RANRAC: Robust Neural Scene Representations via Random Ray Consensus

Benno Buschmann[1,2(✉)], Andreea Dogaru[2], Elmar Eisemann[1], Michael Weinmann[1], and Bernhard Egger[2]

[1] Delft University of Technology, Delft, The Netherlands
b.buschmann@tudelft.nl

[2] Friedrich-Alexander-University Erlangen-Nürnberg, Erlangen, Germany

Abstract. Learning-based scene representations such as neural radiance fields or light field networks, that rely on fitting a scene model to image observations, commonly encounter challenges in the presence of inconsistencies within the images caused by occlusions, inaccurately estimated camera parameters or effects like lens flare. To address this challenge, we introduce RANdom RAy Consensus (RANRAC), an efficient approach to eliminate the effect of inconsistent data, thereby taking inspiration from classical RANSAC based outlier detection for model fitting. In contrast to the down-weighting of the effect of outliers based on robust loss formulations, our approach reliably detects and excludes inconsistent perspectives, resulting in clean images without floating artifacts. For this purpose, we formulate a fuzzy adaption of the RANSAC paradigm, enabling its application to large scale models. We interpret the minimal number of samples to determine the model parameters as a tunable hyperparameter, investigate the generation of hypotheses with data-driven models, and analyse the validation of hypotheses in noisy environments. We demonstrate the compatibility and potential of our solution for both photo-realistic robust multi-view reconstruction from real-world images based on neural radiance fields and for single-shot reconstruction based on light-field networks. In particular, the results indicate significant improvements compared to state-of-the-art robust methods for novel-view synthesis on both synthetic and captured scenes with various inconsistencies including occlusions, noisy camera pose estimates, and unfocused perspectives. The results further indicate significant improvements for single-shot reconstruction from occluded images.

Keywords: Neural scene representations · neural rendering · RANSAC · robust estimation · neural radiance fields · light-field networks

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-73116-7_8.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15134, pp. 126–143, 2025.
https://doi.org/10.1007/978-3-031-73116-7_8

1 Introduction

3D reconstruction is a classical task in computer vision and computer graphics, which has attracted research for decades. It offers numerous applications, including autonomous systems, entertainment, design, advertisement, cultural heritage, VR/AR experiences or medical scenarios. In recent years, neural scene representations and rendering techniques [44,51], including light field networks (LFN) [40] and neural radiance fields (NeRF) [29] have demonstrated great performance in single-view and multi-view reconstruction tasks. The key to the success of such techniques is the coupling of differentiable rendering methods with custom neural field parametrizations of scene properties. However, a common limitation of neural scene reconstruction methods is their sensitivity to inconsistencies among input images induced by occlusions, inaccurately estimated camera parameters or effects like lens flares. Despite the use of view-dependent radiance representations to address view-dependent appearance changes, these inconsistencies severely impact local density estimation, resulting in a poor generalization to novel views.

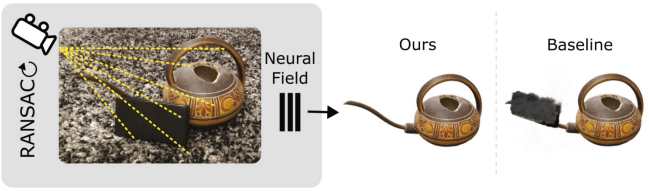

Fig. 1. We propose a robust algorithm for 3D reconstruction from occluded input perspectives that is based on the random sampling of hypotheses. Our algorithm is general and we demonstrate the use for single-shot reconstruction using light field networks or multi-view reconstruction using NeRF. In these cases, it successfully removes the artifacts that normally occur due to occluded input perspectives.

