Depth Supervised Neural Surface Reconstruction from Airborne Imagery

V. Hackstein[1]*, P. Fauth-Mayer[1], M. Rothermel[1], N. Haala[2]

[1] nFrames ESRI, Germany - (vhackstein, pfauthmayer, mrothermel)@esri.com
[2] Institute for Photogrammetry and Geoinformatics, University of Stuttgart, Germany - norbert.haala@ifp.uni-stuttgart.de

Keywords: Neural Radiance Fields (NeRF), Multi-View-Stereo, 3D Scene Reconstruction, Meshed 3D Point Cloud, Airborne Imagery, Depth Supervision

Abstract

While originally developed for novel view synthesis, Neural Radiance Fields (NeRFs) have recently emerged as an alternative to multi-view stereo (MVS). Triggered by a manifold of research activities, promising results have been gained especially for texture-less, transparent, and reflecting surfaces, while such scenarios remain challenging for traditional MVS-based approaches. However, most of these investigations focus on close-range scenarios, with studies for airborne scenarios still missing. For this task, NeRFs face potential difficulties at areas of low image redundancy and weak data evidence, as often found in street canyons, facades or building shadows. Furthermore, training such networks is computationally expensive. Thus, the aim of our work is twofold: First, we investigate the applicability of NeRFs for aerial image blocks representing different characteristics like nadir-only, oblique and high-resolution imagery. Second, during these investigations we demonstrate the benefit of integrating depth priors from tie-point measures, which are provided during presupposed Bundle Block Adjustment. Our work is based on the state-of-the-art framework VolSDF, which models 3D scenes by signed distance functions (SDFs), since this is more applicable for surface reconstruction compared to the standard volumetric representation in vanilla NeRFs. For evaluation, the NeRF-based reconstructions are compared to results of a publicly available benchmark dataset for airborne images.

1. Introduction

Image based 3D surface reconstruction is useful for many applications including urban modeling, environmental studies, simulations, robotics, and virtual reality. Typically, this task is solved by matured photogrammetric pipelines. As a first step, a Structure-from-Motion (SfM) approach estimates camera poses for further processing. While this step already provides a sparse reconstruction from tie point measurement as required by the Bundle Block Adjustment, dense 3D point clouds or meshes are generated by a multi-view stereo (MVS) pipeline in the second step. Examples of state-of-the-art approaches developed during the last decade are COLMAP (Schönberger et al., 2016), PMVS (Furukawa and Ponce, 2010) or SURE (Rothermel et al., 2012). Despite the considerable reconstruction quality available from such pipelines, they still suffer from problems at fine geometric structures, texture-less regions, and especially at non-Lambertian surfaces, i.e. at semi-transparent objects or reflections. Also due to these remaining issues, alternative approaches based on Neural Radiance Fields (NeRF) gained considerable attention. Originally, this technique was developed for synthesizing novel views of complex scenes by using a sparse set of input views (Mildenhall et al., 2021). For this purpose, a neural network provides an implicit representation of the surface geometry as well as the appearance of the scene. This representation is then used to generate synthetic views by neural volume rendering. The neural network is trained to minimize the difference between the observed images and the corresponding virtual view of the scene. While NeRFs were originally motivated by visualisation applications, a considerable part of research work meanwhile focuses on 3D reconstruction (Li et al., 2023, Wang et al., 2023). However, the corresponding experiments are typically limited to reconstructions of close range scenes while investigations using aerial imagery are just emerging (Xu et al., 2024). Such aerial applications make specific demands like the timely processing of large areas including the reconstruction at different scales, or challenging scenarios like street canyons, glass facades or shadowed areas. Furthermore, despite impressive results presented so far, NeRF-based surface reconstruction frequently suffers from challenges for low image redundancy and weak data evidence, while the computational effort for training the neural network is still considerable. To mitigate this problem recent works use additional cues for initialization or supervision during training. One option to support dense reconstruction is to integrate *a priori* structural information as provided from the sparse SfM point cloud to this process. Since the volumetric representation of vanilla NeRFs is sub-optimal for surface reconstruction tasks, approaches as (Yariv et al., 2021, Wang et al., 2021) model 3D scenes using signed distance functions (SDFs). By these means, VolSDF combines the advantages of volume rendering methods during training and implicit surface modeling for geometric reconstruction. As our main contribution, we integrate tie point supervision into VolSdf and evaluate its reconstruction capabilities for typical aerial nadir and oblique image blocks.

The remainder of our paper is as follows: Section 2 gives a brief overview on classical MVS and the state-of-the-art on NeRF based reconstruction. Section 3 then presents our approach, which modifies the framework VolSDF (Yariv et al., 2021) to supervise and thus support the training process using SfM tie points. As discussed in section 2 this framework provides an easy accessible representation of 3D model geometry, which is well suited for regularization during training. Section 4 evaluates our pipeline for three aerial image sets featuring different configurations. These investigations on data typically used in professional aerial mapping are interesting from a practical point of view while investigating specific challenges of NeRF-based surface reconstruction. Such aerial image collections feature limited viewing directions and potentially suffer from re-

* Corresponding author

stricted surface observation due to occlusions. Additional challenges are variances in lighting conditions and moving objects or shadows. We show that training by tie-points supervision is crucial for fast convergence and mitigates convergence to local minima during training. This holds in particular true for demanding scenes featuring vegetation or contradictory data e.g moving shadows. For evaluation, the results of our NeRF based reconstruction are analyzed for three data sets, including a comparison to results of a benchmark on high density aerial image matching (Haala, 2013).

