

# Coloring the Past: Neural Historical Buildings Reconstruction from Archival Photography

Dávid Komorowicz<sup>\*1,3</sup> Lu Sang<sup>\*1</sup> Ferdinand Maiwald<sup>2</sup> Daniel Cremers<sup>1</sup>

<sup>1</sup>Technical University of Munich  
Computer Vision Group

<sup>2</sup>Dresden University of Technology  
Institute of Photogrammetry and Remote Sensing

<sup>3</sup>Friedrich Schiller University Jena  
Chair for Digital Humanities

david.komorowicz, lu.sang, cremers@tum.de    ferdinand.maiwald@tu-dresden.de



Figure 1. **Coloring the Past:** Reconstructing high-quality historical buildings from limited data and recovering color appearance from a majority of gray-scale images. The figure shows photographs of the Hungarian National Theater over a long time period (left), reconstructed mesh with color (middle), and shaded mesh + normal map (right).

## Abstract

Historical buildings are a treasure and milestone of human cultural heritage. Reconstructing the 3D models of these building hold significant value. The rapid development of neural rendering methods makes it possible to recover the 3D shape only based on archival photographs. However, this task presents considerable challenges due to the limitations of such datasets. Historical photographs are often limited in number and the scenes in these photos might have altered over time. The radiometric quality of these images is also often sub-optimal. To address these challenges, we introduce an approach to reconstruct the geometry of historical buildings, employing volumetric rendering techniques. We leverage dense point clouds as a geometric prior and introduce a color appearance embedding loss to recover the color of the building given limited available color images. We aim for our work to spark increased interest and focus on preserving historical buildings. Thus, we also introduce a new historical dataset of the Hungarian National Theater, providing a new

benchmark for the reconstruction method.

## 1. Introduction

Historical buildings embody the unique characteristics of cultural heritage and are seen as landmarks that connect people across time and countries. They make history tangible, yet they are vulnerable to temporal and man-made changes. It is one of the main goals of UNESCO to protect and preserve our cultural heritage which becomes possible because of the rapid development of 3D technologies [42].

(Historical) images are usually processed in a Structure-from-Motion (SfM) workflow to estimate the intrinsic and extrinsic camera parameters and a dense 3D scene is conventionally reconstructed using multi-view stereo (MVS) [40]. With the advent of representing 3D scenes as Neural Radiance Fields (NeRF) [33], the published framework facilitates the synthesis of photo-realistic images from novel viewpoints, using a volumetric scene representation learned from sparse and unstructured 2D images. The ability of NeRF to interpolate and extrapolate from image data introduces an unprecedented potential for reconstructing his-

\* These authors contributed equally.

torical buildings using mostly historical photographs as a source [29].

However, compared to the reconstruction of modern buildings or scenes, recovering synthetic views and the 3D shape of historical buildings comes with several limitations. A significant issue lies in the scarcity and the often-compromised quality of available input data [21, 27, 34]. Many historical sites are documented solely through antiquated photographs, captured with obsolete equipment that takes images with difficult radiometric properties such as blurriness, lack of color, or absence of accurate camera parameters [9, 26, 36]. Additionally, only a limited number of photos can be found and used, coupled with the extensive temporal spread across which they were taken, which means that inconsistencies are in both the structural state of the buildings and in the photographic records themselves [13, 26].

In this paper, we propose a method that tackles 3D reconstruction and colored view syntheses for historical buildings leveraging the sparse and low-quality input images. Fig. 1 shows the reconstruction results of our method. Given a historical image collection dominated by gray-scale images over different time, we are able to recover the colored 3D mesh of the building. We also would like to raise interest in the topic of historic monument reconstruction and the use of historical photographs in the 3D reconstruction community.

In summary, our contributions are as follows:

- We propose a method that is able to 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.
- Our method achieves better reconstruction results by incorporating already existing data.
- We publish a historical dataset that showcases a wide range of properties typically present in historical datasets.

## 2. Related Work

**Multi-view 3D reconstruction** Reconstructing the underlying 3D geometry from multiple images from different viewpoints is a long-studied problem in the computer vision field. Conventional multi-view stereo (MVS) methods [2, 7, 8, 22, 39, 41, 43], consider matching geometry priors such as depth [41] or using voxel as surface representation and project points back to the image to refine the geometry [39, 43]. A limitation of traditional approaches is, that they often rely on discrete representations like depth maps, or voxel volumes to model surface space. This approach can lead to substantial memory usage when dealing with large scenes. Additionally, the process of finding correspondences is generally vulnerable to noise, which can

affect the accuracy of the reconstruction. **Learning-based** multi-view methods [31, 50, 53], usually use networks to learn or extract features from color images with [10, 16, 57] or without [50, 53] geometry priors. They are more robust to noise and generalize across different scenes and objects. However, they require large amounts of training data, which can be computationally intensive and may not be available for historical datasets.

**Surface representation scheme** During 3D reconstruction, the choice of surface representation method is crucial. Traditional approaches often utilize discrete methods like point clouds, polygon meshes, or discrete Signed Distance Fields (SDF) [35, 39, 43]. On the other hand, learning-based techniques, aided by neural networks, can employ continuous surface representation methods. The most commonly used property is density [31, 33], which indicates the transparency value of the given point. However, this approach, while effective for view synthesis, often fails in accurately reconstructing geometry. It calculates an integrated opacity value along a ray, rather than modeling an explicit surface point. Alternatively, continuous SDF, another form of surface representation, offers a more precise approach. It calculates the distance from any point in space to the nearest surface, indicating whether the point is inside (negative) or outside (positive) the object. This representation enables SDF-based methods [44, 50, 53], more accurate identification of surfaces, and better handling of occlusions and view-dependent effects compared to volume-only-based methods. NeuS [49], for instance, uses a differentiable rendering framework that combines SDF with a radiance field, resulting in high-quality images that capture fine geometric details. Neus-Facto [56] propose a network without surface-guided sampling and geometry prior. GeoNeus [16] uses a sparse point cloud to supervise the SDF and a Photometric consistency loss. This is not suitable for historical imagery due to large illumination and appearance inconsistencies between images.

**Historical building dataset** The use of historical public and press photography showing terrestrial scenes is still a rather rare scenario for complex 3D reconstruction [3, 12, 27], whereas historical aerial images are already commonly and increasingly used in Structure-from-Motion workflows [13, 15, 21, 28, 58]. Historical terrestrial images are mainly used for special tasks such as horizon line detection [32], photographer recognition [9], and building height estimation [14]. Other approaches focus on the integration of historical images into further existing data such as terrestrial laser scanning [4, 6], and contemporary photographs [20, 30]. When working with historical images, the standard 3D reconstruction workflow is usually interrupted after sparse point cloud creation and camera pose estimation because conventional MVS strategies fail to generate reasonable surface representations [27].

