcal{I}^{sub}$. NeRF model used for the quality prediction was trained using $\mathcal{I}^{sub}$.

Evaluation Method

To evaluate the validity of the proposed quality score image $Q(c_q)$, we generated an MVS reconstruction error map $R(c_q)$ visualizing the error between M_{mvs} and M_{GT} and compared the maps $Q(c_q)$ and $R(c_q)$. To this end, first, the minimum distance $e(v)$ from a vertex v on M_{mvs} to the dense points on M_{GT} was

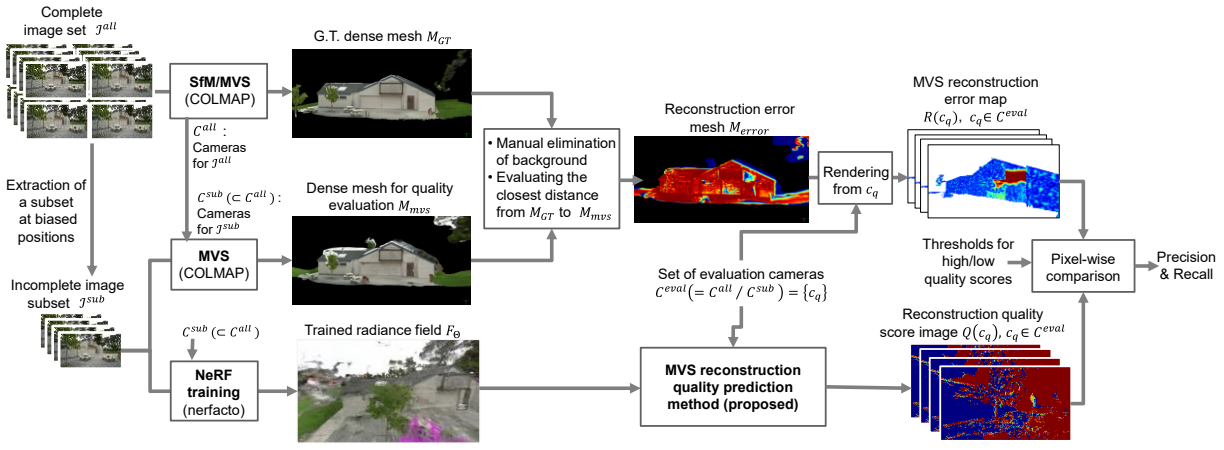

Fig. 3: Flow of the experiments.

SfM	*NeRF*	*Reconstruction quality prediction*	*(MVS)*
11 min	22 min	1.3 sec	57 min

Tab. 1: Difference in processing time between the proposed method and MVS.

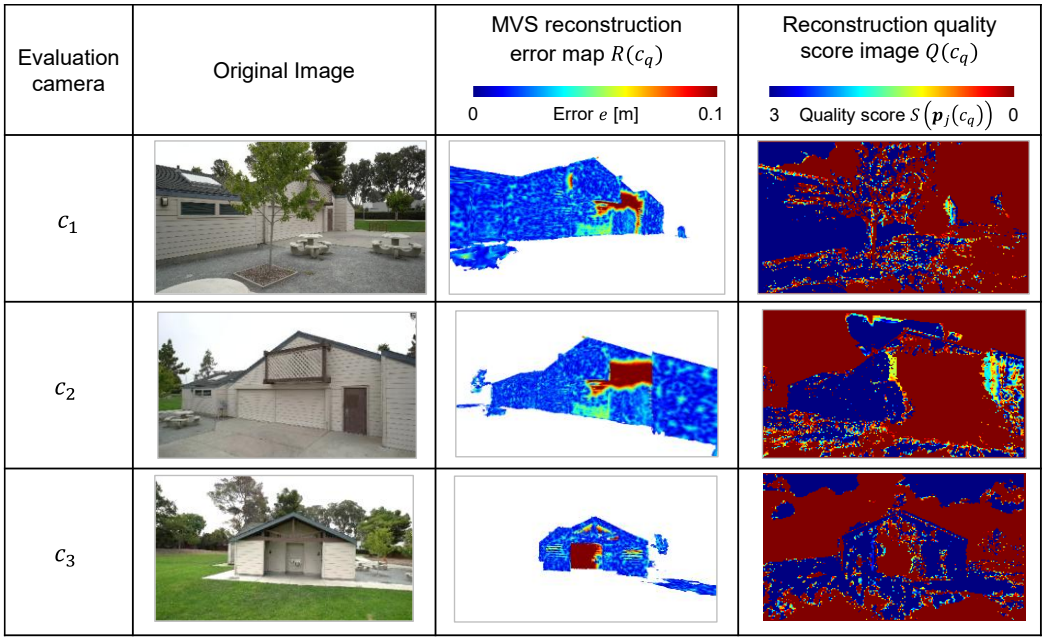

Fig. 4: MVS reconstruction error map $R(cq)$, and reconstruction quality score image $Q(cq)$at three evaluation cameras for Barn dataset.

	Barn		*Truck*	
	Low-quality regions	*High-quality regions*	*Low-quality regions*	*High-quality regions*
Recall	*72.4%*	*85.3%*	*76.4%*	*70.4%*
Precision	16.6%	53.4%	20.9%	91.6%

Tab. 2: Binary classification performances of the reconstruction quality score using pixel-wise precision and recall evaluated from all cameras C^{eval}.

evaluated and treated as the error between the mesh M_{mvs} and M_{GT}. Next, we generated a reconstruction error mesh M_{error} by assigning the error $e(v)$ as a color to the attribute of each vertex v in M_{mvs}. Finally, by rendering M_{error} from the same camera c_q as the map $Q(c_q)$, the MVS reconstruction error map $R(c_q)$ is obtained.