2. Related Work

2.1 Classical MVS

Taking a collection of images and their pixel-accurate poses as input, most prominent MVS systems reconstruct dense geometry in form of points or depth maps. Many approaches use stereo or multi-view stereo algorithms to reconstruct depth maps (Galliani et al., 2015, Schönberger et al., 2016, Rothermel et al., 2012). Another prominent line of work starts with a sparse set of points which are iteratively refined and densified (Furukawa and Ponce, 2010, Goesele et al., 2007). In a second step a globally consistent, topologically valid surface is extracted using volumetric reconstruction such as (Kazhdan and Hoppe, 2013, Labatut et al., 2009, Jancosek and Pajdla, 2011, Fuhrmann and Goesele, 2014, Ummenhofer and Brox, 2015). To enhance detail, meshes can be further refined such that photoconsistency across views is maximized (Vu et al., 2012, Delaunoy et al., 2008). Such reconstruction pipelines rely on a sequence of computational expensive optimization algorithms. Moreover, each module has to be carefully tuned with regard to its parameters and quality of input from upstream modules.

2.2 Neural Implicit Representations

The seminal work introducing NeRF (Mildenhall et al., 2020) opened a new research path in the area of novel view synthesis. (Martin-Brualla et al., 2021) robustify vanilla NeRF for imagery with varying illumination conditions. (Barron et al., 2021) show improved rendering quality by employing a training regime mitigating aliasing and accounting for the fact that pixels capture the scene with different ground resolution. Scalability for larger scenes is addressed in (Tancik et al., 2022, Xiangli et al., 2022, Turki et al., 2022). (Li et al., 2022, Fridovich-Keil et al., 2022, Hu et al., 2023) greatly improve training and inference times, by fully or partly replacing the original representation of scene geometry by spatial data structures such as multi-scale hash encoding or 3D MipMaps.

NeRF style approaches target the task of novel view-synthesis. They implicitly represent 3D geometry as density and the light emitted at a specific position in space, which impedes straightforward surface regularization. Instead, methods targeting 3D reconstruction model the geometry by an implicit surface such as occupancy or signed distance functions. This enables the formulation of surface regularization losses and defines a global threshold required to extract surfaces using marching cubes (Lorensen and Cline, 1987). Early work employed surface rendering (Niemeyer et al., 2020b, Yariv et al., 2020). However, geometry and radiance are only optimized near the surface hampering fast convergence for complex scenes. In contrast (Oechsle et al., 2021, Wang et al., 2021, Yariv et al., 2021) implement volume rendering, optimizing geometry and radiance for an extended scene volume eventually approaching surface vicinity. This stabilizes convergence. Training and inference times can be considerably improved by efficient GPU implementation and incorporating multi scale hash encoding (Li et al., 2023, Wang et al., 2023). Inspired by traditional MVS approaches, (Fu et al., 2022, Darmon et al., 2022) introduce losses encouraging multi-view photometric consistency of surface patches. Contrary to neural surface representations Gaussian splatting (Kerbl et al., 2023) represents the scene by splats (3D points, 3D covariances, color and opaqueness). Differentiable rendering of splats can be efficiently implemented on GPUs which allows for impressive rendering times. To reconstruct surfaces from splats (Guédon and Lepetit, 2023) regularize 3D Gaussians and extract meshes by subsequent Poisson reconstruction.

Training of neural implicit representations is challenging for image collections featuring limited surface observations and challenging appearance. Additional depth cues from monocular depth (Yu et al., 2022b) or RGBD-sensors (Azinović et al., 2022) can mitigate this problem. Similar to (Deng et al., 2022) we supervise our reconstructions with SfM tie points. In the domain of remote sensing (Mari et al., 2022, Qu and Deng, 2023) train NeRFs or neural implicit surfaces for satellite imagery. (Turki et al., 2022) propose a NeRF variant for large aerial scenes but focus on novel view synthesis. Most similar to our work, (Xu et al., 2024) provide an performance and quality evaluation of a scaleable MipNerf (Barron et al., 2021) for aerial images. In contrast in this work we investigate implicit neural surface reconstruction from aerial imagery.

3. Methodology

In this section we first review VolSDF (Yariv et al., 2021) as it is the base method for our implementation. We then explain our extension for depth supervision and implemented training schemes.

3.1 Recap of VolSDF

NeRF. A neural radiance field is a neural function that takes spatial coordinates $\mathbf{x}$ and a viewing direction vector $\mathbf{v}$ as input which is mapped to a scalar density σ and a color vector $\mathbf{c}$.

$$F_\Theta : (\mathbf{x}, \mathbf{v}) \rightarrow (\sigma, \mathbf{c}).$$

The function is modeled by two fully connected multilayer perceptrons (MLPs) encoding geometry and appearance respectively. The color $\hat{\mathbf{C}}$ of a pixel in an arbitrary view encoded by F_Θ can be composed using volume rendering. Let $\mathbf{r}$ (with direction $\mathbf{v}$) be the ray defined by the center of projection of a view and the pixel coordinate. We sample N points $x_i, i \in [0, N]$ along $\mathbf{r}$, the distances between point samples are given by $\delta_i, i \in [0, N-1]$. Using the quadrature based on the rectangle rule (Max, 1995), the discrete formulation of volume rendering is given by

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} T_i o_i \mathbf{c}_i. \tag{1}$$

Thereby

$$o_i = 1 - \exp\left(-\sigma_i \delta_i\right) \tag{2}$$

is a notion of the emitted light or opacity.

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \tag{3}$$

represents the accumulated transparency along the ray up to the current position r_i. Thus, light emitted for samples $j < i$ reduces the contribution of colored light emitted for sample i in the rendering equation 1.