To overcome the difficulties posed by gray-scale, low-quality, and sparse input images, we integrate a dense point cloud as a geometric prior and introduce a novel color appearance embedding loss. We are confident that our methodology holds substantial value in preserving the historical heritage of human culture.

### 3. Historical Dataset

Reconstructing historical buildings based on archival photography provides significant value not only in the research area but also considering 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. Thus, we introduce 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 registered 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 ( Tab. 1). Another significant difference is, that the building can slightly change over the span of decades which is why we provide the capture dates of the images.

Besides its cultural significance to the Hungarian people, this 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. 2 shows the reconstructed point cloud and estimated camera locations using SfM [40] from the National Theater dataset. Fig. 3 shows example images of the facade across time.

Dataset name	Total	Color	Train
National Theater (Ours)	229	16	153
Hotel International [29]	19	1	18
Observatory [29]	37	3	33
St. Michael Church [29]	17	0	16

Table 1. Datasets statistics of National Theater dataset, three historical datasets from [29].

These images are private and can be accessed upon purchase. We will provide the link to these images.

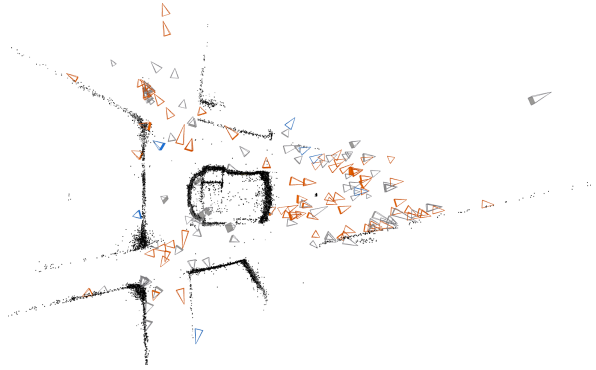


Figure 2. Top down of view of a reconstructed point cloud of the National Theater dataset: blue cameras stand for validation views, orange cameras are training images, gray cameras are images that can be obtained via request.

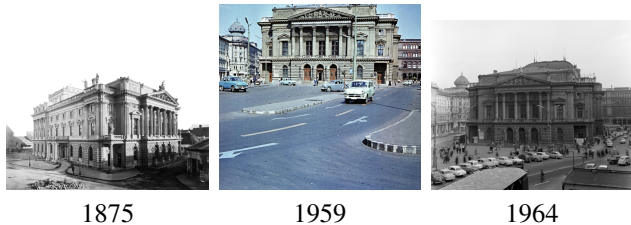


Figure 3. Example images from the Hungarian National Theater dataset.

Tab. 1 summarizes all information for the released historical dataset and three further historical datasets [29]. We use these four datasets to test our methods in Sec. 5. The first column shows the total number of images for the dataset, the second column provides the number of color images, and the third column shows the number of images that we actually use to train our model. Due to the quality of the images, not all of them are suitable. From Tab. 1 we can see that our dataset contains an order of magnitude more images in total and more color images as well.

### 4. Method

The whole pipeline of our method is as follows. Given a set of images  $\{\mathcal{I}_i\}$ , for  $i \in \{0, 1, \dots, n\}$ , we first 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 [40]. We run a segmentation method, similar to [44] to mask out irrelevant objects such as people and cars. We generate two kinds of point clouds, a **sparse** point cloud, directly using SfM [40] and a **dense** point cloud  $\mathcal{P} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  using the estimated cam-

Fortepan by UVATERV/FÖMTERV/Zsolt Pálincás/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

era parameters by multiview stereo fusion [40]. 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.

#### 4.1. Backbones and Geometry Loss

We build our method on top of NeusW [44]. 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  $\mathbf{n} = \nabla \text{MLP}_{\text{SDF}}(\mathbf{x})$ , and a feature vector  $\mathbf{f} \in \mathbb{R}^{f_n}$  with dimension  $f_n$ .

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

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

where  $\mathbf{e}_i$  are appearance embeddings corresponding to each input photo, optimized alongside the parameters of MLPs, see [44] for more details. We first initialize a voxel grid by the sparse point cloud similar to [44]. 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 [16] is

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

where  $|\mathcal{P}_i|$  is the number of points in the point cloud and  $\lambda$  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 [44] 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.

#### 4.2. Color Appearance Embedding

To deal with the situation that most of the input images are available as gray-scale, and only a small portion pro-

vides 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 less-than-ideal appearance embedding and a gray-scale output. The rendered color for a ray  $\mathbf{r}$  is

$$\mathbf{C}'(\mathbf{r}) = \int_0^{+\infty} w(t)c(\mathbf{r}(t), \mathbf{v}, \mathbf{f})dt, \quad (4)$$

where  $w(t)$  is an unbiased and occlusion-aware weight function used in [49]. The color net outputs a three-channel color vector, to supervise it using gray-scale images, we use perceptual weights [51] to convert the output color to gray-scale value, *i.e.*, for  $\mathbf{C}'(\mathbf{r}) = (c_r, c_g, c_b)$ , we propose the function  $g: \mathbb{R}^3 \rightarrow \mathbb{R}$  and

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

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

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

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

### 5. Experiments

**Validation Dataset** We evaluate our method on the **historical datasets** listed in Tab. 1 and one **modern dataset** which shows the Brandenburg Gate [54]. We exclude certain images in the historical dataset where the main object is not visible or the building is demolished. The total number of images used for training is listed in Tab. 1. The historical datasets consist of significantly fewer images compared to the Brandenburg Gate dataset. They also introduce challenges such as limited viewing angles (the Observatory dataset), or a wide range of lighting conditions (the Hotel dataset, see in Fig. 6). The Brandenburg Gate dataset provides corresponding ground truth in the form of LiDAR measurements [44]. Hence, we use it for quantitative evaluation. To close the gap between the modern and historic

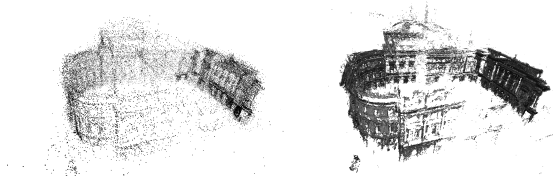


Figure 4. Comparison of sparse (left) and dense (right) point cloud generated by stereo fusion [40].

domains, we transform 90% of the images in the Brandenburg Gate dataset into gray-scale. To show the influence of input data quantity, we sample 10 and 90 images from the dataset.

**Implementation Details** For our method and its comparison to other methods we build upon the implementation of SDFStudio [56]. For dense point cloud generation, we import the camera poses and feature matches obtained by COLMAP [40] and a combination of state-of-the-art key-point detector and feature matching algorithms [11, 23, 47] via the bundler format and apply the segmentation masks. We use 8 layers with 512 hidden units for the geometry MLP and 4 layers with 256 hidden units for the color MLP.