To increase the robustness to potential distractors within the training data, Sabour et al. [38] recently introduced the use of robust losses in the context of training unconditioned NeRF, where distractors in the training data were modeled as outliers of an optimization problem. However, the adaptation of this approach to conditioned neural fields (e.g., pixelNeRF [52]) is not obvious, as no optimization takes place during inference, and the input data is constrained to only a few views. Achieving robustness to data inconsistencies is a well-analyzed problem in computer vision, covered not only by the aforementioned robust loss functions [1], but also other strategies, like the random sample consensus (RANSAC) paradigm [15]. The latter is widely employed for fitting models to outlier-heavy data. The underlying idea is to randomly select subsets of the data to form potential models, evaluating these models against the entire dataset, and identifying the subset that best fits the majority of the data, while disregarding outliers. Despite being the state-of-the-art solution to many

challenges, RANSAC-based schemes are particularly favoured for the fitting of analytical models with a relatively small amount of tuneable parameters. In this paper, we direct our attention to achieving robustness against inconsistencies and occlusions in the observations by using a novel combination of neural scene representation and rendering techniques with dedicated outlier removal techniques such as RANSAC [15]. While downweighting the influence of distractors based on robust losses [1,38] can affect clean samples, representing details, we aim at improving robustness to distractors by only removing the influence of outliers. To this extent, we integrate a RANSAC-based scheme to distinguish inliers and outliers in the data and the inlier-based optimization of the neural fields (Fig. 1); a stochastic scenario characterized by a large-scale, data-driven model that exceeds RANSAC's classical convergence expectations. Instead of guaranteeing convergence to a clean sample set based on a minimal number of samples, we aim for a feasible (cleaner) sample set using a tuneable amount of samples. The proposed algorithm exhibits robustness and versatility, accommodating a wide range of neural fields-based reconstruction methods.

Our method inherits the strengths of RANSAC, such as the ability to handle various classes of outliers without relying on semantics. Yet, it also inherits the need for sufficiently clean samples and the reliance on an iterative scheme. In practice, the first condition is often fulfilled because typically only some of the perspectives are affected by inconsistencies. We validate our approach using synthetic data, focusing on the task of multi-class single-shot reconstruction with LFNs [40], and observe significant quality improvements over the baseline in the presence of occlusions. Furthermore, we showcase robust photo-realistic reconstructions of 3D objects using unconditioned NeRFs from sequences of real-world images in the presence of distractors. In comparison to RobustNeRF [38], we use all available clean data, hence improving the reconstruction quality for single-object scenes. Code and data are available under »https://bennobuschmann.com/ranrac«. Our key contributions are:

- a general, robust RANSAC-based reconstruction method applicable to a variety of neural-fields and handling diverse inconsistencies
- an analysis of the implication to RANSAC's hyperparameters and theoretical convergence expectations, and the experimental study of their effect
- a method for robust photo-realistic object reconstruction using NeRF and for robust single-shot multi-class reconstruction using LFNs
- a qualitative/quantitative evaluation of our method on both synthetic and real-world data with different inconsistencies (occlusions, invalid calibrations, ...) indicating the state-of-the-art performance of our approach

2 Related Work

Among the vast literature on neural fields, the seminal work of Mildenhall et al. [29] opened many avenues in the computer vision community. It contributed to state-of-the-art solutions for novel view synthesis and 3D reconstruction that

have been covered in respective surveys [44,51]. Noteworthy is the more recent contribution of instant neural graphics primitives (iNGP) [31], which uses a hash table of trainable feature vectors alongside a small network for representing the scene. iNGP achieved major run-time improvements, thereby enhancing the feasibility of practical applications for neural fields.

Baseline models are highly sensitive to imperfections in the input data, which led to many works on robustness enhancements of neural fields; addressing a reduced amount of input views [17,21,32,52], errors in camera parameters [4,19, 25,53], variations in illumination conditions across observations [28,42], multi-scale image data [2,3,50], and the targeted removal of floating artifacts [33,47, 48].