Since equation 1 is differentiable, a loss minimizing photo-consistency across all views can be specified by the projection error:

$$\mathcal{L}_{RGB} = \|\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}\|_2. \tag{4}$$

VolSDF - SDF representation. In contrast to NeRF and variants, VolSDF models geometry by a signed distance function which only in a subsequent step is mapped to density. With Ω the 3D space occupied by some object and $\mathcal{M}$ the object boundary, the signed distance function can be formally written by

$$d_\Omega = (-1)^{\mathbf{1}_\Omega(\mathbf{x})} \min_{\mathbf{y}\in\mathcal{M}} \|\mathbf{x} - \mathbf{y}\| \tag{5}$$

with

$$\mathbf{1}_\Omega(\mathbf{x}) = \begin{cases} 1 \text{ if } \mathbf{x} \in \Omega \\ 0 \text{ if } \mathbf{x} \notin \Omega \end{cases}. \tag{6}$$

To be able to employ volume rendering the signed distance is mapped to density with two learnable parameters β and α

$$\sigma(\mathbf{x}) = \alpha \Psi_\beta(-d_\Omega(\mathbf{x})) \tag{7}$$

with

$$\Psi_\beta(s) = \begin{cases} \frac{1}{2}\exp\left(\frac{s}{\beta}\right) & \text{if } s \leq 0 \\ 1 - \frac{1}{2}\exp\left(-\frac{s}{\beta}\right) & \text{if } s > 0. \end{cases} \tag{8}$$

Surface points $\mathbf{x} \in \mathcal{M}$ have a constant density of $\frac{1}{2}\alpha$. Density smoothly decreases for increased distances from the surface. This smoothness is controlled by β. As β approaches 0, Ψ will converge to a step function, and σ will converge to a scaled indicator function that maps all points inside the object to α, and all other points to 0. More intuitively, β can be interpreted as a parameter encoding the confidence in the SDF in the current training stage. When the confidence is low, or equivalent, β is high, points distant from the surface will map to larger densities and contribute to optimization. In later stages when the confidence is higher, only points close to the surface will contribute to optimization. Similar to (Yariv et al., 2021), α is set to $\frac{1}{\beta}$ in all our experiments.

VolSDF - Regularization. The SDF representation allows for regularization of the surface to cope with weak or contradictory image data. Similar to (Gropp et al., 2020) we use the eikonal loss

$$\mathcal{L}_{eik} = (\|\nabla d(\mathbf{x})\| - 1)^2. \tag{9}$$

to enforce smoothness of the signed distance field. Additionally we found that including a prior enforcing consistency of normals (Oechsle et al., 2021) within an increased local neighborhood $\Delta\mathbf{x}$ benefits reconstruction

$$\mathcal{L}_{surf} = \left\|\mathbf{n}(\mathbf{x}) - \mathbf{n}(\mathbf{x} + \Delta\mathbf{x})\right\|_2. \tag{10}$$

The normal $\mathbf{n}$ is the gradient of the signed distance field and can be computed using double backpropagation (Niemeyer et al., 2020a).

VolSDF - Sampling. As in all NeRF-style methods a sampling strategy, ideally sampling points close to the correct surface but at the same time being able to recover/converge from inaccurate states of the NeRF is crucial. VolSDF ultimately places samples based on *inverse transform sampling* of the discrete opacity function $o(i)$. The accuracy of $o(i)$ is influenced by the spatial extent of the ray where the samples are placed. Furthermore, for low sampling densities an approximation error is introduced by quadrature. VolSDF implements an iterative sampling mechanism which (a) bounds the approximation error of $o(i)$ and (b) adapts the sampling extent being closer to the surface with increased confidence of the SDF estimation. For details the reader is referred to the original publication.

3.2 Tie Point Supervision

The input of VolSDF are poses which are computed within SfM or aerial triangulation. As a side product j tie points, each encoding homologous 2D image locations $\mathbf{x}_{i,j}$ across i images and their corresponding 3D point $\mathbf{X}_j$, are generated. For each $\mathbf{x}_{i,j}$ a depth $d_{i,j}$ can be computed by projection. Similarly to (Deng et al., 2022) we use tie points to initialize and supervise the training of VolSDF. More specifically, we follow (Azinović et al., 2022) and sample two set of depths $\mathcal{S}^{tr}$ and $\mathcal{S}^{fs}$ along rays induced by $\mathbf{x}_{i,j}$. $\mathcal{S}^{tr}$ contains samples d_s close to the surface, $|d_{i,j} - d_s| < tr$. $\mathcal{S}^{fs}$ contains samples between the camera center and the surface point $d_{i,j}$ with $d_s \in \{0, d_{i,j} - tr\}$. A first loss enforces the predicted SDF $\hat{d}_s$ to correspond with d_s

$$\mathcal{L}_{tr} = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} \frac{1}{|\mathcal{S}^{tr}|} \sum_{s\in\mathcal{S}^{tr}} \left(d_s - \hat{d}_s\right)^2, \tag{11}$$

where $\mathcal{R}$ is a set of randomly sampled rays over all input images per training batch. (Azinović et al., 2022) encourage a constant SDF value tr in freespace. In contrast we relax that constraint and define a loss which only enforces SDF values larger than t_r

$$\mathcal{L}_{fs} = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}}^{N} \frac{1}{|\mathcal{S}^{fs}|} \sum_{s\in\mathcal{S}^{fs}} ReLU^2\left(tr - \hat{d}_s\right). \tag{12}$$

We note the d_s is only an approximation of the signed distance and rigorously valid in front-to-parallel settings, however we do not introduce any bias on the zero level set. For all experiments we set tr to 30 times the GSD to safely exceed the noise levels.