For the historic datasets, 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 remove floating blobs occluding the view from the validation views disconnected from the object of interest for historic datasets. 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 [24] algorithm with a grid resolution of 1024.

### 5.1. Surface Reconstruction Results

In this section, we show our surface reconstruction results in comparison to other methods. Fig. 7 shows our reconstructed meshes for four different historical datasets and its comparison to other state-of-the-art conventional [1] and learning-based MVS algorithms [44, 56] in terms of reconstruction quality. Tab. 2 shows the quantitative comparison results on the Brandenburg Gate dataset.

In general, all of the methods achieve the best results on the National Theater dataset, especially for the facade. Qualitatively, our method recovers comparable meshes to other methods. However, as we mentioned in Sec. 3, the back side is more challenging due to the lack of images facing this side. We show the results from different viewing angles in Fig. 5. All methods are able to recover the front part of the Theater, but the back part is recovered unsatisfactorily. NeusW [44] generates more noise compared to Neus-facto [56] and ours. We are able to reconstruct the scene without holes as opposed to Neus-facto [56]. For the other datasets, Metashape (using conventional MVS) is only able to recover a small part of the geometry or completely fails, but it tends to provide a closed and clean surface.

For the Observatory dataset, in spite of the limited data and challenging setting, learning-based methods can suc-

cessfully 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. 7) indicate that we are able to recover the pillars correctly while NeusW [44] fails on this part. A similar situation happens in the Hotel and Church datasets, ours is able to recover thin structures such as columns and chimneys. We attribute this to the dense point cloud supervision. Finally, our method can recover the colored meshes for the given datasets as shown in the last column in Fig. 7.

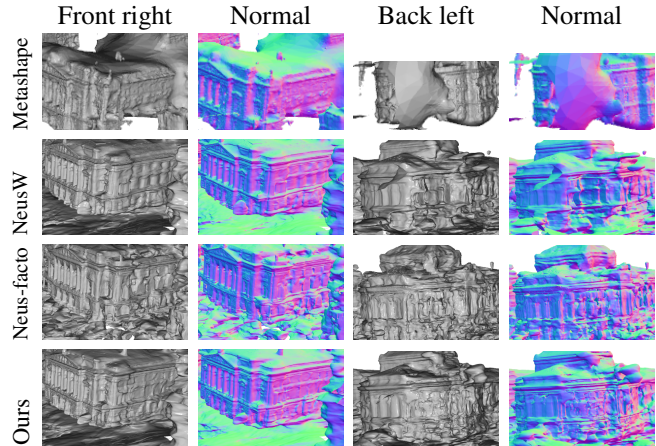


Figure 5. Reconstruction result of different sides of our dataset showing increasing difficulty in terms of number of images. Clearly, the facade can be recovered with intricate detail by all methods, followed by the right side which is satisfactory but smaller deviations arise. The back left corner is the most difficult which is very noisy for almost all methods. Still, ours provides relatively cleaner and complete surface reconstruction.

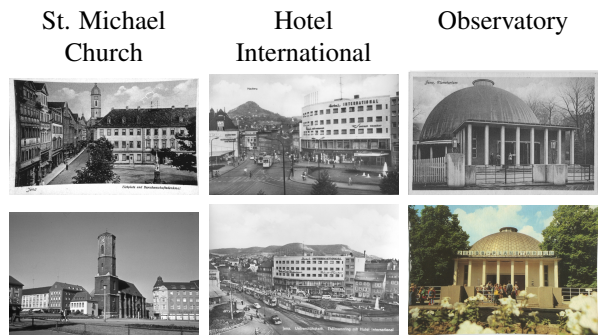


Figure 6. Example images from the historical datasets [29] The lighting and view angles are largely changed in all datasets. Images are far from targeted buildings as well. All the features make reconstruction challenging.

To quantitatively evaluate the 3D geometry reconstruction results, we provide precision (P), recall (R), and F1 score (F1) under three different thresholds of the generated meshes in Tab. 2 following the procedure of [44]. The

