Density-based Geometric Convergence of NeRFs at Training Time: Insights from Spatio-temporal Discretization

Dennis Haitz[1]*, Berk Kıvılcım[1], Markus Ulrich[1], Martin Weinmann[1], Michael Weinmann[2]

[1] Karlsruhe Institute of Technology, Institute of Photogrammetry and Remote Sensing, Germany - (dennis.haitz, markus.ulrich, martin.weinmann)@kit.edu, berk.kivilcim@student.kit.edu
[2] Delft University of Technology, Department of Intelligent Systems, The Netherlands - m.weinmann@tudelft.nl

Technical Commission II

Keywords: Neural Radiance Fields, Density Fields, 3D Reconstruction, Multi-view Stereo, Spatio-temporal Analysis

Abstract

Whereas emerging learning-based scene representations are predominantly evaluated based on image quality metrics such as PSNR, SSIM or LPIPS, only a few investigations focus on the evaluation of geometric accuracy of the underlying model. In contrast to only demonstrating the geometric deviations of models for the fully optimized scene model, our work aims at investigating the geometric convergence behavior during the optimization. For this purpose, we analyze the geometric convergence of discretized density fields by leveraging respectively derived point cloud representations for different training steps during the optimization of the scene representation and their comparison based on established point cloud metrics, thereby allowing insights regarding which scene parts are already represented well within the scene representation at a certain time during the optimization. By demonstrating that certain regions reach convergence earlier than other regions in the scene, we provide the motivation regarding future developments on locally-guided optimization approaches to shift the computational burden to the adjustment of regions that still need to converge while leaving converged regions unchanged which might help to further reduce training time and improve the achieved quality.

1. Introduction

Based on combining principles of machine learning, image formation and image synthesis, recently emerging scene representations such as Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) or 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) are designed to optimize a scene representation in terms of weights of a neural network (in the case of NeRFs) or primitives like 3D Gaussians (in the case of 3DGS) in a supervised manner to match its predicted scene appearance under certain views to the respectively observed input image data for the corresponding view configurations. Whereas 3DGS has been demonstrated to allow efficient, high-quality scene rendering for novel viewpoints, it faces limitations regarding compact scene representation and accurate surface representation due to the large number of Gaussians needed to accurately represent a scene as well as the lacking capability of Gaussians to represent certain geometric characteristics of scenes such as flat surfaces or sharp edges. In contrast, the use of a neural network within NeRFs to predict local density and view-dependent color densely in the scene volume allows focusing the weights to better approximate complex aspects of the scene in a compact scene representation, thereby allowing the derivation of dense point clouds that capture complex scene features. Unfortunately, training NeRFs comes with long training times ranging from hours to days as well as long inference times for novel view synthesis. Therefore, several subsequent works focused on accelerating the training of NeRFs (Deng et al., 2022; Sun et al., 2022; Chen et al., 2022; Fridovich-Keil et al., 2022), culminating in approaches with training times in the order of seconds to minutes as proposed with the Instant-NGP (iNGP) approach (Müller et al., 2022). Methods like Nerfacto (Tancik et al., 2023) and Zip-NeRF (Barron et al., 2023) incorporate techniques from iNGP for training time efficiency, however, additionally tackle problems like reducing aliasing.

Whereas local density gives an indication of the local presence of matter and the respective transparency of that volumetric point and, hence, of the geometric structure in a scene, its combination with view-dependent color information determines view-dependent scene appearance. Whereas scene representation and rendering approaches are typically evaluated based on image-based metrics such as PSNR, SSIM or LPIPS, only a few approaches focus on evaluations of the geometric accuracy of the underlying model (Remondino et al., 2023; Haitz et al., 2024). When focusing on the inference of geometric scene structures, we have to take into account that the density information may be blurred and, in turn, a decision boundary needs to be implemented for determining where a ray intersects with present objects. Since this information can be beneficial for training, proposal networks can be trained within an end-to-end approach for receiving potential surface intersections (Barron et al., 2022; Tancik et al., 2023). In order to extract a point cloud from a trained NeRF, rays from known poses can then be constructed for all possible pixel positions of the corresponding image with its camera intrinsics. Depth rendering is executed by integrating the density, reformulated as weights, of the ray up to the position where the accumulated weights reach a certain value, which is set to 0.5 in Nerfstudio (Tancik et al., 2023). In order to get insights on further reducing training time and improving the achieved quality, we investigate the geometric convergence behavior of discretized density fields to analyze which scene characteristics are already represented well within the scene representation at a certain time during the optimization. The discretization is applied two-fold: (i) Inferring point clouds as explicit geometric scene represent-

* Corresponding author

ations from the NeRF model, and (ii) inferring point clouds at different training iteration steps s throughout training, whereby the steps s follow a logarithmic distribution over a range of 30k steps. We refer to training iteration steps as only steps throughout the rest of this work. Through a quantitative analysis in the form of point cloud comparisons in the course of the training process based on widely used point cloud distance metrics as well as a qualitative analysis in terms of visualizations of point clouds and respective deviations at different steps s we provide a study for the geometric convergence behavior of NeRFs. In particular, we demonstrate that certain regions reach convergence earlier than other regions in the scene, thereby motivating the future development of locally-guided optimization approaches. By this, the overall training time can be reduced with respect to a certain target metric or the overall quality can be improved.