3.3 Implementation and Training Details

Our model builds on the VolSDF implementation (Yu et al., 2022a). It is composed of learnable multi-resolution hash grid encoding (Müller et al., 2022) and two MLPs with two layers of 256 neurons each. We set the leaf size of the grid to match the GSD of each evaluated dataset. Our final training loss consists of five terms:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{RGB} + \lambda_{eik}\mathcal{L}_{eik} + \lambda_{surf}\mathcal{L}_{surf} \\ +\lambda_{fs}\mathcal{L}_{fs} + \lambda_{tr}\mathcal{L}_{tr},\end{aligned} \tag{13}$$

where λ_* are hyperparameters controlling the contribution of the respective loss terms. Their values were found by grid

search, are constant in all experiments and listed in table 1. The network is optimized using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $lr = 5\text{e}{-4}$ and exponential decay with rate 0.1. The batch size remains constant and is set to 4096 rays per training iteration. Training and inferences were run on an AMD Ryzen 3960X 24-Core CPU and a Nvidia RTX 3090.

Parameter	Symbol	Values	
		1st stage	2nd stage
Eikonal factor	λ_{eik}	0	5e-4
Surface smooth. factor	λ_{surf}	1e-2	5e-3
Surface smooth. radius	R_{surf}	35 GSD	35 GSD
Free-space factor	λ_{fs}	10	10
Signed-distance factor	λ_{tr}	60	60
Initial β	β_0	0.001	0.001

Table 1. Training parameters.

The training is split into two stages: a fast geometric initialization with depth supervision and smoothness regularization only, and a second stage additionally activating photometric supervision. The duration of the first stage is 1 k epochs for all experiments. We evaluate results after training for 30 k and 100 k epochs.

4. Evaluation

4.1 Datasets

We qualitatively evaluate our method on three datasets. These include two image blocks captured by professional large-format cameras in nadir only (*Frauenkirche*) and oblique (*Domkirk*) configuration. Images of *Frauenkirche* are part of a benchmark on high density aerial image matching (Haala, 2013). Furthermore, we run tests on precisely georeferenced, high-resolution UAV images provided within a recent benchmark *Hessigheim 3D* (Kölle et al., 2021, Haala et al., 2022). More details can be found in table 2. The image collections cover challenging urban scenes, including thin structures, low data evidence (e.g. limited views, occlusions), photometric inconsistency and ambiguity (e.g. moving objects, shadows, and vegetation). We generate tie points for each dataset using a commercial AT software (Esri, 2023).

4.2 Qualitative Evaluation

In a warm-up stage we train the network with depth supervision and smoothness regularization only. These models (figure 1) serve as geometric initialization for the main training stage with photometric supervision enabled. The warm-up optimization typically converges within 1 k epochs, which equals approximately 6 minutes of training on our hardware. We found the parameters in table 1 to be a good balance between detail provided by tie points and completeness of reconstructed surfaces. Despite the sparsity of tie points, completeness of reconstructions is rather impressive.

Figures 2, 3 and 4 display the results for VolSDF with depth prior after 30 k epochs (first column) and vanilla VolSDF after 100 k epochs (second column) for all evaluated datasets. We found that additional training improves details but does not fix erroneous topology of extracted surfaces. The boxes in figures 2, 3 and 4 highlight areas of sub-optimal reconstruction (red) and improvements (black) achieved by using depth supervised VolSDF.

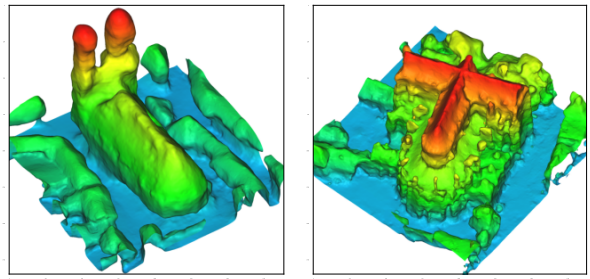

Figure 1. Intermediate reconstructions based on depth and smoothness supervision used as geometric initialization for the main training stage.

Without depth prior VolSDF has difficulties reconstructing complete surfaces when trained on *Frauenkirche*. The ground level is incomplete, presumably due to (moving) shadows and weak texture. Furthermore, VolSDF fails in the reconstruction of facades (figure 2, boxes 4, 5 and 7). We assume that this is caused by the combination of limited number of observations and repetitive structure. Topology significantly improves when using depth-supervised VolSDF. The geometric initialization ensures more complete ground surface. Despite the limited number of tie points on facades, depth supervision improves completeness (figure 2, boxes 2 and 3). In areas where no tie points are computed, no improvements can be observed (box 6).

For the *Domkirk* dataset impressive detail is reconstructed for both approaches (figure 3, box 1 and 2). Again VolSDF gets stuck in local minima in weakly observed and low-texture areas (box 6 and 5). In both cases depth supervision facilitates reconstruction of a more correct surface (box 2 and 3).

Similar to *Frauenkirche* depth supervised VolSDF delivers more complete reconstructions for the *Hessigheim 3D* scene. Furthermore, VolSDF struggles to reconstruct areas for which appearance across views is dissimilar or contradictory, e.g vegetation (figure 4, box 2). For such areas the depth prior resolves ambiguities and constrains the optimization resulting in more faithful surfaces. We note our approach seems rather robust to imprecise or outlier contaminated tie points and only in rare cases generates artifacts as spikes (box 1).