Results and Discussion

Table 1 shows the execution times for each processing step in the proposed prediction method and MVS. It was confirmed that the proposed prediction method performed approximately 2.6 times faster than MVS processing. In this case, NeRF training was set to 20,000 epochs, but the model still performs well with even fewer epochs.

Fig. 4 compares the reconstruction quality score image $Q(c_q)$ and MVS reconstruction error map $R(c_q)$ from the evaluation cameras c_1, c_2 and c_3 in *Barn* dataset. For all cameras, the areas with missing portions or large reconstruction errors (red) in the error map $R(c_q)$ could be predicted as low-quality regions (red) on the score image $Q(c_q)$. This suggests the proposed prediction method was effective.

Furthermore, for the quantitative verification, we performed a binary classification of the pixels both in the error map $R(c_q)$ and in the quality score image $Q(c_q)$. We classified pixels of the image $Q(c_q)$ in the top 10% as high-quality, and those in the bottom 10% as low-quality. In contrast, for the error map $R(c_q)$, we reversed this order. Table 2 presents the pixel-wise binary classification results for all evaluation cameras C^{eval} ($|C^{eval}| = 279$ for *Barn*, and 147 for *Truck*). For these datasets, recall for both low- and high-quality regions, as determined by the reconstruction quality score, was relatively high, ranging from 70% to 85%. Furthermore, precision for high-quality regions was also high, reaching 91% on *Truck* dataset. This indicates that the proposed score can predict high-quality regions on the MVS-reconstructed mesh with relatively high accuracy and that it also misses relatively few low-quality regions. On the other hand, precision for low-quality regions was low, ranging from 16% to 20%. This implies that the proposed score over-detects regions on the MVS-reconstructed mesh as "low-quality" even when they are not actually low-quality.

Conclusions:

This paper proposed a method to predict MVS reconstruction quality in advance by leveraging the relationship between the NeRF-learned depth map of the radiance field from the input image and the camera configuration. Its effectiveness was verified through experiments. The results qualitatively demonstrated that the proposed method could predict reconstruction quality, especially in the high-quality region, with acceptable accuracy, more rapidly than MVS.

Moving forward, we intend to address the over-detection issue for low-quality regions by incorporating the richness of the object's surface texture into the evaluation metrics, along with the geometric relationship between the camera and the surface.

ORCID:
Satoshi Kanai, https://orcid.org/0000-0003-3570-1782
Kutai Ito, https://orcid.org/0009-0005-1014-5558
Hiroaki Date, https://orcid.org/0000-0002-6189-2044
Yasuhito Niina, https://orcid.org/0009-0008-5901-1229
Ryohei Honma, https://orcid.org/0009-0008-2246-4849
Kazuo Oda, https://orcid.org/0009-0001-8711-0135

References:

[1] Furukawa, Y.; Hernández, C.: Multi-View Stereo: A Tutorial, Foundations and Trends in Computer Graphics and Vision, 9(1-2), 2015, 1–148. https://doi.org/10.1561/0600000052

[2] Knapitsch, A.; Park, J.; Zhou, Q. Y.; Koltun, V.: Tanks and temples: benchmarking large-scale scene reconstruction, ACM Trans. Graph., 36(4), 2017, 78. https://doi.org/10.1145/3072959.3073599

[3] Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; Ng., R.: NeRF: representing scenes as neural radiance fields for view synthesis, Commun. ACM, 65(1), 2022, 99–106. https://doi.org/10.1145/3503250

[4] Moritani, R.; Kanai, S.; Date, H.; Niina, Y.; Honma, R.: Plausible reconstruction of an approximated mesh for next-best view planning of SfM-MVS, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLIII-B2-2020, 2020, 465–471. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-465-2020

[5] Schonberger, J. L.; Jan-Frahm, J. M.; Structure-from-Motion revisited, IEEE Conference on Computer Vision and Pattern Recognition, 2016, 4104-4113. https://doi.org/10.1109/CVPR.2016.445.

[6] Smith, N.; Moehrle, N.; Goesele, M.; Heidrich, W.: Aerial path planning for urban scene reconstruction: a continuous optimization method and benchmark, ACM Trans. Graph., 37(6), 2018, 183. https://doi.org/10.1145/3272127.3275010

[7] Tancik, M.; Weber, E.; Ng, E.; Li, R.; Yi, B.; Wang, T.; Kristoffersen, A.; Austin, J.; Salahi, K.; Ahuja, A.; Mcallister, D.; Kerr, J.; Kanazawa, A.: Nerfstudio: A Modular Framework for Neural Radiance Field Development, ACM SIGGRAPH 2023, 72, 2023, 1–12. https://doi.org/10.1145/3588432.3591516

[8] Yamazaki, K.; Okahara, K.; Minezawa, A.: View Planning using Geospatial Information for 3d Reconstruction with Unmanned Aerial Vehicles, IEEE International Geoscience and Remote Sensing Symposium, 2023, 4760-4763. https://doi.org/10.1109/IGARSS52108.2023.10282238.