2. Related Work

In recent years, neural scene representation and rendering techniques have gained a lot of attention for scene modeling and novel view synthesis (Tewari et al., 2020; Tewari et al., 2022). In particular, Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) and respective extensions (Tewari et al., 2022) have been demonstrated to offer a high potential for accurate scene representation. Major extensions focus on improving rendering quality and, hence, also model quality by reducing aliasing (Barron et al., 2021; Wang et al., 2022; Barron et al., 2022; Barron et al., 2023) as well as the acceleration of the training of the underlying network (Müller et al., 2022; Chen et al., 2022), enabling the inference of scene models within seconds (Müller et al., 2022). Further approaches focused on improving robustness to inconsistencies in the input data (Sabour et al., 2023; Buschmann et al., 2025), handling photo collections taken *in-the-wild* (Martin-Brualla et al., 2021), handling large-scale scenarios (Tancik et al., 2022; Turki et al., 2022; Xiangli et al., 2022; Mi and Xu, 2023; Xie et al., 2023; Xu et al., 2024; Chen et al., 2024), real-time training with fast point cloud extraction (Haitz et al., 2023) and the refinement or complete estimation of camera pose parameters for the input images (Yen-Chen et al., 2021; Wang et al., 2021b; Lin et al., 2021; Jeong et al., 2021; Bian et al., 2023; Chen and Lee, 2023). Whereas conventional photogrammetric techniques still outperform NeRF and respective variants in case of well-textured and partially textured objects (Remondino et al., 2023; Hillemann et al., 2024), NeRF approaches outperform conventional approaches for challenging scenarios (Condorelli et al., 2021; Balloni et al., 2023; Pepe et al., 2023; Llull et al., 2023; Remondino et al., 2023) including texture-less, metallic, highly reflective, and transparent objects. Recent NeRF extensions even allow the detection and appropriate handling of mirroring surfaces in scenes (Holland et al., 2025). Alternatives to the involvement of a network to predict volumetric fields for density and view-dependent color information include the representation of scenes in terms of implicit surfaces (Wang et al., 2021a; Wang et al., 2023; Ge et al., 2023) as well as explicit representations in terms of meshes (Munkberg et al., 2022) or 3D Gaussians (Kerbl et al., 2023).

The aforementioned scene representation and rendering approaches are typically evaluated based on generating synthesized views from certain test configurations and subsequently evaluating the quality of the synthesized images using image-based metrics such as PSNR, SSIM or LPIPS. Thereby, the reported values typically correspond to the average across the generated views for the test configurations. Instead, several applications also rely on the geometric accuracy of the underlying model. Whereas only few investigations include evaluations of the achieved geometric accuracy, e.g. (Remondino et al., 2023; Haitz et al., 2024; Hillemann et al., 2024), these only depict the final deviations from ground truth models, but not the geometric convergence throughout the training process. Our study particularly investigates the latter aspect of geometric convergence behavior during the NeRF optimization process. Reaching geometric convergence at different times for different regions would mean that such regions should then not be further optimized. Similar to seminal works on image denoising based on deep image priors (Ulyanov et al., 2018), that rely on fitting a neural network to represent a single image, we demonstrate that convergence is reached at different timesteps for different regions of the scene. Thereby, our study motivates future developments on adaptively freezing the optimization of certain regions that have already converged in the reconstruction and spending the later optimization steps on refining uncertain regions, similar to locally-guided image denoising approaches (Bode et al., 2022).

3. Methodology

In this study, we investigate the geometric convergence behavior during the NeRF optimization process. Thereby, we intend to analyze whether convergence is reached earlier for certain regions in the scene than for other regions, which could motivate future locally-guided geometric optimization.

3.1 Neural Radiance Fields

Given a set of N input images with corresponding camera parameters (i.e. camera intrinsics and pose), Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) aim at novel view synthesis by optimizing an underlying continuous volumetric scene function. Using a feed-forward network to predict view-dependent radiance $\mathbf{c}(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^3$ and volume density $\sigma(\mathbf{x}) \in \mathbb{R}$ for a given spatial 3D location $\mathbf{x} \in \mathbb{R}^3$ and the view direction $\mathbf{d} \in \mathbb{R}^3$, the color observed in a particular pixel in the image is obtained by integrating along the respective viewing ray $\mathbf{r}(t) = \mathbf{o} + t\,\mathbf{d}$ in the volume, where the origin $\mathbf{o}$ of the ray coincides with the projective center of the camera. For this integration, the conventional NeRF approach (Mildenhall et al., 2020) exploits the sampling of $K \in \mathbb{N}$ positions $t_1, ..., t_K \in \mathbb{R}$ along the ray (Max, 1995). This allows optimizing the underlying neural network based on a loss function that penalizes the deviations of the synthesized images from their corresponding reference images, where the original formulation leverages the mean squared error between the rendered color $\hat{C}(\mathbf{r})$ and the corresponding color $C(\mathbf{r})$ from the input image for a batch of camera rays.