4.3 Quantitative Evaluation

The number of publically available benchmarks for aerial surface reconstruction is very limited. The main challenge is to collect precise and geo-referenced 3D ground truth data featuring sufficient density to evaluate reconstruction quality of sharp depth discontinuities and small details. We base our quantitative evaluation on DSM raster data of the *Frauenkirche* scene. As ground truth we use the benchmark DSM provided by (Haala, 2013). This DSM was generated by robust fusion of eight DSMs computed by eight independent reconstruction pipelines across academia and industry.

Error metrics. We evaluate the correctness of the reconstruction in terms of *accuracy* and *completeness* as defined in the ETH3D benchmark (Schöps et al., 2017). For both metrics, we evaluate differences between corresponding height values of ground truth DSM and a DSM derived from our reconstruction. For the latter we generate point clouds on the SDF zero level set and rasterize the highest surface points in the area of

Dataset	GSD	Configuration	Images	Pixels	Dimensions
Frauenkirche	10 cm	Aerial Nadir	35	86 MPix	95^3 m^3
Domkirk	3 cm	Aerial Oblique	226	1362 MPix	42^3 m^3
Hessigheim 3D	1.5 cm	UAV	217	1500 MPix	27^3 m^3

Table 2. Datasets used for evaluation.

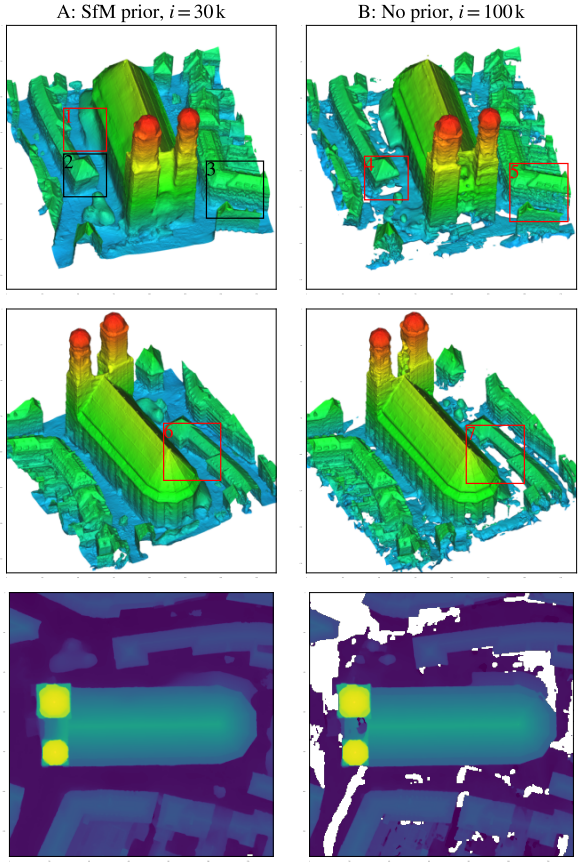

Figure 2. Results after 30 k (left) and 100 k (right) training epochs for the *Frauenkirche* dataset using the sparse, depth priors from tie points (left) or only RGB input (right) for training. The first two rows display extracted meshes from different viewpoints. Row three shows extracted DSMs.

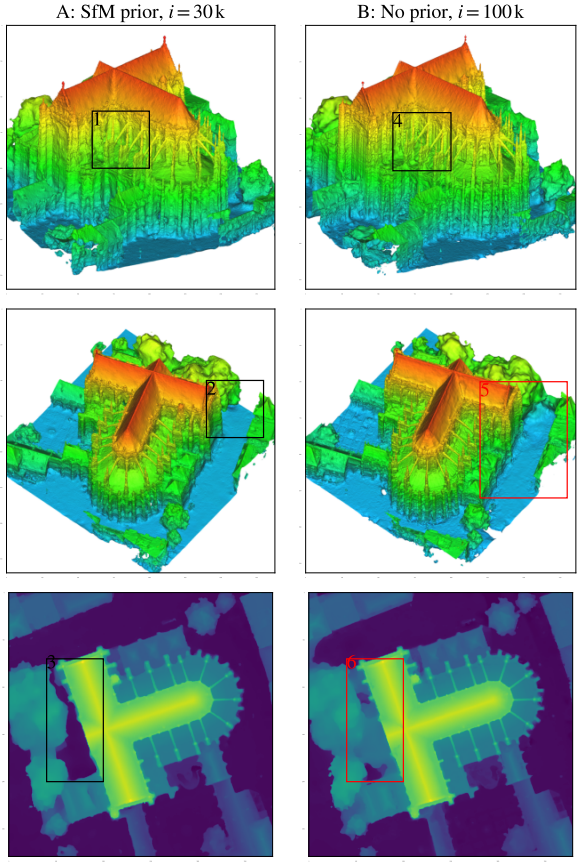

Figure 3. Results after 30 k (left) and 100 k (right) training epochs for the *Domkirk* dataset using the sparse, depth priors from tie points (left) or only RGB input (right) for training. The first two rows display extracted meshes from different viewpoints. Row three shows extracted DSMs.

interest. Both metrics are evaluated over a range of tolerance thresholds, ranging from 1 GSD to 30 GSD.

Additionally, we measure the noise of our surfaces in a robust fashion. The *NMAD* metric measures the similarity of the prediction and the ground truth within the error band, i.e. is an indicator of the amount of noise within the tolerance. More specifically, NMAD is the Normalized Median Absolute Deviation and is a robust estimator for the standard deviation in normally distributed data (Rousseeuw and Croux, 1993).

Reconstruction Quality. Figures 5, 6, 7 and 8 display metric scores of the models trained on *Frauenkirche* data. We show scores for VolSDF and the depth-supervised variant after 30 k and 100 k epochs.

