Coloring the Past: Neural Historical Monuments Reconstruction from Archival Photography

Dávid Komorowicz*[1,3], Lu Sang*[1,4],
Ferdinand Maiwald[2], and Daniel Cremers[1,4]

[1] Technical University of Munich
[2] Dresden University of Technology
[3] Friedrich Schiller University Jena
[4] Munich Center for Machine Learning
{david.komorowicz,lu.sang,cremers}@tum.de
ferdinand.maiwald@tu-dresden.de

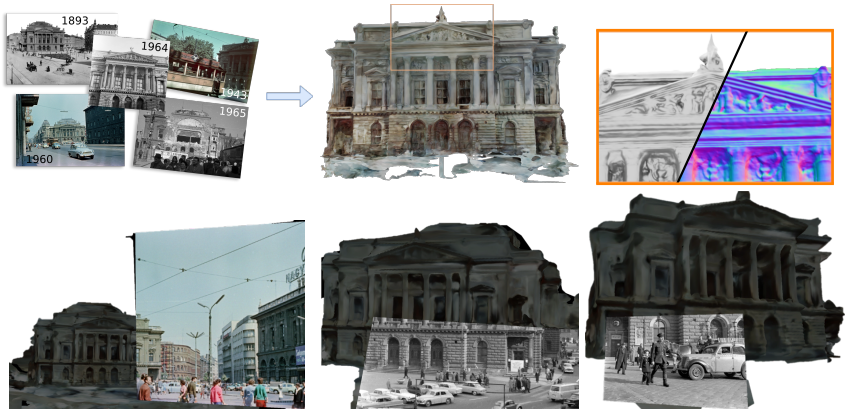


Fig. 1: The first row shows photographs of the Hungarian National Theater over a long period (left), reconstructed mesh with color (middle), shaded mesh, and normal map (right). The second row displays the aligned mesh alongside historical images. Our method to restore historical monuments has meaningful applications.

Abstract. Historical monuments are a treasure and milestone of cultural heritage. Reconstructing the 3D models of these buildings holds significant value. The rapid development of neural rendering methods makes it possible to recover the original 3D shape exclusively based on archival photographs. However, this task presents considerable challenges due to the properties of available color images. Historical pictures are often limited in number and the scenes in these photos might have altered over time. The radiometric quality of these images is often sub-optimal for using automatic methods. To address these challenges, we introduce an approach to reconstruct the geometry of historical buildings from limited input images. We leverage dense point clouds as a geometric prior and introduce a color appearance embedding loss in volumetric rendering to recover the color of the building. We aim for our work to spark increased interest and focus on preserving historic buildings. Together with the proposed method, we introduce a new historical dataset of the Hungarian National Theater, providing a new benchmark for 3D reconstruction.

Keywords: Neural rendering · Historical monuments · 3D reconstruction.

1 Introduction

Historical monuments are unique cultural heritage features and are seen as landmarks connecting people across time and countries. They make history tangible and provide insight into the past. However, they are vulnerable to temporal and man-made changes. It is one of the main goals of UNESCO[5] to protect and preserve our cultural heritage which becomes possible because of the rapid development of 3D technologies [12] such as 3D reconstruction from 2D images [13, 16, 18]. However, these methods depend on the availability of numerous high-quality color images with precise calibration information from the cameras used to capture them. On the contrary, historical images have their own characteristics that hinder the previous 3D reconstruction techniques. For example, images of historical monuments, especially those of buildings that no longer exist, are limited and often exclusively available in gray-scale. Many historical sites are documented solely through antiquated photographs, captured with obsolete equipment that took images with difficult radiometric properties such as blurriness, lack of color, or absence of accurate camera parameters [2, 8, 10]. Furthermore, the appearance of historical monuments may change over time, making it difficult to find correspondences across multi-temporal images. Identifying correspondences is crucial for most 3D reconstruction techniques, as camera poses and 3D geometry features depend on them.