3.2 Geometric Convergence of NeRFs

In order to early detect geometrically converged scene structure at training time for a NeRF model, we investigate spatial changes at different subsequent training steps. For this purpose, we use point clouds as a discrete geometric scene representation, derived from the NeRF model as described in the following. The highest training signals in the form of backward gradients usually occur within the first 10% of the training steps. Therefore, we extract point clouds at a higher frequency in the beginning phase of the training, achieved through an empirically determined log_{10}-based distribution of write-out-steps

$\mathcal{S}$. We follow the median-based method in Nerfstudio (Tancik et al., 2023) for depth rendering and subsequent point cloud extraction. Therefore, weights w_i along each ray are created based on the density σ_i at each sampled position i on the ray with a distance δ_i between neighboring samples according to:

$$w_i = (1 - \exp(-\delta_i \sigma_i)) \cdot \exp\left(-\sum_{j=1}^{i-1} \delta_j \sigma_j\right) \tag{1}$$

Based on the computed weights at each sample position i, the median position is utilized as the depth d at pixel position (x, y).

Given this depth estimate obtained per ray, we can backproject the color information to a 3D point, thereby creating a colored point cloud $\mathcal{P}$. As implemented in the original NeRF (Mildenhall et al., 2020), depth per pixel can also be obtained via *expected depth*, which is based on the actual volume rendering procedure and computes the weighted sum of sample-point distances along the ray, divided by the total weight. The *expected depth* is also available in Nerfstudio (Tancik et al., 2023) as an alternative to the median depth computation.

In our experiments, however, the median and expected depth methods showed similar deviations with respect to a reference scan (e.g. as provided in the *Tanks and Temples* datasets (Knapitsch et al., 2017)) for the overall point cloud. We observed the *median*-based depth estimation to be slightly more robust than the approach based on *expected depth*, which motivated us to focus on using the *median*-based approach.

In order to compare the progression of the geometric convergence over training time, we calculate point-to-point distance metrics. The point cloud $\mathcal{P}$, generated from the optimized NeRF, is set as the reference point cloud to which the intermediate point clouds $\mathcal{Q}_s$ are compared. For that matter, the distance metrics are calculated from $\mathcal{Q}_s$ to $\mathcal{P}$. We utilize the mean $L2$ distance d_{L2} as well as the median and standard deviation (std). Besides that, we also compute the Chamfer and Hausdorff distances d_c and d_h for comparison:

$$d_{L2}(\mathcal{Q}_s, \mathcal{P}) = \frac{1}{|\mathcal{Q}_s| \cdot |\mathcal{P}|} \sum_{\mathbf{x} \in \mathcal{Q}_s} \sum_{\mathbf{x}' \in \mathcal{P}} \|\mathbf{x}' - \mathbf{x}\|_2 \tag{2}$$

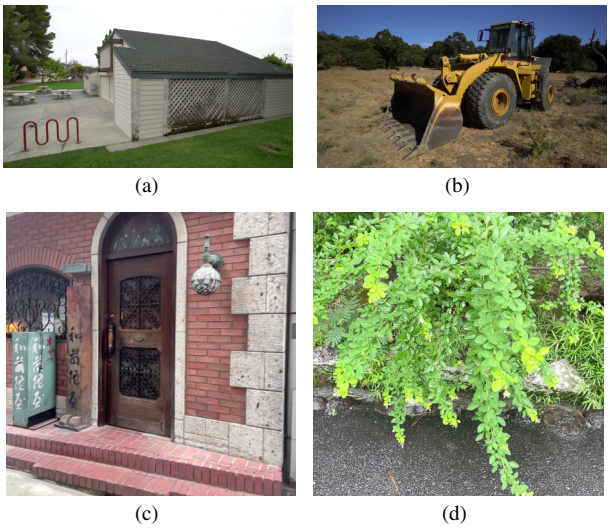
$$\begin{aligned} d_c(\mathcal{Q}_s, \mathcal{P}) = &\frac{1}{|\mathcal{Q}_s|} \sum_{\mathbf{x} \in \mathcal{Q}_s} \min_{\mathbf{x}' \in \mathcal{P}} \|\mathbf{x} - \mathbf{x}'\| + \\ &\frac{1}{|\mathcal{P}|} \sum_{\mathbf{x}' \in \mathcal{P}} \min_{\mathbf{x} \in \mathcal{Q}_s} \|\mathbf{x}' - \mathbf{x}\| \end{aligned} \tag{3}$$

$$\begin{aligned} d_h(\mathcal{Q}_s, \mathcal{P}) = &\frac{1}{2} \max_{\mathbf{x} \in \mathcal{Q}_s} \|\mathbf{x} - \mathrm{NN}(\mathbf{x}, \mathcal{P})\| + \\ &\frac{1}{2} \max_{\mathbf{x}' \in \mathcal{P}} \|\mathbf{x}' - \mathrm{NN}(\mathbf{x}', \mathcal{Q}_s)\| \end{aligned} \tag{4}$$

with

$$\mathrm{NN}(\mathbf{x}, \mathcal{P}) = \arg\min_{\mathbf{x}' \in \mathcal{P}} \|\mathbf{x} - \mathbf{x}'\|. \tag{5}$$

Equations (2)-(5) are referenced in (Williams, 2022) and (Zhou et al., 2018), from which the software implementations were used for our comparisons.