Figure 6 displays the *accuracy* and *completeness* scores. In terms of *completeness*, our model trained for 30 k epochs significantly outperforms the reference model trained for 100 k epochs. The respective scores converge with a difference of around 10 %. This validates the observation that local minima and incompleteness are mitigated in early training stages already. The reconstruction *accuracy* of our model is higher than that of the reference model throughout the entire tolerance interval when both are evaluated after 30 k epochs, verifying accelerated convergence. Furthermore, our model after 30 k iterations is almost on par with the reference trained for 100 k epochs, which only delivers a slightly better score for lower GSD ranges. Our model trained for 100 k epochs achieves the best *accuracy* throughout the entire evaluation range. Figure 7 visualizes the signed differences between the reconstructed and ground truth DSMs for VolSDF with depth prior (left column) to vanilla SDF (right column). After 30 k epochs of training we observe improved quality of the depth supervised variant (A and B), in particular for the streets around the building. Furthermore, additional training further improves quality of both solutions although the progress is rather slow. We note that very inaccurate areas are not improved even beyond 100 k epochs.

Accelerated convergence for depth supervision can also be ob-

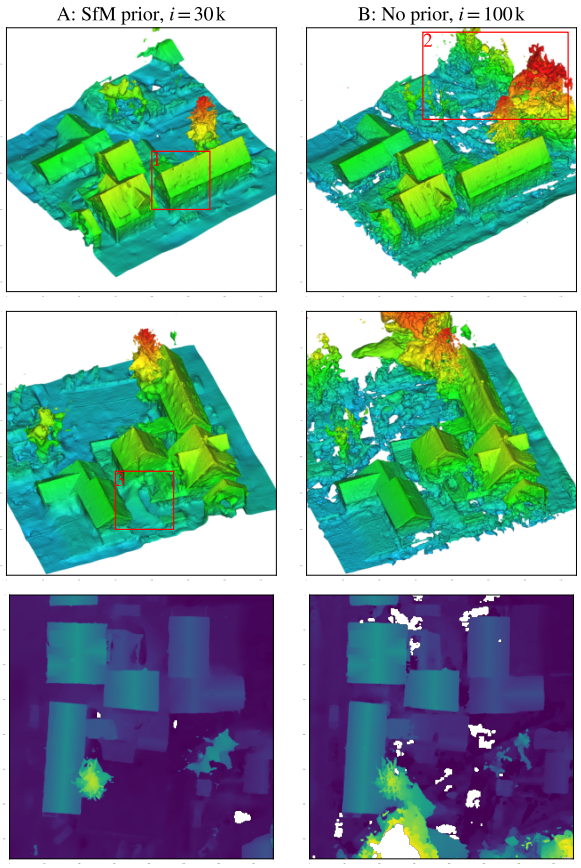

Figure 4. Results after 30 k (left) and 100 k (right) training epochs for the *Hessigheim 3D* dataset using the sparse, depth priors from tie points (left) or only RGB input (right) for training. The first two rows display extracted meshes from different viewpoints. Row three shows extracted DSMs.

served in the loss curves over training time (figure 5). Right after initialization, the loss of the depth-guided model drops significantly faster compared to the baseline. After the first hour of training, however, the values of both curves are only slowly decreasing. This underlines the observation that in early training stages the depth priors rapidly guide the reconstruction to a faithfull solution. In later training stages, details are refined which for both approaches still demands considerable computation time.

Depth-supervised VolSDF outperforms vanilla VolSDF in terms of *NMAD* scores (figure 8). Notably even after 30 k iteations the *NMAD* is slightly better than training its counterpart for 100 k iterations. After 100 k iterations we achieve a *NMAD* score below 3 GSD.

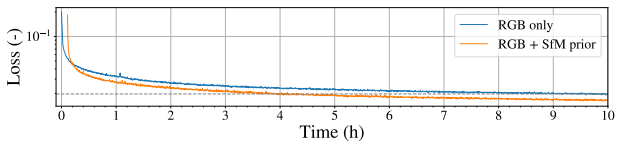

Figure 5. Loss curves for training based on *Frauenkirche* data using RGB input only (blue), or additional depth priors (orange) from SfM/AT in form of tie points (TPs). The grey, dashed line shows the loss after 10 hours of training when using RGB only.

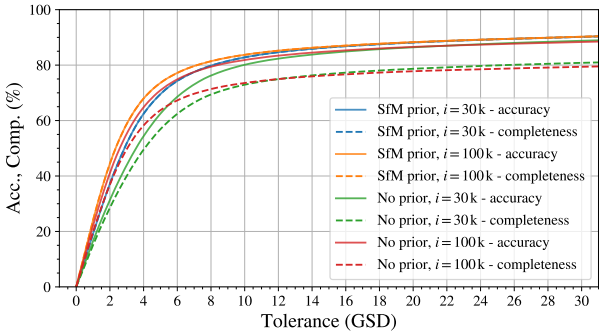

Figure 6. *Accuracy* (solid line, left) and *completeness* (dashed line, left) scores for the *Frauenkirche* dataset using the sparse, depth priors from tie points for training.