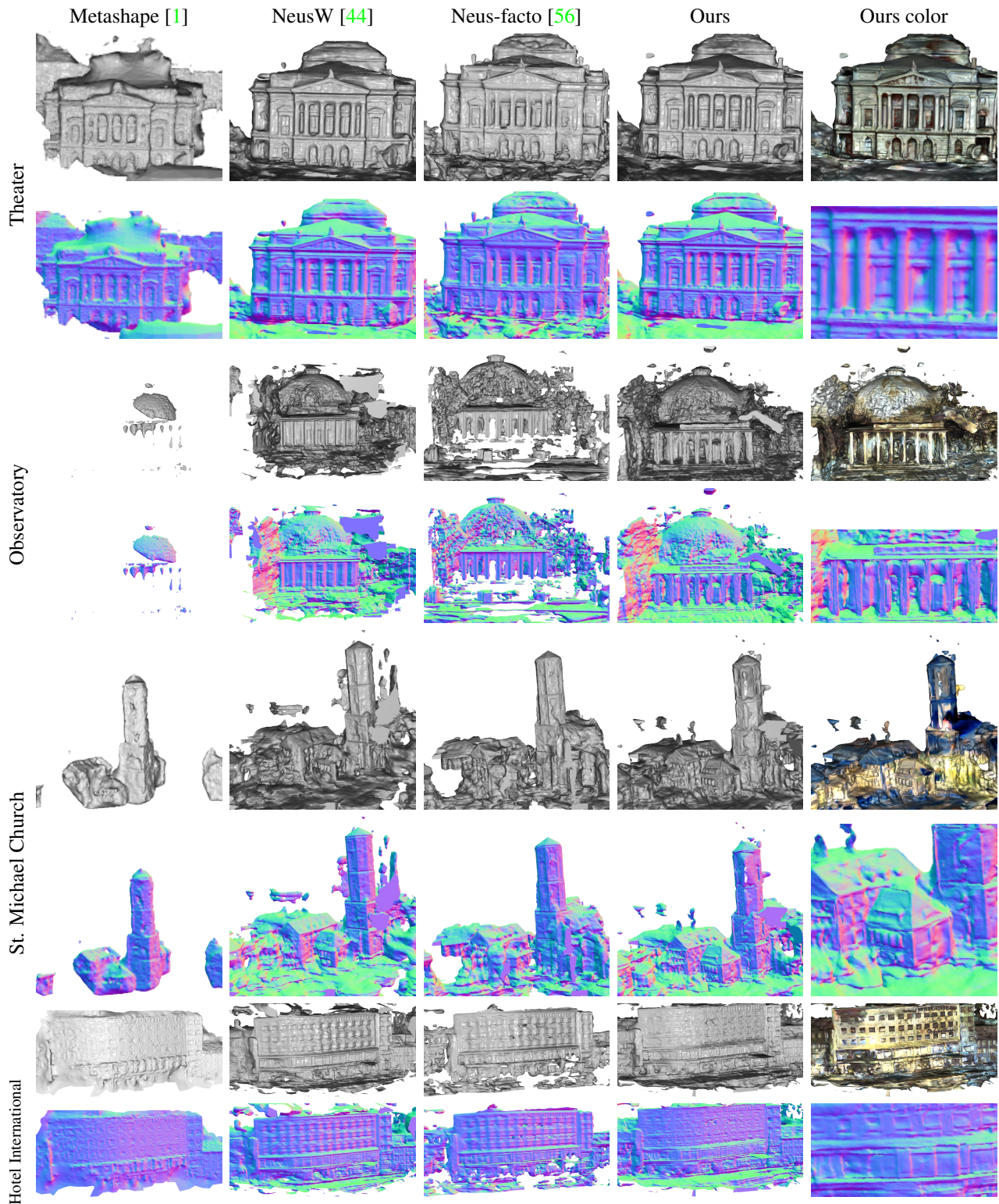


Figure 7. Reconstructed mesh results compared to other methods. Metashape [1] can get clean geometry reconstruction but fails to get the details. Our method is able to provide comparable mesh reconstructions while additionally recovering the color of the mesh.

Settings		Low			Medium			High			All (AUC)		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
10 frames	Metashape	<b>73.0</b>	8.0	14.4	<b>85.0</b>	15.4	26.0	<b>91.2</b>	19.1	31.6	<b>88.7</b>	22.2	35.0
	NeusW	37.7	<u>26.6</u>	31.2	55.4	<b>44.0</b>	<b>49.1</b>	67.0	<b>55.0</b>	<b>60.4</b>	70.8	<b>61.5</b>	<b>65.6</b>
	NeuS-Facto	<u>40.1</u>	25.7	<u>31.3</u>	<u>56.8</u>	40.8	47.5	<u>68.0</u>	51.3	58.5	<u>72.2</u>	58.2	<u>64.2</u>
	Ours	38.6	<b>27.8</b>	<b>32.3</b>	55.0	<u>43.5</u>	<u>48.6</u>	66.0	<u>53.0</u>	<u>58.8</u>	70.3	<u>59.2</u>	64.1
90 frames	Metashape	<b>72.9</b>	14.9	24.7	<b>85.1</b>	34.1	48.7	<b>90.9</b>	43.0	58.4	<b>88.6</b>	46.8	60.0
	NeusW	59.7	44.7	51.1	73.7	<u>59.4</u>	<u>65.8</u>	81.1	66.0	<u>72.8</u>	81.3	<u>68.6</u>	74.1
	NeuS-Facto	<u>64.9</u>	<b>47.0</b>	<b>54.5</b>	<u>79.1</u>	<b>61.4</b>	<b>69.1</b>	<u>86.4</u>	<b>69.0</b>	<b>76.7</b>	<u>85.0</u>	71.5	<b>77.3</b>
	Ours	60.1	<u>45.6</u>	<u>51.8</u>	73.4	<u>59.4</u>	65.6	81.1	<u>66.1</u>	<u>72.8</u>	81.5	<b>69.0</b>	<u>74.4</u>

Table 2. Quantitative Comparison: We compare against other methods on two different settings of the Brandenburg Gate dataset. We report the precision (P), recall (R), and F1 scores. The best results are in bold, and the second best are underlined. The numbers indicate that our method is comparable to or outperforms other methods in terms of recall, especially in the 10-image setting, which is the most closed case with historical datasets. For other cases, we still achieve comparable results.

precision indicates the accuracy of the reconstructed mesh compared to the ground truth mesh and the recall indicates the completeness of the results, the F1-score is a weighted score computed using precision and recall. Fig. 8 illustrates these metrics. The three thresholds (Low, Medium, and High) correspond to 0.1, 0.2, and 0.3 meters respectively. Additionally, the area under the curve (AUC) combines all

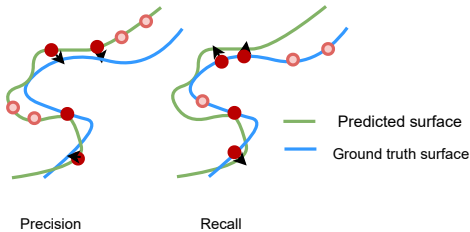


Figure 8. Demonstration of precision and recall. The red dots are under the chosen threshold, so they are used to compute the precision or recall, pink dots are over the threshold thus they are ignored.

thresholds into a single metric.

The quantitative comparison of the Brandenburg Gate dataset shows that we achieve results comparable to other state-of-the-art methods on mesh reconstruction tasks. Our method outperforms other methods in the 10-image scenario (which is the most comparable case with historical datasets) at Low threshold for R and F1-score, which means that the mesh is more complete and closer to the ground truth mesh. Metashape [1] gains high scores for precision because it recovers the overall shape accurately. However, the mesh is incomplete bringing down the total F1 score as a result. Note that the image quality of the Brandenburg Gate dataset is still superior to the images in historical datasets.

## 5.2. Ablation Study

To analyze the influence of our proposed loss functions and the number of input images, we train our model using 10 images and 90 images from the Brandenburg Gate dataset where we set 90% of images to gray-scale. To quantitatively evaluate the 3D geometry reconstruction results, we also provide precision (P), recall (R), and F1 value (F1) of the generated meshes in Tab. 3. The baseline setting is NeusW [44] without geometrical loss term (3). We gradually add only the color appearance loss (+ color), only the geometrical loss based on two different point clouds (+ sparse geo, + dense geo), and finally our setting, *i.e.* with color appearance loss and dense geometric loss. Tab. 3 shows that the color embedding loss slightly degrades the geometry.

The dense point cloud supervision consistently outperforms the sparse point cloud supervision using the geometry loss. Fig. 9 visualizes the meshes for the ablation study. The color appearance loss does not degrade the mesh qualitatively and results in small gains. That is in the 10-image case it recovers the legs of the horses and recovers one or more holes in both 10 and 90 images situations. In the case of the 10-image input (comparable to historical datasets), the sharp structure on the top left roof gets filled in the baseline case. This is almost entirely eliminated with the color appearance embedding loss.

## 6. Conclusion

**Summary** We introduced a new historical dataset that has significantly more images than previous datasets and also provide its point cloud along with camera information which is generated via SfM. We propose a method that tackles challenges such as sparse and low-quality inputs, when reconstructing 3D shapes using archival historical datasets.