(a) (b) (c) (d)

Figure 1. Sample images of the four training datasets with number of images in braces: (a) *Barn* (410), (b) *Caterpillar* (383), (c) *Storefront* (1030) and (d) *Vegetation* (463).

4. Experiments

For the experiments, we use two scenes (i.e. *Barn* and *Caterpillar* scenes (see Figures 1(a) and 1(b))) from the *Tanks and Temples* dataset (Knapitsch et al., 2017) as well as two scenes (i.e. the *Storefront* and *Vegetation* scenes (see Figures 1(c) and 1(d))) from the *Nerfstudio* dataset (Tancik et al., 2023). Camera poses and intrinsics for all scenes were obtained via the implementation of Structure-from-Motion in COLMAP (Schönberger and Frahm, 2016). Furthermore, we train the Nerfacto model with default parameters and set the number of training steps to 30k. All results are based on the subset of intermediate steps $\mathcal{S} = \{78, 240, 499, 1073, 10123, 29999\}$, whereby 29999 denotes the last step and therefore corresponds to the reference point cloud $\mathcal{P}$. At each intermediate step s, a point cloud $\mathcal{Q}_s$ is generated from the Nerfacto model according to the procedure described in Section 3.

5. Results

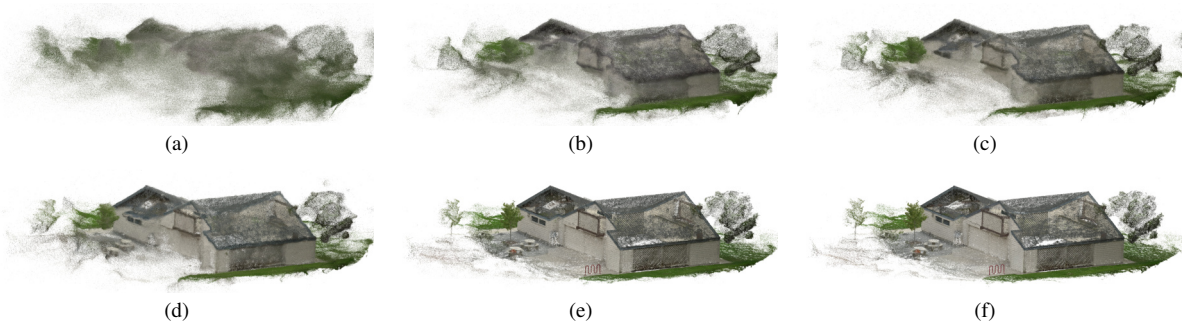
The results of the previously described experiments are shown qualitatively in Figures 2-9 as screen captures of point clouds and color-encoded point cloud distances, respectively. For quantitative comparison, we provide respective results for the distance metrics represented by Equations (2)-(4) in Tables 1-4. Note that we omit step 29999 in the tables since both point clouds are identical in that case.

#steps	mean distance	std	median distance	Chamfer distance	Hausdorff distance
78	0.02357	0.02796	0.01462	0.02810	0.38339
240	0.01630	0.02197	0.00813	0.02069	0.37498
499	0.01421	0.02167	0.00625	0.01780	0.37986
1073	0.00758	0.01265	0.00351	0.01046	0.37078
10123	0.00118	0.00152	0.00078	0.00221	0.34503

Table 1. Distance measures between the reference point cloud of the *Barn* scene and the point clouds obtained for different steps.

6. Discussion

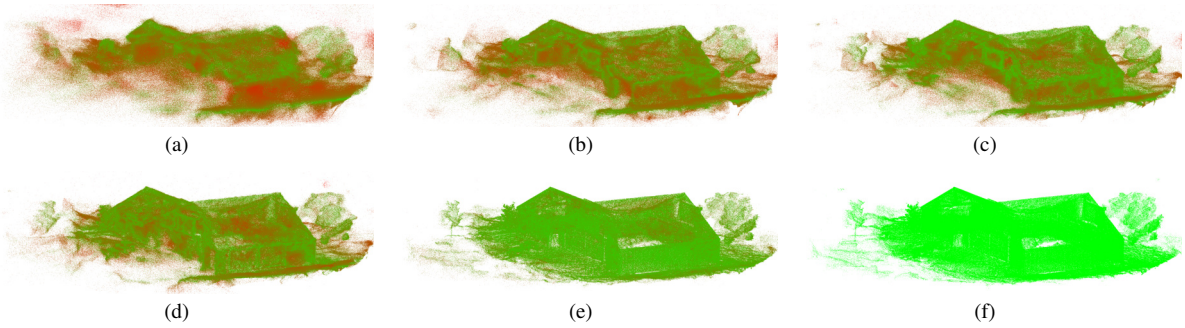
For all quantitative results in Tables 1-4 the observation can be drawn, that the distance of the point clouds derived for subsequent steps $s \in \mathcal{S}$ reduces with respect to the reference point



(a) (b) (c)

(d) (e) (f)

Figure 2. Extracted point clouds for the *Barn* scene at steps s: (a) 78, (b) 240, (c) 499, (d) 1073, (e) 10123, (f) 29999. Subfigure (f) shows the point cloud after the last training step.