Fewer works solve the reconstruction task in the presence of inconsistencies between observations. Bayes' Rays [16] provides a framework to quantify uncertainty of a pretrained NeRF by approximating a spatial uncertainty field. It handles missing information due to self-occlusion or missing perspectives well, but cannot deal with inconsistencies caused by noise or distractors. Similarly, NeuRay [27] only supports missing, but not inconsistent information. Naive occlusion handling via semantic segmentation requires the occluding object types to be known in advance [28,36,37,43,46]. Solutions to learn semantic priors on transience exist [23] but separating occlusions via semantic segmentation without manual guidance is ill posed. Occ-NeRF [54] considers any foreground element as occlusion and removes them via depth reasoning, but their removal leaves behind blurry artifacts. Alternatively, some methods do not remove dynamic distractors, but reconstruct them together with the rest of the scene using time-conditioned representations [9,26,34,49]. Closely related to our work, RobustNeRF [38] considers input-image distractors as outliers of the model optimization task. The authors employ robust losses improved via patching to preserve high-frequency details. RobustNeRF does not rely on prior assumptions about the nature of distractors, nor does it require preprocessing of the input data or postprocessing of the trained model. Nevertheless, their method comes at the cost of loosing view-dependent details and a reduced reconstruction quality in undistracted scenes. Furthermore, their method is limited to unconditioned models that overfit to a single scene. A generalization to conditioned NeRFs, such as pixelNeRF [52], is not obvious, as no further optimization takes place during inference.

Conditioned neural fields offer a distinct advantage in their ability to generalize to novel scenes by leveraging knowledge acquired from diverse scenes during learning. This results in a more robust model that requires as few as one input view for inference, showcasing the efficiency and adaptability of the approach. Contrary to PixelNeRF [52], which relies on a volumetric parametrization of the scene, demanding multiple network evaluations along the ray, Light Field Networks (LFNs) [40], which succeed Scene Representation Networks [41], take a different approach. LFNs represent the scene as a 4D light field, enabling a more efficient single evaluation per ray for inference. The network takes as input a ray represented in Plücker coordinates and maps it to an observed radiance, all within an autoencoder framework used for conditioning.

None of the mentioned methods can deal with occlusions in single-shot reconstruction and no prior work exists on robust LFNs or robustness of other conditioned neural fields for single-shot reconstruction, which we address via the RANSAC paradigm [15]. Since its introduction in 1981, RANSAC has gained attention for fitting analytical models with a small number of parameters, such as homography estimation in panorama stitching [7]. Among the few direct applications of classical RANSAC to larger models is robust morphable face reconstruction [13,14]. Other common expansions and applications include differentiable RANSAC [5] for camera parameter estimation in a deep learning pipeline, locally optimized RANSAC [11] to account for the requirement of a descriptive sample set, and adaptive real-time RANSAC [35].

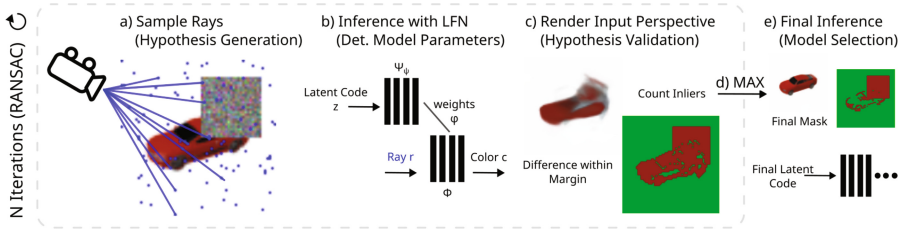

Fig. 2. The RANRAC algorithm for LFNs samples random hypotheses by selecting a set of random samples from the given perspective (a), and inferring the latent representation of these rays using the autodecoder of a pretrained LFN (b). The obtained light field is then used to predict an image from the input perspective (c). Based on this prediction, confidence in the random hypothesis is evaluated via the Euclidean distance between the predicted ray colors and the remaining color samples of the input image. The amount of samples which are explained by each hypothesis up to some margin are used to determine the best hypothesis (d). All samples explained by the selected hypothesis are used for a final inference with the LFN to obtain the final model and latent representation (e).

3 Method

In this section, we present our approach to increase the robustness of neural scene representations to inconsistencies in the input data. First, we recap RANSAC, its theoretical convergence and hyperparameters, and the required adaptions for its application to high-dimensional data-driven models. We then introduce a general scheme for the random sampling and validation of neural fields. Based hereon, we formulate a robust algorithm using LFNs for 3D reconstruction from a single image with occlusions. Finally, using NeRF, we formulate an algorithm for robust photo-realistic reconstruction from multiple views in the presence of common sources of inconsistencies.