In this paper, we propose a method for 3D reconstruction and color view synthesis of historical buildings using sparse and low-quality input images. We employ a dense point cloud recovered with the Structure from Motion (SfM) and Multi-view Stereo (MVS) algorithms as a geometric prior and introduce color appearance embedding to utilize the minority of available color images for recovering the 3D model's color. As shown in Fig. 1, our method can reconstruct a colored 3D mesh of a building from a historical image collection predominantly composed of gray-scale images taken over different periods. We aim to spark interest in historic monument reconstruction and the use of historical photographs within the 3D reconstruction community.

In summary, our contributions are as follows:

- We propose a method that can reconstruct satisfactory 3D geometry of historical buildings by leveraging sparse and low-quality images.
- We propose a color appearance embedding loss to obtain a color synthetic view when the majority of photos are gray-scale.
- We introduce a historical dataset that showcases a wide range of properties typically present in historical datasets.

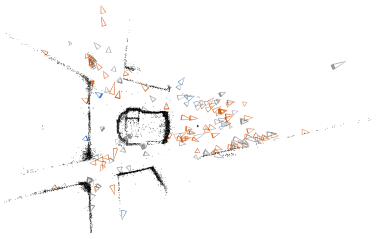
[5] State of the Ocean Report: understand, educate, protect (`https://www.unesco.org/`)

2 New Historical Dataset

Reconstructing historical buildings based on archival photography provides significant value, not only in the research area but also in the protection and preservation of cultural heritage. However, historical images of the same building are often scattered in multiple archives with often unresolved copyrights and only a few historical datasets are available for research purposes. Here, we introduce a new historical image dataset: the Hungarian National Theater dataset.

This dataset includes 229 images of the Hungarian National Theater directly released by us and another 136 images for which the access link is provided. Additionally, we provide a dense point cloud and camera poses which are estimated using Structure from Motion (SfM). All photos were taken between 1875-1965. During this period, the availability of color photography was limited. Thus, different from the modern building image datasets, the vast majority of photos are gray-scale (over 90%) and only a small portion is available in color (Fig. 2b). Another significant difference is, that the building can slightly change over decades. We provide the capture dates of the images to allow selections according to the years. Besides its cultural significance to the Hungarian people, this historical dataset is a rare case of having a complete photo collection covering the whole area around an old building that is no longer present. All four sides appear in different numbers of images in the dataset.

This makes the dataset suitable as a benchmark to evaluate the algorithms' performance regarding the number and quality of the input images. Fig. 2a shows the reconstructed point cloud and estimated camera locations using SfM [11] from the National Theater dataset. Fig. 3 shows example images of the facade across time. Fig. 2b summarizes all released historical datasets so far from [9] and our new released dataset (first row). The first column shows the total number of images in each dataset, the second column indicates the number of color images, and the third column lists the number of images with sufficient quality for training (for the proposed method). From Fig. 2b, it is evident that our dataset includes a greater number of images and a higher proportion of color images. We hope this newly released dataset enriches the historical reconstruction benchmark and encourages further research in this area.



(a)

Dataset name	Total	Color	Train
National Theater (Ours)	229	16	153
Hotel International [9]	19	1	18
Observatory [9]	37	3	33

(b)

Fig. 2: (a) down view of a reconstructed point cloud of the National Theater dataset. blue cameras stand for validation views, orange cameras are training images, gray cameras are images that can be obtained via request[6]. (b) statistics of the National Theater dataset, three historical datasets from Jena dataset [9]

1875 1959 1964

Fig. 3: Example images[7] from the **Hungarian National Theater** dataset.

3 Method

The whole pipeline of our method is as follows. Given a set of images $\{\mathcal{I}_i\}$, for $i \in \{0, 1, \ldots, n\}$, we first preprocess the input images, i.e., resize the images to the same size, since the historical datasets typically contain images with varying resolution. Then, the corresponding extrinsic (poses) and intrinsic camera parameters are estimated using SfM [11]. We run a segmentation method, similar to [13] to mask out irrelevant objects such as people and cars. We generate two kinds of point clouds, a **sparse** point cloud, directly using SfM [11] and a **dense** point cloud $\mathcal{P} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ using the estimated camera parameters by multiview stereo fusion [11]. We use the dense point cloud as geometry prior together with images to train an SDF-based differential rendering network with color appearance embedding loss to estimate the geometry. The dense point cloud prior helps to recover a better quality mesh when the input image number is limited and enable rendering views from unseen camera poses.