Settings		Low			Medium			High			All (AUC)		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
10 frames	Baseline	37.7	26.6	31.2	55.4	44.0	49.1	67.0	55.0	<u>60.4</u>	70.8	61.5	65.6
	+ Color loss	37.4	26.7	31.1	54.9	43.6	48.6	66.6	55.0	60.3	70.0	61.8	65.4
	+ Sparse geo	<u>38.9</u>	<u>29.7</u>	<u>33.7</u>	<u>55.9</u>	<u>46.6</u>	<u>50.8</u>	<u>67.4</u>	<u>57.4</u>	62.0	<u>71.4</u>	<b>63.2</b>	<u>66.9</u>
	+ Dense geo	<b>42.9</b>	<b>33.0</b>	<b>37.3</b>	<b>59.5</b>	<b>49.3</b>	<b>53.9</b>	<b>69.5</b>	<b>58.3</b>	<b>63.4</b>	<b>72.9</b>	<u>63.1</u>	<b>67.5</b>
	Ours	38.6	27.8	32.3	55.0	43.5	48.6	66.0	53.0	58.8	70.3	59.2	64.1
90 frames	Baseline	59.7	44.7	51.1	<b>73.7</b>	59.4	65.8	<u>81.1</u>	66.0	72.8	81.3	68.6	74.1
	+ Color loss	<u>60.1</u>	45.3	51.7	<u>73.4</u>	59.2	65.6	80.8	65.8	72.5	81.0	68.3	73.9
	+ Sparse geo	59.5	<u>45.7</u>	51.7	<u>73.4</u>	<u>60.1</u>	<u>66.1</u>	80.9	<u>67.1</u>	<u>73.3</u>	81.1	<u>70.0</u>	<u>74.9</u>
	+ Dense geo	<b>60.8</b>	<b>46.6</b>	<b>52.8</b>	<b>73.7</b>	<b>60.4</b>	<b>66.4</b>	<b>81.2</b>	<b>67.4</b>	<b>73.6</b>	<b>81.8</b>	<b>70.1</b>	<b>75.2</b>
	Ours	<u>60.1</u>	45.6	<u>51.8</u>	<u>73.4</u>	59.4	65.6	<u>81.1</u>	66.1	72.8	<u>81.5</u>	69.0	74.4

Table 3. Ablation Study: We report the precision (P), recall (R), and F1 scores over two different settings of the Brandenburg Gate dataset for the ablation study. The best results are in bold, and the second best are underlined. The numbers indicate that the geometric priors, especially the dense point cloud, contribute to the reconstruction. It gives the best results in overall settings for different scores in most of the cases. Even though the color embedding appearance loss slightly deteriorates the results, it still gives comparable accuracy to the best scores.

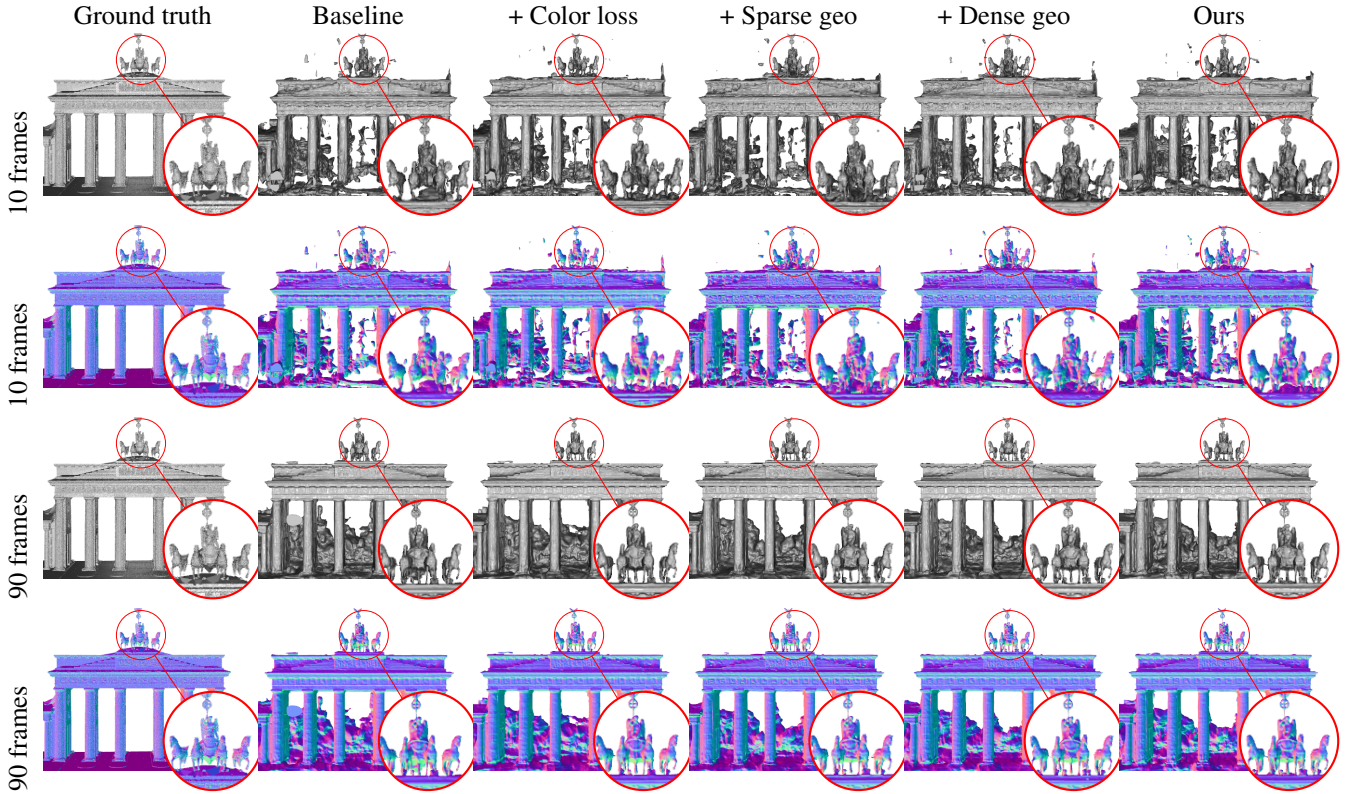


Figure 9. Ablation study: reconstructed mesh and normal maps on the proposed losses and the number of input images on the Brandenburg Gate dataset. The dense point cloud provides a better geometric prior compared to sparse situations, especially for the sparse input (10 frames) case (first two rows). Horse legs are recovered using the proposed method. Under the more dense input (90 frames) situation, the baseline and other settings have provided relatively good results. Quantitatively, with only dense geometric loss still helps improve the accuracy.