(a) (b) (c)

(d) (e) (f)

Figure 3. Differences in $L2$ distance from $\mathcal{Q}_s$ to $\mathcal{P}$ at steps s of the *Barn* scene: (a) 78, (b) 240, (c) 499, (d) 1073, (e) 10123, (f) 29999. Red color indicates high differences, green low differences. Note that the last subfigure shows a self-comparison and is depicted for completeness only.

cloud derived from the fully trained NeRF. Within the first 1000 steps, the scene geometry undergoes a comparatively larger change than in the following 9000 steps, meaning the geometric convergence behavior is not linear. Changes between steps 10123 and 29999 mostly correspond to noise removal and slight adaptation of scene surfaces.

#steps	mean distance	std	median distance	Chamfer distance	Hausdorff distance
78	0.03245	0.04609	0.01595	0.04576	0.54157
240	0.02309	0.03774	0.00989	0.03378	0.55355
499	0.02022	0.03252	0.00861	0.02691	0.53502
1073	0.01358	0.02381	0.00490	0.01877	0.51768
10123	0.00227	0.00556	0.00099	0.00415	0.37694

Table 2. Distance measures between the reference point cloud of the *Storefront* scene and the point clouds obtained for different steps.

#steps	mean distance	std	median distance	Chamfer distance	Hausdorff distance
78	0.04264	0.07927	0.01462	0.04964	0.79381
240	0.01197	0.03617	0.00393	0.01809	0.78457
499	0.00808	0.02697	0.00266	0.01333	0.76313
1073	0.00595	0.02074	0.00198	0.00975	0.74919
10123	0.00155	0.00346	0.00092	0.00284	0.74306

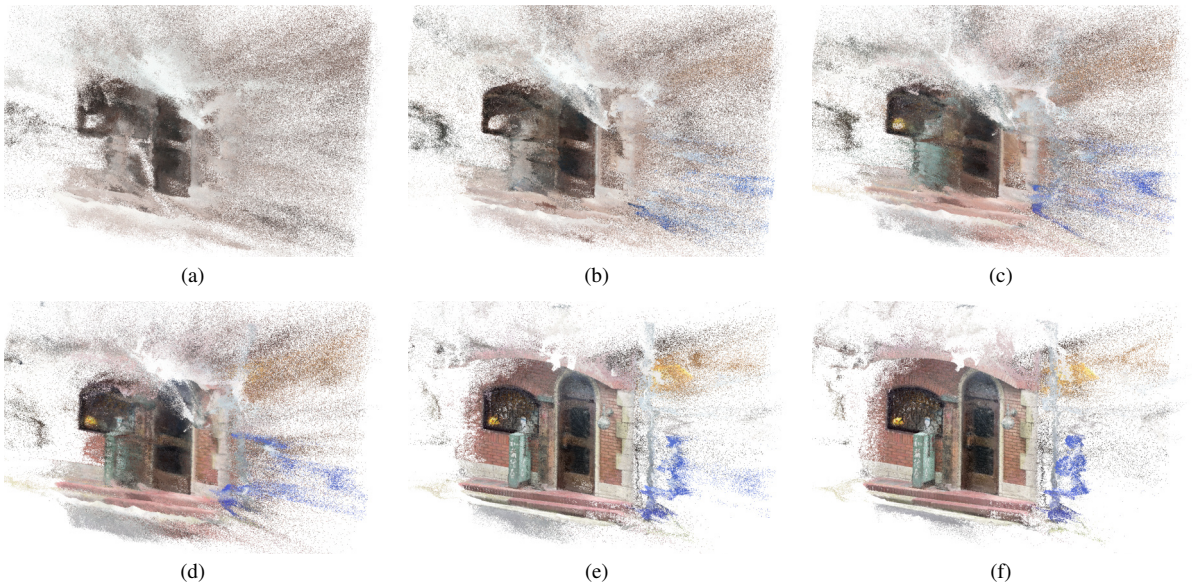
Table 3. Distance measures between the reference point cloud of the *Caterpillar* scene and the point clouds obtained for different steps.

#steps	mean distance	std	median distance	Chamfer distance	Hausdorff distance
78	0.04122	0.05307	0.01909	0.06979	0.70703
240	0.02737	0.04150	0.01045	0.05273	0.52334
499	0.01957	0.03351	0.00705	0.03866	0.50020
1073	0.01540	0.03134	0.00531	0.02987	0.50358
10123	0.00340	0.00606	0.00198	0.00613	0.39375

Table 4. Distance measures between the reference point cloud of the *Vegetation* scene and the point clouds obtained for different steps.

For the *Barn* scene, Table 1 depicts that almost all distance values decrease over training time. Steps 240 and 499 show a small increase in the Hausdorff distance. As the difference is comparably small, this can be attributed to the strong geometric variation that occurs in the early training phase, because the values then again drop further as training progresses. From the quantitative results of Table 1, the conclusion can be drawn that most geometric change happens within the first 1000 training steps to varying degrees. The qualitative results in Figures 2 and 3 reflect the quantitative results. As mentioned above, from step 78 to 1073, the scene is represented barely recognizable at first (cf. Figure 2(a)), then with a lot of wave-like surface approximation, until at step 1073 (cf. Figure 2(d)) the scene becomes pronounced in geometry. An interesting observation is that between step 10123 and the last step there seem to be only few visible changes, even though the former represents the training progress at only around 30%. In Figure 3, the point-to-point $L2$ distances are visualized, from which the mean, std and median metrics in Table 1 are derived. Similar observations can be obtained for the other considered scenes as becomes visible from the Figures 4-9.