3.1 Backbones and Geometry Loss

We build our method on top of NeusW [13] and our network architecture consists of two parts, an SDF net and a color prediction net. The SDF net estimates the signed distance value $d \in \mathbb{R}$ and a geometric feature $\mathbf{f} \in \mathbb{R}^{f_n}$, for f_n is the dimension of the feature vector. Given point $\mathbf{x} \in \mathbb{R}^3$, the color prediction net outputs the rendered color $\mathbf{c}$. In detail, given points $\mathbf{x}$, viewing direction $\mathbf{v} \in \mathbb{S}^2$, we compute normal $\mathbf{n} = \nabla \mathbf{MLP}_{\mathrm{SDF}}(\mathbf{x})$, and a feature vector $\mathbf{f} \in \mathbb{R}^{f_n}$ with dimension f_n.

$$(d, \mathbf{f}) = \mathbf{MLP}_{\mathrm{SDF}}(\mathbf{x}), \tag{1}$$

$$\mathbf{c}_i = \mathbf{MLP}_{\mathrm{COLOR}}(\mathbf{x}, \mathbf{v}, \mathbf{n}, \mathbf{e}_i, \mathbf{f}). \tag{2}$$

where $\mathbf{e}_i$ are appearance embeddings corresponding to each input photo, optimized alongside the parameters of MLPs, see [13] for more details. We first

[6] These images can be accessed upon purchase. We will provide the link to these images.

[7] Fortepan by UVATERV/FÖMTERV/Zsolt Pálinkás/Pál Breuer/Lajos Miklós and Budapest City Archives: HU.BFL.XV.19.d.1.05.103/HU.BFL.XV.19.d.1.07.020 under CC-BY-SA-3.0

initialize a voxel grid by the sparse point cloud similar to [13]. For image $\mathcal{I}_i$ with camera center $\mathbf{o}$, we shoot a ray from its pixels. The ray $\mathbf{r}$ with direction $\mathbf{v}$ is $\{\mathbf{r}(s) = \mathbf{o} + \mathbf{v}s | s \geq 0\}$. We pass the points along the ray to the SDF net to get the geometry feature $\mathbf{f}$ and then pass these points to the color net to get the color estimation. We reuse the SDF net for geometry loss as well. For image $\mathcal{I}_i$ where we sampled ray from, we find all points from the dense point cloud $\mathcal{P}$, which are visible from this image, denoted as $\mathcal{P}_i$. The geometry loss [4] is

$$l_g(\mathbf{x}) = \lambda \frac{1}{|\mathcal{P}_i|} \sum_{\mathbf{x} \in \mathcal{P}_i} |\mathbf{MLP}_{\mathrm{SDF}}(\mathbf{x})| \,, \tag{3}$$

where $|\mathcal{P}_i|$ is the number of points in the point cloud and λ is a learnable parameter. During training, we sample rays across multiple images for one batch and randomly choose one image to compute the geometry loss for the point cloud visible from that image. The geometry loss ensures that the SDF net is guided by the **dense point cloud**.

We use a dense point cloud instead of the sparse point cloud because we believe the dense point cloud provides complementary information, see Fig. 4. Directly sampling at the dense point cloud points to optimize the SDF net allows us to bypass the ray marching procedure. In NeusW [13] and our case, the sampling is directly dependent on the SDF values. Good geometry prior, *i.e.* dense point cloud will benefit SDF estimation first, and the improved SDF will improve sampling again.