We showed that incorporating existing data such as dense point clouds can significantly improve the geometry reconstruction. The dense point cloud supervision enhances the reconstruction, especially the scenes with few images. It enables thin structures and flat, texture-less wall segments to be reconstructed and also recovers structures that are temporally changing. Moreover, we propose a color appearance embedding loss to recover the color of the generated mesh of the historical buildings.

**Limitations and future works** The color appearance embedding loss decreases the mesh accuracy quantitatively. The capability of dealing with sparse input images still need to be improved to be able to recover detailed 3D meshes under more extreme situation. Next, we plan to explore methods that are especially targeted at few-shot view synthesis [17, 52] and reconstruction methods.

## References

- [1] Agisoft LLC. Agisoft Metashape (Version 2.0.0 build 15597) [Software]. <http://www.agisoft.com/>, 2022. Accessed: 2023-11-16. 5, 6, 7, 12, 13
- [2] M. Agrawal and L.S. Davis. A probabilistic framework for surface reconstruction from multiple images. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (CVPR), volume 2, pages II–II, 2001. 2
- [3] C. Beltrami, D. Cavezzali, F. Chiabrando, A. Iaccarino Idelson, G. Patrucco, and F. Rinaudo. 3d digital and physical reconstruction of a collapsed dome using sfm techniques from historical images. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W11:217–224, may 2019. 2
- [4] M. G. Bevilacqua, G. Caroti, A. Piemonte, and D. Ulivieri. Reconstruction of lost architectural volumes by integration of photogrammetry from archive imagery with 3-d models of the status quo. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W9:119–125, jan 2019. 2
- [5] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4160–4169, 2023.
- [6] G. Bitelli, M. Dellapasqua, V. A. Girelli, S. Sbaraglia, and M. A. Tinia. Historical photogrammetry and terrestrial laser scanning for the 3d virtual reconstruction of destroyed structures: A case study in italy. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-5/W1:113–119, may 2017. 2
- [7] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo - stereo matching with slanted support windows. In BMVC, January 2011. 2
- [8] Jeremy S. De Bonet. Poxels: Probabilistic voxelized volume reconstruction. 1999. 2
- [9] Kateryna Chumachenko, Anssi Männistö, Alexandros Iosifidis, and Jenni Raitoharju. Machine learning based analysis of finnish world war ii photographers. IEEE Access, 8:144184–144196, 2020. 2
- [10] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12882–12891, 2022. 2
- [11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In CVPR Deep Learning for Visual SLAM Workshop, 2018. 5, 12
- [12] Peter L. Falkingham, Karl T. Bates, and James O. Farlow. Historical photogrammetry: Bird’s paluxy river dinosaur chase sequence digitally reconstructed as it was prior to excavation 70 years ago. PLoS ONE, 9(4):e93247, apr 2014. 2

- [13] E. M. Farella, L. Morelli, F. Remondino, J. P. Mills, N. Haala, and J. Crompvoets. THE EUROSDR TIME BENCHMARK FOR HISTORICAL AERIAL IMAGES. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B2-2022:1175–1182, may 2022. [2](#)
- [14] Elisa Mariarosaria Farella, Emre Özdemir, and Fabio Remondino. 4d building reconstruction with machine learning and historical maps. Applied Sciences, 11(4):1445, feb 2021. [2](#)
- [15] D. Feurer and F. Vinatier. Joining multi-epoch archival aerial images in a single SfM block allows 3-d change detection with almost exclusively image information. ISPRS Journal of Photogrammetry and Remote Sensing, 146:495–506, dec 2018. [2](#)
- [16] Qiancheng Fu, Qingshan Xu, Yew-Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction, 2022. [2](#), [4](#)
- [17] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. CoRR, abs/2104.00677, 2021. [9](#)
- [18] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 5885–5894, October 2021.
- [19] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis, 2021.
- [20] P. Kalinowski, F. Both, T. Luhmann, and U. Warnke. Data fusion of historical photographs with modern 3d data for an archaeological excavation – concept and first results. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B2-2021:571–576, jun 2021. [2](#)
- [21] Friedrich Knuth, David Shean, Shashank Bhushan, Eli Schwat, Oleg Alexandrov, Christopher McNeil, Amaury Dehecq, Caitlyn Florentine, and Shad O’Neel. Historical structure from motion (HSfM): Automated processing of historical aerial photographs for long-term topographic change analysis. Remote Sensing of Environment, 285:113379, feb 2023. [2](#)
- [22] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, ECCV, pages 796–811, Cham, 2018. Springer International Publishing. [2](#)
- [23] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In ICCV, 2023. [5](#), [12](#)
- [24] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’87, page 163–169, New York, NY, USA, 1987. Association for Computing Machinery. [5](#)
- [25] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. SIGGRAPH Comput. Graph., 21(4):163–169, aug 1987.
- [26] Ferdinand Maiwald. 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. [2](#)
- [27] Ferdinand Maiwald. A window to the past through modern urban environments — Developing a photogrammetric workflow for the orientation parameter estimation of historical images. PhD thesis, Technische Universität Dresden, 2022. [2](#)
- [28] Ferdinand Maiwald, Denis Feurer, and Anette Eltner. Solving photogrammetric cold cases using ai-based image matching: New potential for monitoring the past with historical aerial images. ISPRS Journal of Photogrammetry and Remote Sensing, 206:184–200, 2023. [2](#)
- [29] Ferdinand Maiwald, Dávid Komorowicz, Iqra Munir, Clemens Beck, and Sander Münster. Semi-automatic generation of historical urban 3d models at a larger scale using structure-from-motion, neural rendering and historical maps. In Sander Münster, Aaron Pattee, Cindy Kröber, and Niebling Florian, editors, Research and Education in Urban History in the Age of Digital Libraries, pages 107–127, Cham, 2023. Springer Nature Switzerland. [2](#), [3](#), [5](#), [12](#)
- [30] Ferdinand Maiwald, Theresa Vietze, Danilo Schneider, Frank Henze, Sander Münster, and Florian Niebling. Photogrammetric analysis of historical image repositories for virtual reconstruction in the field of digital humanities. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 42:447–452, 2017. [2](#)
- [31] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In CVPR, 2021. [2](#)
- [32] Sebastian Mikolka-Flöry and Norbert Pfeifer. Horizon line detection in historical terrestrial images in mountainous terrain based on the region covariance. Remote Sensing, 13(9):1705, apr 2021. [2](#)
- [33] 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](#)
- [34] L. Morelli, F. Bellavia, F. Menna, and F. Remondino. Photogrammetry now and then – from hand-crafted to deep-learning tie points –. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-2/W1-2022:163–170, dec 2022. [2](#)
- [35] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In ISMAR, pages 127–136. IEEE Computer Society, 2011. [2](#)
- [36] D. Poli, C. Casarotto, M. Strudl, E. Bollmann, K. Moe, and K. Legat. USE OF HISTORICAL AERIAL IM-