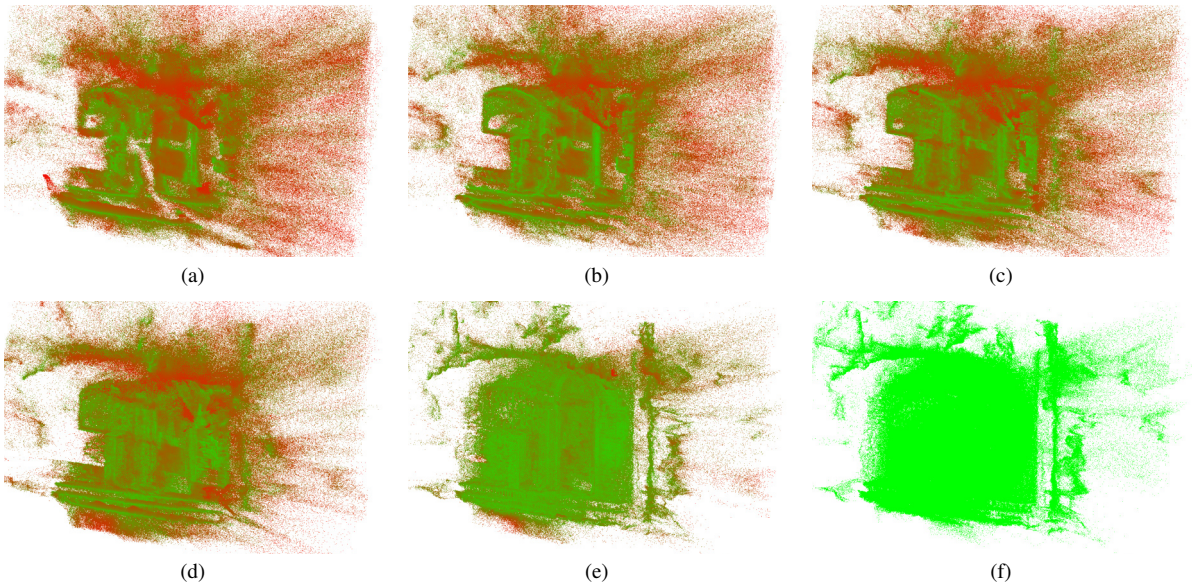
Generally, we can identify an initial rapid evolution from a very noisy representation across the whole scene to a low-frequency representation of the most significant structure of the underlying scene. In the following steps, the noise is gradually removed whereas significant scene structures get more and more refined. From the visual depictions, it seems as if the scene parts around more significant edges and boundaries seen from multiple views tend to converge earlier.



(a) (b) (c)

(d) (e) (f)

Figure 4. Extracted point clouds for the *Storefront* scene at steps s: (a) 78, (b) 240, (c) 499, (d) 1073, (e) 10123, (f) 29999. Subfigure (f) shows the point cloud after the last training step.



(a) (b) (c)

(d) (e) (f)

Figure 5. Differences in $L2$ distance from $\mathcal{Q}_s$ to $\mathcal{P}$ at steps s of the *Storefront* scene: (a) 78, (b) 240, (c) 499, (d) 1073, (e) 10123, (f) 29999. Red color indicates high differences, green low differences. Note that the last subfigure shows a self-comparison and is depicted for completeness only.

7. Conclusion

In this paper, we have investigated the geometric convergence behavior of NeRFs during the optimization of the underlying neural network in the training stage. For this purpose, we have derived discretized density fields in terms of point clouds corresponding to specific training iteration steps.

The comparison of those point clouds allows reasoning about which scene characteristics have reached convergence within the scene representation at a certain time step during the optimization. Since certain regions in the scene reach convergence earlier than other regions, future developments will be dedicated to locally-guided optimization to shift the computational burden to the adjustment of regions that need to converge further while leaving already converged regions unchanged. Thereby, a further reduction of training time may be achieved as well as an improvement of the resulting quality.

References

Balloni, E., Gorgoglione, L., Paolanti, M., Mancini, A., Pierdicca, R., 2023. Few shot photogrametry: A comparison between NeRF and MVS-SfM for the documentation of cultural heritage. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-M-2-2023, 155–162.

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. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5835–5844.

Figure 6. Extracted point clouds for the *Caterpillar* scene at steps s: (a) 78, (b) 240, (c) 499, (d) 1073, (e) 10123, (f) 29999. Subfigure (f) shows the point cloud after the last training step.