3.2 Color Appearance Embedding

To deal with the situation that most of the input images are available as gray-scale, and only a small portion provides color channels, we propose a color appearance embedding loss to recover color output. Previous methods treat gray-scale images as color by setting the three channels to equal values. This results in a less-than-ideal appearance embedding and a gray-scale output. The rendered color for a ray $\mathbf{r}$ is

$$\boldsymbol{C}'(\mathbf{r}) = \int_0^{+\infty} w(t) c(\mathbf{r}(t), \mathbf{v}, \mathbf{f}) \mathrm{d}t \,, \tag{4}$$

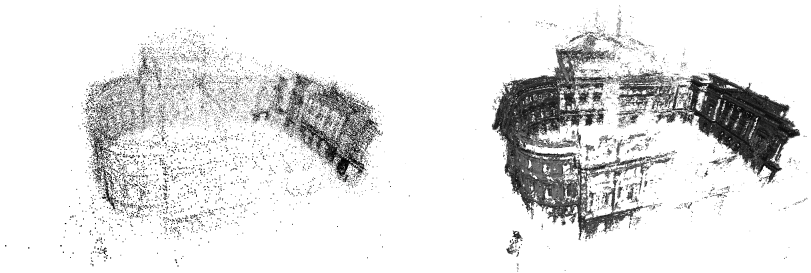


Fig. 4: Comparison of sparse (left) and dense (right) point cloud generated by stereo fusion [11].

where $w(t)$ is an unbiased and occlusion-aware weight function used in [15]. The color net outputs a three-channel color vector, to supervise it using gray-scale images, we use perceptual weights [17] to convert the output color to gray-scale value, *i.e.*, for $\boldsymbol{C}'(\mathbf{r}) = (c_r, c_g, c_b)$, we propose the function $g : \mathbb{R}^3 \to \mathbb{R}$ and

$$g(\boldsymbol{C}'(\mathbf{r})) = w_r c_r + w_g c_g + w_b c_b \,, \tag{5}$$

where $W_r = 0.2126$, $W_g = 0.7152$ and $W_b = 0.0722$. The loss for ray color $\boldsymbol{C}'(\mathbf{r})$ in image with true color $\boldsymbol{C}(\mathbf{r})$ is

$$l_c(\mathbf{r}) = \begin{cases} \frac{1}{2}|\boldsymbol{C}(\mathbf{r}) - g(\boldsymbol{C}'(\mathbf{r}))|^2, & \text{if } \mathbf{r} \text{ is gray-scale}\,, \\ \frac{1}{2}|\boldsymbol{C}(\mathbf{r}) - \boldsymbol{C}'(\mathbf{r})|^2, & \text{otherwise}\,. \end{cases} \tag{6}$$

With the color appearance embedding loss, we weakly supervise on gray-scale images and strongly on color images.

4 Experiments

In this section, we present the experimental results of the proposed method on 3 historical scenes, including the proposed theater dataset, listed in Fig. 2b.

Implementation details For our method and its comparison to other methods we build upon the implementation of SDFStudio [19]. We use 8 layers with 512 hidden units for the geometry MLP and 4 layers with 256 hidden units for the color MLP. For dense point cloud generation and to obtain the camera poses, we use COLMAP [11] and feature matches obtained by a combination of state-of-the-art keypoint detector and feature matching algorithms [3, 6, 14]. For Metashape [1], we import the feature matches using the bundler format and apply the segmentation masks.

Training details We select a sampling radius V_{sfm} roughly 2 times the radius of the encapsulating sphere of the main object or building of interest. For the geometry loss Eq. (3) we set $\lambda = 0.1$. The voxel grid used for accelerated sampling is updated every 5k iterations. We sample the color network Eq. (2) at the vertices and save it as vertex color. During inference, we use the average appearance embedding vector for the color network. We run all experiments on 4 NVIDIA A100 GPUs for 100.000 iterations with a batch size of 2048 per GPU. For the final output mesh, we only extract a mesh within the V_{sfm} radius using Marching Cubes [7] algorithm with a grid resolution of 1024.

4.1 Mesh Reconstruction

In this section, we show our results on historical datasets and compare to other state-of-the-art methods. We choose one classical method, Metashape [1] and learning-based algorithms NeusW [19] and Neus-factor [13]. Fig. 1 shows the aligned 3D mesh with images that have been taken from that angle. The figure demonstrated how our work can bring the historical monuments alive and