- AGES 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. [2](#)
- [37] Chen Quei-An. Nerf-pl: a pytorch-lightning implementation of nerf, 2020.
- [38] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12932–12942, 2022.
- [39] L Sang, B Haefner, X Zuo, and D Cremers. High-quality rgb-d reconstruction via multi-view uncalibrated photometric stereo and gradient-sdf. In IEEE Winter Conference on Applications of Computer Vision (WACV), Hawaii, USA, January 2023. [2](#)
- [40] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In Conference on Computer Vision and Pattern Recognition (CVPR), 2016. [1](#), [3](#), [4](#), [5](#), [12](#)
- [41] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, ECCV, pages 501–518, Cham, 2016. Springer International Publishing. [2](#)
- [42] Maria Skublewska-Paszowska, Marek Milosz, Pawel Powroznik, and Edyta Lukasik. 3d technologies for intangible cultural heritage preservation—literature review for selected databases. Heritage Science, 10(1), jan 2022. [1](#)
- [43] C Sommer, L Sang, D Schubert, and D Cremers. Gradient-SDF: A semi-implicit surface representation for 3d reconstruction. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022. [2](#)
- [44] Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. Neural 3D reconstruction in the wild. In SIGGRAPH Conference Proceedings, 2022. [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [12](#), [13](#)
- [45] Matthew Tancik, Vincent Casser, Xinchun Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8248–8258, 2022.
- [46] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH ’23, 2023.
- [47] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. Advances in Neural Information Processing Systems, 33, 2020. [5](#), [12](#)
- [48] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. CVPR, 2022.
- [49] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689, 2021. [2](#), [4](#)
- [50] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023. [2](#)
- [51] Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. IEEE Transactions on Image Processing, 20(5):1185–1198, 2011. [4](#)
- [52] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization, 2023. [9](#)
- [53] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In Thirty-Fifth Conference on Neural Information Processing Systems, 2021. [2](#)
- [54] Kwang Moo Yi. Image matching: Local features and beyond 2020, 2020. Accessed: 2023-11-16. [4](#), [12](#), [13](#)
- [55] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In CVPR, 2021.
- [56] Zehao Yu, Anpei Chen, Bozidar Antic, Songyou Peng, Apratim Bhattacharyya, Michael Niemeyer, Siyu Tang, Torsten Sattler, and Andreas Geiger. Sdfstudio: A unified framework for surface reconstruction, 2022. [2](#), [5](#), [6](#), [12](#), [13](#)
- [57] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. Advances in Neural Information Processing Systems (NeurIPS), 2022. [2](#)
- [58] Lulin Zhang, Ewelina Rupnik, and Marc Pierrot-Deseilligny. Feature matching for multi-epoch historical aerial images. ISPRS Journal of Photogrammetry and Remote Sensing, 182:176–189, dec 2021. [2](#)

## A1. Used Code and Datasets

Tab. A.4 summarizes the code and datasets we use for evaluation and comparison. Our code and recorded datasets will be made publicly available upon acceptance. For the pre-processing of the dataset, *i.e.*, mask out humans and irrelevant objects,

	Name	type	year	link	license
[1]	Metashape	code	2021	<a href="https://www.agisoft.com/">https://www.agisoft.com/</a>	Proprietary
[44]	NeuralRecon-W	code	2022	<a href="https://github.com/zju3dv/NeuralRecon-W">https://github.com/zju3dv/NeuralRecon-W</a>	Apache-2.0
[56]	NeuS-Facto, NeusW	code	2022	<a href="https://docs.nerf.studio/nerfology/methods/nerfacto.html">https://docs.nerf.studio/nerfology/methods/nerfacto.html</a>	Apache-2.0
[40]	COLMAP	code	2016	<a href="https://colmap.github.io/">https://colmap.github.io/</a>	new BSD
[11,23,47]	Hierarchical Localization Toolbox	code	2023	<a href="https://github.com/cvg/Hierarchical-Localization/">https://github.com/cvg/Hierarchical-Localization/</a>	Apache-2.0
[54]	Brandenburg Gate	dataset	2020	<a href="https://www.cs.ubc.ca/~kmyi/imw2020/data.html">https://www.cs.ubc.ca/~kmyi/imw2020/data.html</a>	-
[29]	Historic Building	dataset	2022	<a href="https://www.gw.uni-jena.de/en/faculty/juniorprofessur-fuer-digital-humanities/research/jena4d-stadtgeschichtsbuch">https://www.gw.uni-jena.de/en/faculty/juniorprofessur-fuer-digital-humanities/research/jena4d-stadtgeschichtsbuch</a>	Creative Commons
Ours	National Theater	dataset	2023	will be public upon acceptance	Creative Commons

Table A.4. Used datasets and code in our submission, together with reference, link, and license.

we use the NeuralRecon-W [44] codebase. For NeusW we used the SDF-Studio implementation.

## A2. Mesh Visualization

We show in this section the reconstructed meshes of Brandenburg Gate [54] dataset, corresponding to Tab. 2.

## A3. Rendering Visualization

In this section, we show the rendered view, which is the by-product of our method. We use volumetric rendering to get high-fidelity surface which does not need root-finding when extracting meshes. In Fig. A.11 we show the results with and without color loss on the Brandenburg Gate dataset corresponding to Fig. 9 (the second and the last columns) and Tab. 3. The network can already recover the colors with only one color image.

Fig. A.12 shows the rendered images, depth, and normal maps of our method on the historical datasets corresponding to the last column of Fig. 7. Note that the depth and normal maps are rendered directly without extracting meshes, different from the mesh normal we show in Fig. 7. We are able to recover the color images as well, despite the limited color input images. For the St. Micheal Church dataset, however, the recovered color is less successful compared to the other historical datasets due to the fact that the dataset does not contain any color images. The color is only projected to gray-scale using the perceptual weights without color supervision.