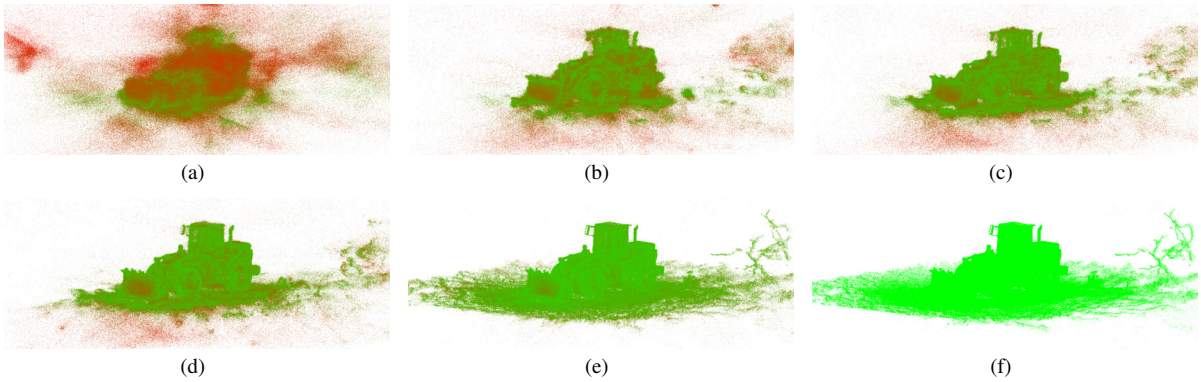


Figure 7. Differences in $L2$ distance from $\mathcal{Q}_s$ to $\mathcal{P}$ at steps s of the *Caterpillar* scene: (a) 78, (b) 240, (c) 499, (d) 1073, (e) 10123, (f) 29999. Red color indicates high differences, green low differences. Note that the last subfigure shows a self-comparison and is depicted for completeness only.

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., Hedman, P., 2022. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5460–5469.

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., Hedman, P., 2023. Zip-NeRF: Anti-aliased grid-based neural radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19697–19705.

Bian, W., Wang, Z., Li, K., Bian, J.-W., Prisacariu, V. A., 2023. NoPe-NeRF: Optimising neural radiance field with no pose prior. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4160–4169.

Bode, L., Merzbach, S., Kaltheuner, J., Weinmann, M., Klein, R., 2022. Locally-guided neural denoising. *Graphics and Visual Computing*, 7, 200058.

Buschmann, B., Dogaru, A., Eisemann, E., Weinmann, M., Egger, B., 2025. RANRAC: Robust neural scene representations via random ray consensus. *Proceedings of the European Conference on Computer Vision*, Springer, 126–143.

Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H., 2022. TensoRF: Tensorial radiance fields. *Proceedings of the European Conference on Computer Vision*, 333–350.

Chen, K., Dong, B., Wang, Z., Cheng, P., Yan, M., Sun, X., Weinmann, M., Weinmann, M., 2024. PriNeRF: Prior constrained neural radiance field for robust novel view synthesis of urban scenes with fewer views. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215, 383-399.

Chen, Y., Lee, G. H., 2023. DBARF: deep bundle-adjusting generalizable neural radiance fields. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24–34.

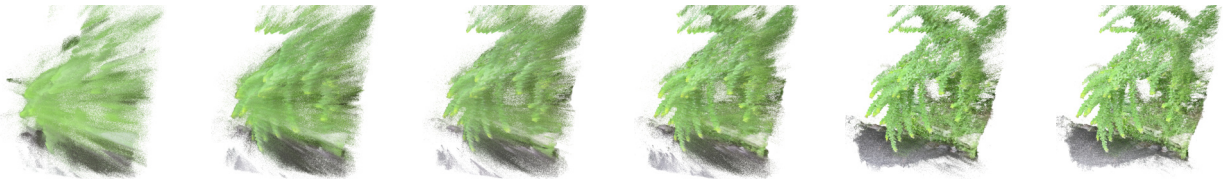
Condorelli, F., Rinaudo, F., Salvadore, F., Tagliaventi, S., 2021. A comparison between 3d reconstruction using NeRF neural networks and MVS algorithms on cultural heritage images. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B2-2021, 565–570.

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*, 12882–12891.

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*, 5491–5500.

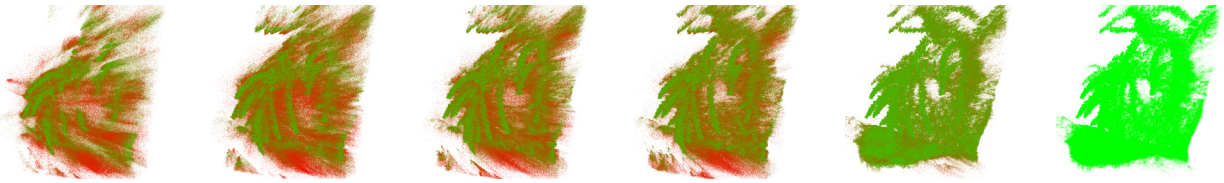
Ge, W., Hu, T., Zhao, H., Liu, S., Chen, Y.-C., 2023. Ref-NeuS: Ambiguity-reduced neural implicit surface learning for multi-view reconstruction with reflection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4228–4237.

Haitz, D., Hermann, M., Roth, A. S., Weinmann, M., Weinmann, M., 2024. The potential of neural radiance fields and 3D Gaussian splatting for 3D reconstruction from aerial imagery. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, X-2-2024, 97–104.



(a) (b) (c) (d) (e) (f)

Figure 8. Extracted point clouds for the *Vegetation* scene at steps s: (a) 78, (b) 240, (c) 499, (d) 1073, (e) 10123, (f) 29999. Subfigure (f) shows the point cloud after the last training step.