3.1 RANSAC Convergence on Complex Models

Classical RANSAC [10,12,15] follows an iterative process. Initially, a minimal set of samples is randomly selected to determine the model parameters, known as the hypothesis generation phase. Then, the hypothesis is evaluated by assessing the number of observations it explains, within a specified margin. These steps are repeated until the best hypothesis is chosen to constitute the consensus set, which comprises all of its inliers.

This paradigm cannot be directly applied to complex models such as neural fields, as a significant amount of samples is required to obtain decent initial model parameters and additional clean samples improve the quality further. This imposes a challenge regarding the expected amount of clean initial sample sets, S_{clean}:

$$\mathbb{E}[\#S_{clean}] = N * \prod_{m=1}^{M} \frac{s_{img} - s_{img}^{occ} - m}{s_{img} - m}, \tag{1}$$

where s_{img} denotes the total amount of samples (e.g., image pixels), s_{img}^{occ} represents the occluded samples, and N/M are the number of iterations/samples. The expected amount exponentially decreases with the initial number of samples.

The samples and respective requirements for analytic and data-driven models vary a lot. The effect of individual samples is less traceable in data-driven models and the information entropy varies more significantly across samples. When using a model that projects onto a latent space, some very atypical outliers do not show an effect at all if the latent space is not expressive enough to explain them in an overall loss-reducing way, yet, outliers close to the object or its color, or larger chunks of outliers, will usually be distracting. At the same time, samples of small-scale high-frequency details are important for the reconstruction and contain a lot of information, whereas multiple samples of larger-scale lower-frequency details contribute much less. The amount of initial samples for the hypothesis generation becomes a tunable hyperparameter trading initial reconstruction quality for likelihood of finding desired sample sets. This invalidates the classical convergence idea [10,12] where the RANSAC iterations N with

$$N \geq \frac{\log(1-p)}{\log(1-t^M)} \tag{2}$$

are chosen such that at least one clean sample set is found with a probability p, given the expected ratio of clean samples t and the amount of initial samples M. There is not only the need to find a clean sample set, but one that captures all important details. At the same time, a completely clean sample set is not required at all, as long as the contained outliers are not represented by the local minimum of the latent space or the model itself, depending on the concrete scene representation.

3.2 Random Sampling Neural Fields

We propose a general strategy for RANSAC-like, iterative, robust reconstruction with neural fields, before formulating respective algorithms for LFNs and NeRF.

Sampling Hypotheses. Based on the application a feasible sampling domain is chosen *e.g.* pixels/rays, or observations. Depending on the requirements of the neural field an appropriate sample size is determined. The initial sample size is chosen as small as possible to allow for a reasonable convergence expectation to cleaner sets, while being large enough for a coarse fit of the model expressive enough to enable the discrimination of outliers.

Determine Model Parameters. The determination of the model parameters in a classical RANSAC application corresponds to the inference of the sample sets with a neural field: In case of unconditioned neural fields (such as NeRF) this corresponds to (over-)fitting the model to the scene, in case of conditioned fields this corresponds to obtaining a latent representation. Full convergence is not required, the reconstruction is only used to evaluate the outlier contamination.

Validation of Hypotheses. For the validation, the rays/perspectives not used for the inference are rendered using the obtained model and compared to the input. The quality of each hypothesis is evaluated based on the amount of other samples explained by it up to some margin. Depending on the sampling/inference strategy, different similarity metrics can be used to distinguish outliers from other sources of noise such as the coarse inference. The initial sample set of the best hypothesis together with all its inliers is used for a final complete model fit.

As will be demonstrated in our evaluation in Sect. 5, the major benefit of our RANSAC-like neural field approach over formulations based on robust loss functions [1] is the possibility to reliably filter outliers and inconsistencies in the input data in comparison to the down-weighting of their influence.