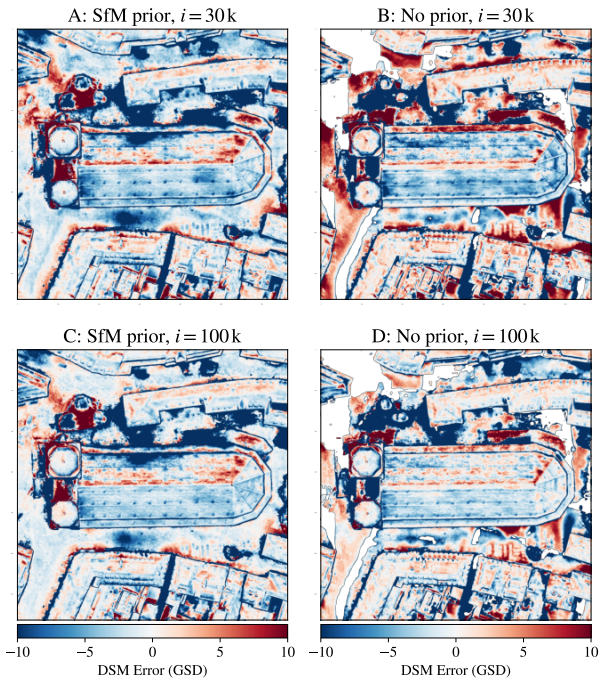

Figure 7. Differences between the extracted DSMs and the GT DSM, where the colors saturate outside of ±10 GSD.

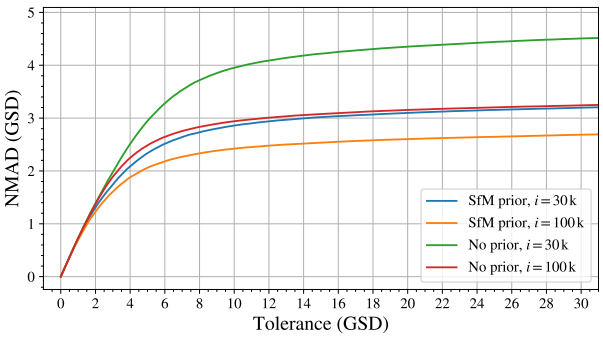

Figure 8. *NMAD* scores for the *Frauenkirche* dataset using the sparse, depth priors from tie points for training.

Processing Time. The use of tie point priors improves convergence in the early training stages, thus decreases runtimes. We note that it is difficult to rigorously compare training times across the approaches since surface states as well as final solutions differ. However, for the majority of scene parts comparable visual reconstruction quality of vanilla VolSDF and its depth supervised variant is obtained after 100k iterations vs 30k iterations respectively. For both approaches, training for 30 k and 100 k iterations takes about 3:20 h and 11 h respectively.

Furthermore, we profiled the algorithm to identify most computationally expensive routines. Figure 9 shows the relative time requirements of the different model components within one training epoch. The additional depth-supervision terms do not generate noticeable computational overhead. The main bottleneck is VolSDF's sampling routine accounting for 70 % of training time, which suggests future optimization.

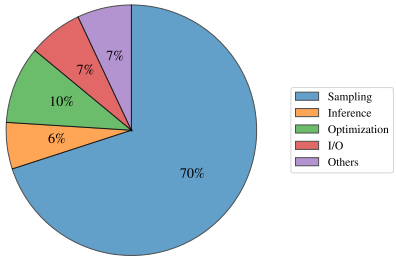

Figure 9. Relative time requirements of different model components within one epoch. The sampling algorithm accounts for 70 % of training time.

5. Conclusion

We present the applicability of VolSDF, a NeRF variant modeling implicit neural surfaces, for 3D reconstruction from airborne imagery. We demonstrated that supervising VolSFD by tie points improves reconstructions: we observed faster convergence in early training stages and better quality in terms of completeness and accuracy. This is in particular true for challenging areas featuring only limited data evidence for which VolSDF tends to get stuck in local minima or does not converge at all. Reconstructed surfaces of an example nadir scene featured less than 4 GSD deviations to traditional MVS pipelines in terms of *NMAD*. To completely converge and recover full detail prolonged training times are still required. This hampers practical application. However, we obtain topologically correct surfaces in reasonable time which could be subject to subsequent mesh post-processing. Sampling routines are the main bottle neck in the evaluated implementation and subject to future work. On the one hand efficient GPU implementation could speed up this process (Wang et al., 2023), on the other hand we want to investigate possibilities to dynamically reinforce sampling in areas with a large potential for improvements (Kerbl et al., 2023). Neural implicit surface reconstruction is still an active research topic and we hope that this article encourages future work also in the domain of geometric reconstruction from aerial imagery and other remote sensing applications.

References

Azinović, D., Martin-Brualla, R., Goldman, D. B., Nießner, M., Thies, J., 2022. Neural rgb-d surface reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P. P., 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. *ICCV*.

Darmon, F., Bascle, B., Devaux, J.-C., Monasse, P., Aubry, M., 2022. Improving neural implicit surfaces geometry with patch warping. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Delaunoy, A., Prados, E., Piracés, P. G. I., Pons, J.-P., Sturm, P., 2008. Minimizing the multi-view stereo reprojection error for triangular surface meshes. *BMVC 2008-British Machine Vision Conference*, BMVA, 1–10.

Deng, K., Liu, A., Zhu, J.-Y., Ramanan, D., 2022. Depth-supervised NeRF: Fewer views and faster training for free. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Esri, 2023. Arcgis reality studio, version 2023.1. https://www.esri.com/en-us/arcgis/products/arcgis-reality/products/arcgis-reality-studio (13 June 2023).

Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A., 2022. Plenoxels: Radiance fields without neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fu, Q., Xu, Q., Ong, Y.-S., Tao, W., 2022. Geo-Neus: Geometry-Consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*.

Fuhrmann, S., Goesele, M., 2014. Floating scale surface reconstruction. *ACM Transactions on Graphics (ToG)*, 33(4), 46.

Furukawa, Y., Ponce, J., 2010. Accurate, Dense, and Robust Multi-View Stereopsis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(8), 1362–1376.