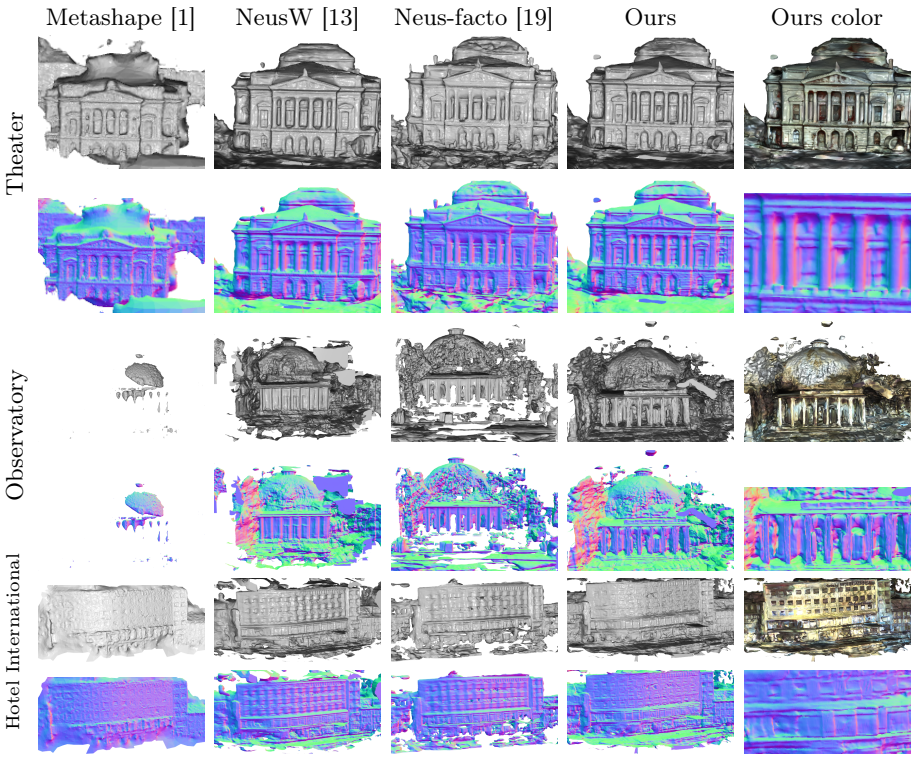


Fig. 5: Reconstructed mesh results compared to other methods. Metashape [1] can get clean geometry reconstruction but fails to get the details. Our method is able to provide comparable mesh reconstructions while additionally recovering the color of the mesh.

positively influence future study. Fig. 5 shows our reconstructed meshes for four historical datasets. In general, all the methods achieve the best results on the National Theater dataset, especially for the facade. Qualitatively, our method recovers comparable meshes to other methods. For the Observatory dataset, despite the limited data and challenging setting, learning-based methods can successfully recover the main building with varying degrees of artifacts. Our method gives the most complete and round dome. However, the normal meshes (4th-row in Fig. 5) indicate that we are able to recover the pillars correctly while NeusW [13] fails on this part. A similar situation happens in the Hotel dataset, ours can recover thin structures such as columns and chimneys. We attribute this to the dense point cloud supervision.

5 Colorization Accuracy

We compare the rendered results of the baseline, an unconditional colorization method [5], and our color appearance loss with the ground truth color image which was converted to gray-scale during training. For the unconditional colorization case we colorize all images before training. Fig. 6 shows the differ-

ent results of colorization. While the unconditional colorization results in a

Fig. 6: Colorization comparison: The baseline is gray-scale and therefore unnatural, the unconditional colorization results in plausible but historically inaccurate colors. The Color Appearance loss results in the closest color scheme to the ground truth.

plausible-looking colorization, it's not historically accurate, since it doesn't retain any colors from the available color images. Our method with the color appearance loss achieves the best performance in terms of PSNR and SSIM values but falls short in LPIPS as shown in Tab. 1.