3.3 Robust Light Field Networks

In the following, we propose a novel fast and robust single-shot multi-class reconstruction algorithm based on LFNs [40]. LFNs are globally conditioned, meaning that the supported subset of 3D consistent scenes is represented by a single global latent vector. Therefore, when the latent space is not expressive enough to represent object and distractor correctly, large inconsistencies cause global damage instead of local artifacts. LFNs intrinsically support parallel inference, allowing to jointly process all hypotheses rather than following an iterative scheme. Our algorithm consists of the following steps (Fig. 2):

1. **Hypothesis Consensus Set:** Given the input image I as a set of pixel color values $\boldsymbol{c}_i$, and the intrinsic and extrinsic camera parameters, the set of rays R – one ray $\boldsymbol{r}_i$ for every pixel – represented by Plücker coordinates, is generated. In the first step, N initial consensus sets S_n are drawn using a uniform distribution, where each consensus set consists of M random samples:

$$(\boldsymbol{c}_n^m, \boldsymbol{r}_n^m) \in_R \{(\boldsymbol{c}_i, \boldsymbol{r}_i) \mid \boldsymbol{c}_i \in I, \boldsymbol{r}_i \in R\} \tag{3}$$

where $n \in \{1, \ldots, N\}$, $m \in \{1, \ldots, M\}$, and $\in_R$ denotes a sample randomly drawn from the set without replacement according to a uniform distribution.

2. **Hypothesis Inference:** The autodecoder of an LFN Φ with hypernetwork Ψ and pretrained hypernetwork weights ψ, is used to infer the latent codes $\boldsymbol{z}_n$ for each of the initial sample sets in parallel.

$$\{\boldsymbol{z}_n\} = \underset{\{\boldsymbol{z}_n\}}{\operatorname{argmin}} \sum_n \sum_m \|\Phi(\boldsymbol{r}_n^m \mid \Psi_\psi(\boldsymbol{z}_n)) - \boldsymbol{c}_n^m\|_2^2 + \lambda_{lat}\|\boldsymbol{z}_n\|_2^2 \tag{4}$$

   λ_{lat} determines the strength of the Gaussian prior on the latent space [40]. An exponential learning rate schedule speeds up the inference. The inferred latent codes form the hypotheses.
3. **Hypothesis Prediction:** Each of the hypotheses is used to render an entire image I_n^{pred} from the perspective of the input image, each consisting of the pixel color values $\boldsymbol{c}_{n,i}^{pred} = \Phi(\boldsymbol{r}_i \mid \Psi_\psi(\boldsymbol{z}_n))$, using again the set of rays $R = \{\boldsymbol{r}_i\}$ obtained from the camera parameters. The rendered pixels resemble the predictions for the remaining observations under each hypothesis.
4. **Hypothesis Validation:** The obtained predictions are compared to the input image to validate the hypothesis. For each pixel in each predicted image, we calculate the Euclidean distance in color space: $e_{n,i} = \|\boldsymbol{c}_{n,i}^{pred} - \boldsymbol{c}_i\|_2$. For each image, using these distances, we collect, up to some margin ϵ, the observations explained by the model (inliers): $S_n^{inlier} = \{(\boldsymbol{c}_i, \boldsymbol{r}_i) \mid e_{n,i} < \epsilon\}$
5. **Model Selection:** We select the best hypothesis sample set S_{best} based on the number of inliers $\#S_n^{inlier}$. The model is inferred once more, similar to the second step, to obtain the final latent code z_{cons}. The inference is based on the final consensus set $S_{cons} = S_{best} \cup S_{best}^{inlier}$, the initial sample set of the strongest hypothesis S_{best} together with all its inliers S_{best}^{inlier}. The final output consists of both the latent code z_{cons} and the final consensus set S_{cons} of the selected model, and can be used to render arbitrary new perspectives.

3.4 Robust Neural Radiance Fields

In the following, we propose a novel robust multi-view reconstruction approach based on NeRFs. NeRFs are fit to a single scene based on a set of observations. As they support view-dependent radiance, one might expect inconsistencies in the input observations to only have an effect on specific perspectives. However, the density is only spatially parametrized, hence, inconsistencies lead to significant ghosting and smearing artifacts in more than just the inconsistent perspective. This can be leveraged in the hypothesis validation. The following steps are performed for N iterations:

1. **Hypothesis Consensus Set:** A NeRF fit requires a large set of rays, making a sampling in ray space infeasible, as even obtaining significantly clean*er* sets becomes unlikely. In order to obtain a sampling domain with reasonably sized initial consensus sets S_n, we sample M observations from the given sets of images $\mathcal{I}$ and corresponding camera poses $\mathcal{C}$ in every iteration:

$$(I_n^m, E_n^m) \in_R \{(I_i, E_i) \mid I_i \in \mathcal{I}, E_i \in \mathcal{C}\} \tag{5}$$

   where $n \in \{1, \ldots, N\}$, $m \in \{1, \ldots, M\}$, and $\in_R$ denotes a sample randomly drawn from the set without replacement according to a uniform distribution.