Galliani, S., Lasinger, K., Schindler, K., 2015. Massively parallel multiview stereopsis by surface normal diffusion. *2015 IEEE International Conference on Computer Vision (ICCV)*.

Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S. M., 2007. Multi-view stereo for community photo collections. 2275–2290.

Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y., 2020. Implicit geometric regularization for learning shapes. *Proceedings of Machine Learning and Systems*, 3569–3579.

Guédon, A., Lepetit, V., 2023. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering.

Haala, N., 2013. The landscape of dense image matching algorithms. *Photogrammetric Week '13*.

Haala, N., Kölle, M., Cramer, M., Laupheimer, D., Zimmermann, F., 2022. Hybrid georeferencing of images and LiDAR data for UAV-based point cloud collection at millimetre accuracy. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 4, 100014.

Hu, W., Wang, Y., Ma, L., Yang, B., Gao, L., Liu, X., Ma, Y., 2023. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields.

Jancosek, M., Pajdla, T., 2011. Multi-view reconstruction preserving weakly-supported surfaces. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 3121–3128.

Kazhdan, M., Hoppe, H., 2013. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3), 29.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 1–14. https://inria.hal.science/hal-04088161.

Kingma, D. P., Ba, J., 2015. Adam: A method for stochastic optimization. Y. Bengio, Y. LeCun (eds), *3rd International Conference on Learning Representations ICLR*.

Kölle, M., Laupheimer, D., Schmohl, S., Haala, N., Rottensteiner, F., Wegner, J. D., Ledoux, H., 2021. The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3D point clouds and textured meshes from UAV LiDAR and Multi-View-Stereo. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 1.

Labatut, P., Pons, J.-P., Keriven, R., 2009. Robust and efficient surface reconstruction from range data. *Computer graphics forum*, Wiley Online Library, 2275–2290.

Li, C., Wu, B., Pumarola, A., Zhang, P., Lin, Y., Vajda, P., 2022. Ingeo: Accelerating instant neural scene reconstruction with noisy geometry priors.

Li, Z., Müller, T., Evans, A., Taylor, R. H., Unberath, M., Liu, M.-Y., Lin, C.-H., 2023. Neuralangelo: High-fidelity neural surface reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lorensen, W. E., Cline, H. E., 1987. Marching cubes: A high resolution 3d surface construction algorithm. *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, Association for Computing Machinery, New York, NY, USA, 163–169.

Mari, R., Facciolo, G., Ehret, T., 2022. Sat-NeRF: Learning Multi-View Satellite Photogrammetry With Transient Objects and Shadow Modeling Using RPC Cameras. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, New Orleans, LA, USA, 1310–1320.

Martin-Brualla, R., Radwan, N., Sajjadi, M. S. M., Barron, J. T., Dosovitskiy, A., Duckworth, D., 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections.

Max, N., 1995. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2), 99-108.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. *ECCV*.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM*, 65(1), 99–106. https://doi.org/10.1145/3503250.

Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 41(4), 102:1–102:15.

Niemeyer, M., Lars, M., Michael, O., Geiger, A., 2020a. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3504–3515.

Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A., 2020b. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oechsle, M., Peng, S., Geiger, A., 2021. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Qu, Y., Deng, F., 2023. Sat-Mesh: Learning Neural Implicit Surfaces for Multi-View Satellite Reconstruction. *Remote Sensing*, 15, 4297.

Rothermel, M., Wenzel, K., Fritsch, D., Haala, N., 2012. Sure: Photogrammetric surface reconstruction from imagery. *In Proceedings LC3D Workshop*.

Rousseeuw, P., Croux, C., 1993. Alternatives to Median Absolute Deviation. *Journal of the American Statistical Association*, 88, 1273 - 1283.

Schönberger, J. L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo. B. Leibe, J. Matas, N. Sebe, M. Welling (eds), *Computer Vision – ECCV 2016*, Springer International Publishing, Cham, 501–518.

Schöps, T., Schönberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A., 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P., Barron, J. T., Kretzschmar, H., 2022. Block-NeRF: Scalable Large Scene Neural View Synthesis. *arXiv:2202.05263 [cs.CV]*.

Turki, H., Ramanan, D., Satyanarayanan, M., 2022. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12922–12931.

Ummenhofer, B., Brox, T., 2015. Global, dense multiscale reconstruction for a billion points. *Proceedings of the IEEE International Conference on Computer Vision*, 1341–1349.

Vu, H.-H., Labatut, P., Pons, J.-P., Keriven, R., 2012. High accuracy and visibility-consistent dense multiview stereo. *IEEE transactions on pattern analysis and machine intelligence*, 34(5), 889–901.

Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W., 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *NeurIPS*.

Wang, Y., Han, Q., Habermann, M., Daniilidis, K., Theobalt, C., Liu, L., 2023. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., Dai, B., Lin, D., 2022. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. *European conference on computer vision*, Springer, 106–122.

Xu, N., Qin, R., Huang, D., Remondino, F., 2024. Multi-tiling neural radiance field (nerf) – geometric assessment on large-scale aerial datasets.

Yariv, L., Gu, J., Kasten, Y., Lipman, Y., 2021. Volume rendering of neural implicit surfaces. *Thirty-Fifth Conference on Neural Information Processing Systems*.

Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., Lipman, Y., 2020. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. *Advances in Neural Information Processing Systems*, 33.

Yu, Z., Chen, A., Antic, B., Peng, S., Bhattacharyya, A., Niemeyer, M., Tang, S., Sattler, T., Geiger, A., 2022a. Sdfstudio: A unified framework for surface reconstruction.

Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A., 2022b. MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*.