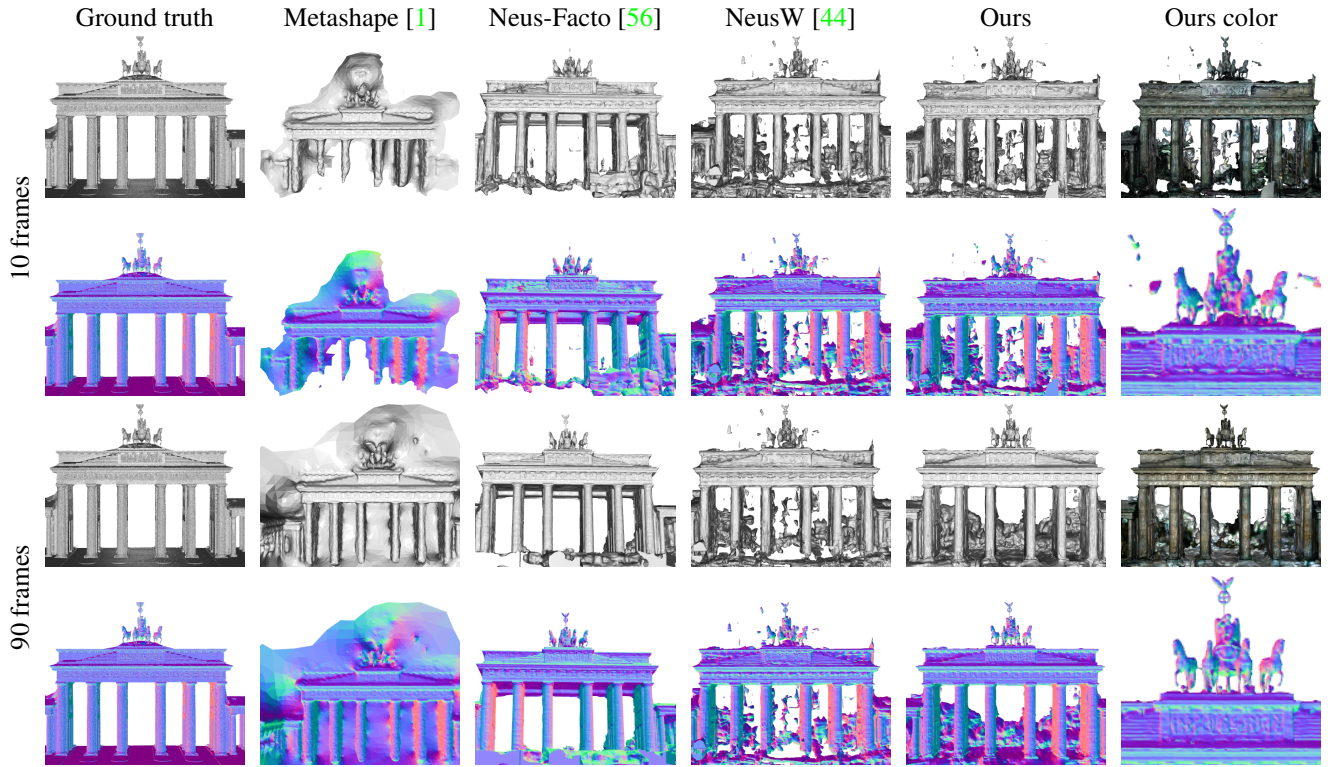


Figure A.10. Reconstructed meshes comparison results for Brandenburg Gate [54] dataset.



Figure A.11. Comparison results of novel view synthesis of Brandenburg Gate for baseline NeusW [44] and our method. With the color appearance embedding loss, we can recover the color images as well.

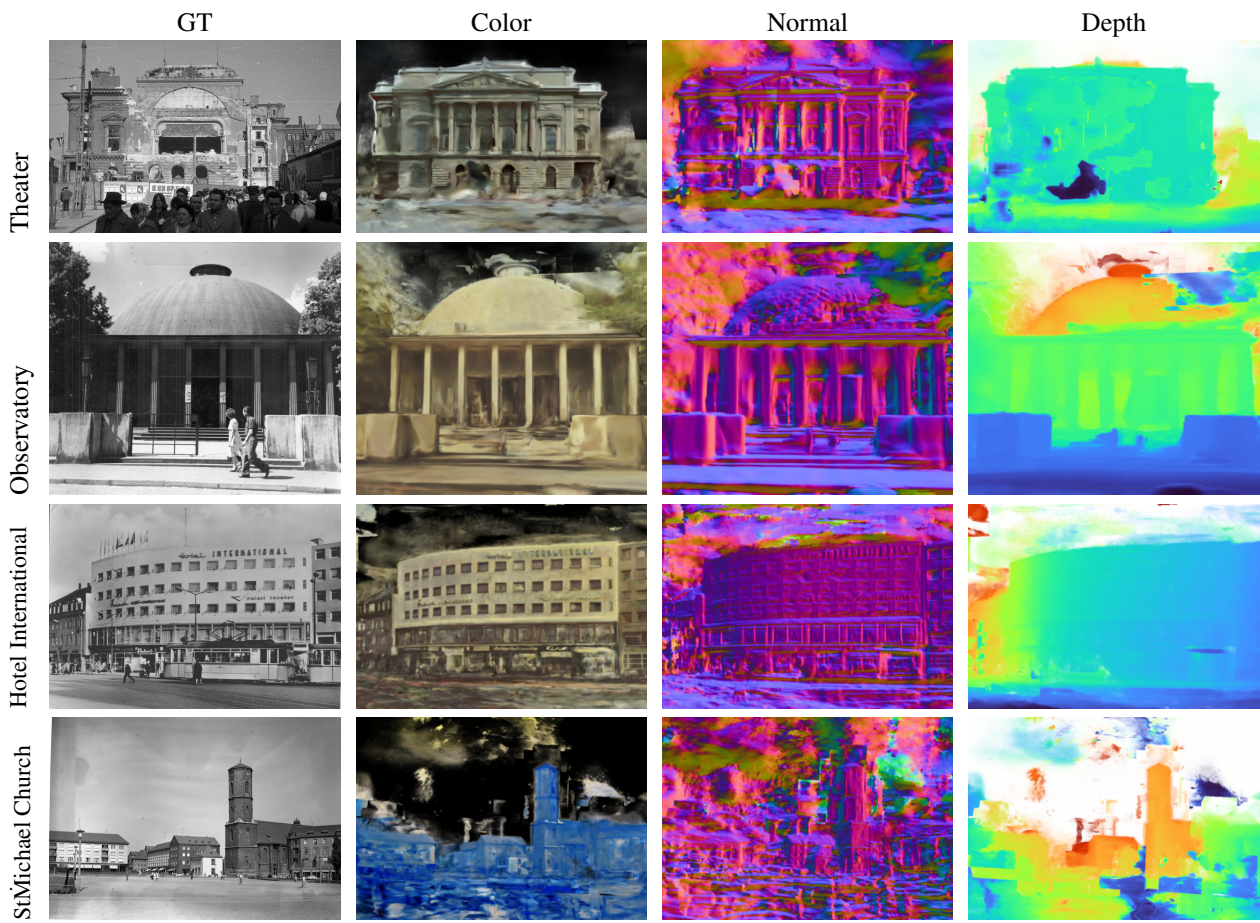


Figure A.12. Novel View Synthesis of historical datasets. The first column shows the ground truth view, the second column the rendered color images, the third and fourth columns are normal and depth maps. The normal maps reflect more details for each ray than normal from meshes. The normal of the meshes are restricted by the Marching Cubes resolutions.