2. **Hypothesis Inference:** The sampled observations are used to fit a hypothesis neural radiance field F_n.
3. **Hypothesis Prediction:** The obtained NeRF F_n is used to render predictions $I_{n,i}^{pred}$ with pixel colors $\boldsymbol{p}_{n,i,x}^{pred}$ under the hypothesis for all unseen input perspectives E_i.
4. **Hypothesis Validation:** The chosen sample space requires a careful evaluation of the hypotheses. We propose a two-step evaluation, where, first, the pixels inliers $P_{n,i}$ up to some margin ϵ_{pix} are determined for every observation based on the Euclidean distance in color space to the pixels $\boldsymbol{p}_{i,x}$ of the input images I_i: $P_{n,i} = \{\boldsymbol{p}_{n,i,x}^{pred} \mid \|\boldsymbol{p}_{n,i,x}^{pred} - \boldsymbol{p}_{i,x}\|_2 < \epsilon_{pix}\}$ and, second, the observations themselves are labeled as inliers or outliers based on the amount of pixel inliers, again up to some margin ϵ_{img}, to obtain the consensus set S_n^{inlier} under the hypothesis: $S_n^{inlier} = \{(I_i, E_i) \mid \#P_{n,i} > \epsilon_{img}\}$. The binary metric for pixels ensures that smaller mispredictions (due to, e.g., view-dependent lighting effects) do not introduce noise into the evaluation.
5. **Model Selection:** The strongest hypothesis, with its initial sample set S_{best}, is selected based on the number of inliers $\#S_n^{inlier}$ to obtain the final consensus set $S_{cons} = S_{best} \cup S_{best}^{inlier}$. The final model is obtained with one more NeRF fit of the consensus set.

3.5 Hyperparameters

For LFNs, the amount of initial samples and random hypotheses to evaluate (iterations), are determined experimentally. Without fine-tuning per class, the experimentally determined parameters are 90 initial samples and 2000 iterations, which supersedes the theoretical value for a convergence because the latent space introduces an intrinsic robustness. The inlier margin balances the amount of slight high-frequency variations that are being captured and the capability of separating outliers that are similar to the object. A margin of 0.25 in terms of the Euclidean distance of the predicted colors to the input samples in an RGB color space normalized to the range $(-1, 1)$ has been found to be optimal. Please refer to the supplementary material for further experimental results.

For NeRFs, the parameters behave more natural and the amount of perspectives required for a meaningful, not completely artifact-free fit of the model lies around 25 observations [29, Table 2]. With fewer samples, more artifacts are introduced that get harder to separate from the ones caused by inconsistencies, and the samples get more dependent on being evenly spaced. For real-world captures with 10% inconsistent perspectives, as few as 50 iterations are sufficient. With the color space normalized to $(0, 1)$, a pixel margin between around 0.15 in terms of Euclidean distance worked well for the determination of actual artifacts. We consider an observation an inlier based on a margin of 90% - 98% pixel inliers, which proved to be a good choice to separate minor artifacts (due to the sparse sampling) from artifacts caused by actual inconsistencies. For different datasets or inconsistencies, these values can be adapted.

4 Implementation and Preprocessing

For LFNs, we build on top of the original implementation [40], with a slight adaptation to enable a parallel sub-sampled inference. We furthermore use the provided pretrained multi-class model. The camera parameters are known. For efficiency reasons, the steps of the algorithm are not performed iteratively, but multiple hypotheses are validated in parallel. To further speed up the inference, an exponential learning rate schedule is used for the auto-decoding, leading to a total runtime of about a minute on a single GPU.

For the robust reconstruction of objects from lazily captured real-world data, one has to estimate the camera parameters and extract foreground masks before applying the algorithm. For the estimation of the camera parameters, we used the COLMAP structure-from-motion package [39]. We extracted foreground masks using Segment Anything [22]. However, only foreground masks, containing the objects and the occlusions, are extracted. Segment Anything is not capable of removing arbitrary occlusions in an automatized way. After these preprocessing steps, the robust reconstruction algorithm can be applied as described. Erroneous estimates of the camera parameters or foreground masks are excluded by our algorithm, thus making the entire reconstruction pipeline robust.