(a) (b) (c) (d) (e) (f)

Figure 9. Differences in $L2$ distance from $\mathcal{Q}_s$ to $\mathcal{P}$ at steps s of the *Vegetation* scene: (a) 78, (b) 240, (c) 499, (d) 1073, (e) 10123, (f) 29999. Red color indicates high differences, green low differences. Note that the last subfigure shows a self-comparison and is depicted for completeness only.

Haitz, D., Jutzi, B., Ulrich, M., Jäger, M., Hübner, P., 2023. Combining HoloLens with Instant-NeRFs: Advanced real-time 3D mobile mapping. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-1/W1-2023, 167–174.

Hillemann, M., Langendörfer, R., Heiken, M., Mehltretter, M., Schenk, A., Weinmann, M., Hinz, S., Heipke, C., Ulrich, M., 2024. Novel view synthesis with neural radiance fields for industrial robot applications. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-2-2024, 137–144.

Holland, L. V., Weinmann, M., Müller, J., Stotko, P., Klein, R., 2025. NeRFs are mirror detectors: Using structural similarity for multi-view mirror scene reconstruction with 3d surface primitives. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., Park, J., 2021. Self-calibrating neural radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5826–5834.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3d Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 139:1–139:14.

Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V., 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 37(4), 78:1–78:13.

Lin, C.-H., Ma, W.-C., Torralba, A., Lucey, S., 2021. BaRF: Bundle-adjusting neural radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5741–5751.

Llull, C., Baloian, N., Bustos, B., Kupczik, K., Sipiran, I., Baloian, A., 2023. Evaluation of 3d reconstruction for cultural heritage applications. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 1642–1651.

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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7206–215.

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

Mi, Z., Xu, D., 2023. Switch-NeRF: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. *Proceedings of the 11th International Conference on Learning Representations*, 1–15.

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. *Proceedings of the European Conference on Computer Vision*, 405–421.

Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4), 102:1–102:15.

Munkberg, J., Hasselgren, J., Shen, T., Gao, J., Chen, W., Evans, A., Müller, T., Fidler, S., 2022. Extracting triangular 3d models, materials, and lighting from images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8270–8280.

Pepe, M., Alfio, V. S., Costantino, D., 2023. Assessment of 3d model for photogrammetric purposes using AI tools based on NeRF algorithm. *Heritage*, 6(8), 5719–5731.

Remondino, F., Karami, A., Yan, Z., Mazzacca, G., Rigon, S., Qin, R., 2023. A critical analysis of NeRF-based 3d reconstruction. *Remote Sensing*, 15(14), 3585:1–3585:22.

Sabour, S., Vora, S., Duckworth, D., Krasin, I., Fleet, D. J., Tagliasacchi, A., 2023. RobustNeRF: Ignoring distractors with robust losses. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20626–20636.

Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Sun, C., Sun, M., Chen, H.-T., 2022. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5459–5469.

Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B. P., Srinivasan, P., Barron, J. T., Kretzschmar, H., 2022. Block-NeRF: Scalable large scene neural view synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8238–8248.

Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Kerr, J., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., McAllister, D., Kanazawa, A., 2023. Nerfstudio: A modular framework for neural radiance field development. *SIGGRAPH'23: ACM SIGGRAPH 2023 Conference Proceedings*, 72:1–72:12.

Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., Martin-Brualla, R., Simon, T., Saragih, J., Nießner, M., Pandey, R., Fanello, S., Wetzstein, G., Zhu, J.-Y., Theobalt, C., Agrawala, M., Shechtman, E., Goldman, D. B., Zollhöfer, M., 2020. State of the art on neural rendering. *Computer Graphics Forum*, 39 (2), 701–727.

Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Yifan, W., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., Simon, T., Theobalt, C., Nießner, M., Barron, J. T., Wetzstein, G., Zollhöfer, M., Golyanik, V., 2022. Advances in neural rendering. *Computer Graphics Forum*, 41(2), 703–735.

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*, 12912–12921.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2018. Deep image prior. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9446–9454.

Wang, C., Wu, X., Guo, Y.-C., Zhang, S.-H., Tai, Y.-W., Hu, S.-M., 2022. NeRF-SR: High quality neural radiance fields using supersampling. *Proceedings of the 30th ACM International Conference on Multimedia*, 6445–6454.

Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W., 2021a. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 354, 27171–27183.

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*, 3272–3283.

Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V. A., 2021b. NeRF–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*.

Williams, F., 2022. Point cloud utils. https://www.github.com/fwilliams/point-cloud-utils.

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. *Proceedings of the European Conference on Computer Vision*, 106–122.

Xie, S., Zhang, L., Jeon, G., Yang, X., 2023. Remote sensing neural radiance fields for multi-view satellite photogrammetry. *Remote Sensing*, 15(15), 3808:1–3808:17.

Xu, N., Qin, R., Huang, D., Remondino, F., 2024. Multi-tiling neural radiance field (NeRF) – Geometric assessment on large-scale aerial datasets. arXiv preprint arXiv:2310.00530v3.

Yen-Chen, L., Florence, P., Barron, J. T., Rodriguez, A., Isola, P., Lin, T.-Y., 2021. iNeRF: Inverting neural radiance fields for pose estimation. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1323–1330.

Zhou, Q.-Y., Park, J., Koltun, V., 2018. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*.