	PSNR	SSIM	LPIPS
Baseline	18.01	0.7299	0.2871
Unconditional Colorization	17.30	0.7408	**0.2992**
Appearance Embedding Loss (Ours)	**19.02**	**0.7528**	0.2654

Table 1: Quantitative evaluation of colorization performance compared to the baseline, unconditional colorization and our proposed color embedding loss

6 Conclusion

Summary We present a novel historical dataset with a substantially larger image count compared to prior datasets. Additionally, we provide its point cloud data along with camera information, generated through Structure from Motion (SfM). Our method addresses challenges inherent in reconstructing 3D shapes from sparse and low-quality inputs found in archival historical datasets. We demonstrate that incorporating pre-existing data, such as dense point clouds, significantly enhances geometry reconstruction. This supervision of dense point clouds improves reconstruction outcomes, particularly in scenes with limited image coverage. It facilitates the reconstruction of thin structures, texture-less wall segments, and temporally changing structures. Furthermore, we propose a color appearance embedding loss to restore the color of generated meshes representing historical buildings to some extent.
Limitations The color appearance embedding loss has been observed to negatively impact mesh accuracy in quantitative terms. Moreover, there is a need for enhanced capabilities in handling sparse input images to enable the recovery of detailed 3D meshes, especially under more challenging conditions.

Acknowledgement. The research upon which this paper is based has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101135556 (INDUX-R).

References

1. Agisoft LLC: Agisoft Metashape (Version 2.0.0 build 15597) [Software]. http://www.agisoft.com/ (2022), accessed: 2023-11-16
2. Chumachenko, K., Männistö, A., Iosifidis, A., Raitoharju, J.: Machine learning based analysis of finnish world war ii photographers. IEEE Access **8**, 144184–144196 (2020). https://doi.org/10.1109/ACCESS.2020.3014458
3. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: CVPR Deep Learning for Visual SLAM Workshop (2018)
4. Fu, Q., Xu, Q., Ong, Y.S., Tao, W.: Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction (2022)
5. Kang, X., Yang, T., Ouyang, W., Ren, P., Li, L., Xie, X.: Ddcolor: Towards photo-realistic image colorization via dual decoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 328–338 (2023)
6. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: LightGlue: Local Feature Matching at Light Speed. In: ICCV (2023)
7. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques. p. 163–169. SIGGRAPH '87, Association for Computing Machinery, New York, NY, USA (1987)
8. Maiwald, F.: Generation of a Benchmark Dataset Using Historical Photographs for an Automated Evaluation of Different Feature Matching Methods. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. **XLII-2/W13**, 87–94 (2019). https://doi.org/10.5194/isprs-archives-XLII-2-W13-87-2019
9. Maiwald, F., Komorowicz, D., Munir, I., Beck, C., Münster, S.: Semi-automatic generation of historical urban 3d models at a larger scale using structure-from-motion, neural rendering and historical maps. In: Münster, S., Pattee, A., Kröber, C., Niebling, F. (eds.) Research and Education in Urban History in the Age of Digital Libraries. pp. 107–127. Springer Nature Switzerland, Cham (2023)
10. Poli, D., Casarotto, C., Strudl, M., Bollmann, E., Moe, K., Legat, K.: USE OF HISTORICAL AERIAL IMAGES FOR 3d MODELLING OF GLACIERS IN THE PROVINCE OF TRENTO. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLIII-B2-2020**, 1151–1158 (aug 2020). https://doi.org/10.5194/isprs-archives-xliii-b2-2020-1151-2020
11. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
12. Skublewska-Paszkowska, M., Milosz, M., Powroznik, P., Lukasik, E.: 3d technologies for intangible cultural heritage preservation—literature review for selected databases. Heritage Science **10**(1) (jan 2022)
13. Sun, J., Chen, X., Wang, Q., Li, Z., Averbuch-Elor, H., Zhou, X., Snavely, N.: Neural 3D reconstruction in the wild. In: SIGGRAPH Conference Proceedings (2022)
14. Tyszkiewicz, M., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. Advances in Neural Information Processing Systems **33** (2020)
15. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)

16. Wang, Y., Han, Q., Habermann, M., Daniilidis, K., Theobalt, C., Liu, L.: Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
17. Wang, Z., Li, Q.: Information content weighting for perceptual image quality assessment. IEEE Transactions on Image Processing **20**(5), 1185–1198 (2011)
18. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021)
19. Yu, Z., Chen, A., Antic, B., Peng, S., Bhattacharyya, A., Niemeyer, M., Tang, S., Sattler, T., Geiger, A.: Sdfstudio: A unified framework for surface reconstruction (2022)