Our algorithm is not limited to a specific NeRF implementation. The chosen sampling domain eases integration into arbitrary existing NeRF implementations, which commonly expect images instead of unstructured ray sets. However, using a fast NeRF variant is advantageous when applying an iterative scheme. We used the instant NGP implementation [31] of the instant NSR repository [18], which includes some accelerations [24,30]. Other, (specifically fast) variants are likely good choices as well. Antialiased and unbounded, but slow variants, such as MipNeRF360 [3], are not feasible. For further implementation details please refer to the supplementary.

5 Experiments

5.1 Inconsistencies, Baselines and Datasets

Multi-View Reconstruction (NeRF). We provide a comprehensive quantitative and qualitative analysis of RANRAC's multi-view reconstructions compared to NeRF without any method of robustness (baseline) and RobustNeRF [38] (state of the art). We conduct experiments for various common sources of inconsistencies such as occluded perspectives, noisy camera parameter estimates, and blurred perspectives. RobustNeRF [38] targets unbounded scenes with multiple objects and small amounts of distractors in every perspective. In comparison, our RANSAC-based approach deals well with single-object reconstruction, even with heavy occlusions, as long as enough clean perspectives are available. As their dataset reflects the algorithm's properties, we cannot provide a fair comparison. Instead, we demonstrate the applicability using a custom dataset of a single object with a controlled amount of deliberately occluded perspectives. For

the other inconsistencies we use of-the-shelf datasets and add noise to the camera parameters and blur to the images. Furthermore, we implement the robust losses [38] on the same NeRF variant as RANRAC to provide a fair comparison of the method's robustness, independent of the NeRF variant. For further details on the implementation, please refer to the supplementary material.

Single-Shot Reconstruction (LFN). We benchmark against the original LFN implementation of Sitzmann et al. [40] as baseline, as there are no other robust methods for LFNs or conditioned neural fields, nor are there methods for robust single-shot multi-class reconstruction in general. Furthermore, we use the same pretrained LFN for the baseline and for the application of our method. The LFN is pretrained on the thirteen largest ShapeNet classes [8].

We provide a detailed qualitative and quantitative performance comparison under different amounts of occlusion for three representative classes (plane, car, and chair), while just stating reconstruction performance in a fixed environment without additional tuning of the hyperparameters for the others. The plane class is mostly challenging due to the low-frequency shape, while the car class contains a lot of high-frequency color details. The chair class represents shapes that are generally problematic for vanilla LFNs, even without occlusions. We provide a complementing analysis of the hyperparameters in the supplementary material. If not stated otherwise, we evaluate using 50 randomly selected images of the corresponding class. All comparisons use the same images.

The occlusions are created synthetically. They consist of randomly generated patches while controlling two metrics of occlusion: Image occlusion and object occlusion. The former is the naive ratio of occluded over total pixels. The latter are the occluded pixels on the object compared to the total pixels covered by the object. We use both metrics to take the vastly different information entropy of samples across the image into account. For further details on the generation of the synthetic occlusions, please refer to the supplemental.

5.2 Evaluation

For LFNs, our approach leads to a significant improvement in occluded scenarios of up to 8dB in PSNR and a similarly strong improvement for the SSIM. The improvement is most significant in heavily distracted scenarios (Fig. 3). In clean scenarios a slight performance penalty can be observed, but even with small amounts of object occlusion (information loss), our algorithm outperforms the baseline, leading to numerically better results up to 50% information loss (Fig. 3). The effect is not only measurable, but also well visible (Fig. 4). Increasing amounts of occlusion slowly introduce local artifacts into our reconstruction while preserving a reasonable shape estimate even for larger amounts of occlusion. In contrast, the reconstruction of LFNs breaks rather early in a global fashion. Still, our consensus set (Fig. 4), reveals that some high-frequency details were wrongfully excluded, explaining the slight performance decrease on clean images. In general, the benefit of RANRAC is best observable for classes that

Table 1. RANRAC obtains a significant quantitative improvement in PSNR and SSIM (higher is better) compared to the baseline. We compare RANRAC to vanilla LFNs for the 13 largest ShapeNet classes (find plane, car, and chair with more detail in Fig. 3). The results are based on a moderate amount of occlusion of approximately 25% object occlusion and about 5% image occlusion. The reported results are conservative, as higher amounts of image occlusion result in a more significant performance increase (Fig. 3). No hyperparameter tuning has been performed for these classes; the same configuration obtained from the analysis of the other three classes is used.

Metric	Model	Bench	Boat	Cabin.	Displ.	Lamp	Phone	Rifle	Sofa	Speak.	Table
PSNR↑	RANRAC	**19.21**	**22.92**	**21.92**	17.65	**19.99**	**18.45**	**21.35**	**21.61**	**20.7**	**20.44**
	LFN	17.89	19.2	20.5	**18.85**	19.09	17.81	18.46	20.12	19.99	20.29
SSIM↑	RANRAC	**0.767**	**0.858**	**0.801**	0.699	**0.764**	**0.75**	**0.853**	**0.805**	**0.761**	**0.784**
	LFN	0.724	0.791	0.767	**0.73**	0.748	0.726	0.795	0.775	0.743	0.777

can be well described by LFNs, as evidenced by the lower improvements the chair class exhibits, compared to the plane and car class. The same effect is visible in the quantitative evaluation on the other classes without additional tuning (Table 1). The only outlier is the display class, on which LFNs struggle the most on [40], even in unoccluded scenarios. It is not consistently represented in the latent space and robust reconstruction amplifies this effect by reducing the samples to a consistent set.

The application to NeRF shows the significantly improved reconstructions of RANRAC (Table 2) for different types of inconsistencies such as occlusions and blur, as well as noisy camera parameters on captured and synthetic datasets. RANRAC consistently outperforms RobustNeRF [38] for object reconstruction from inconsistent inputs. RobustNeRF specifically struggles in capturing view-dependent appearance and at concavities while RANRAC seamlessly reconstructs them. The effects are not only measurable but also well-visible (Fig. 5).

6 Limitations and Future Work

For the application to NeRF, the iterations imply the use of a fast inferable variant, ruling out MipNeRF360 [3] and other high-quality variants for unbounded scenes. Foreground separation is a must, limiting this application of RANRAC to object-centric scenes. Novel fast, unbounded methods might resolve this [20].

By following a RANSAC-like approach, we inherit the requirement of sufficient clean perspectives, which could be lifted via NeRF variants that require fewer perspectives [32,45] or by using different sampling domains. In return, our method is not limited to specific kinds of inconsistencies, and is robust to arbitrarily heavy distractions or inconsistencies in the impure perspectives.

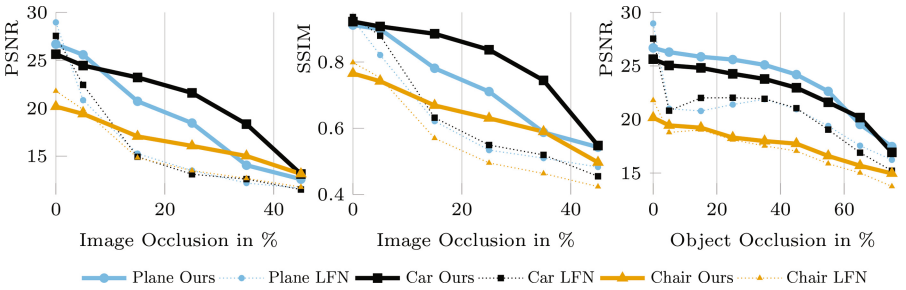

Fig. 3. RANRAC (solid lines) leads to a quantitative improvement in PSNR and SSIM (higher is better) for occluded inputs compared to vanilla LFNs (dashed lines). The same hyperparameter configuration and LFN is used for all classes. On the left and in the middle, the amount of image occlusion is increased, while the object occlusion is constant at 25%. On the right, the amount of object occlusion is increased while the image occlusion is kept low. For the car class, a large improvement is observed over the entire occlusion spectrum. For the plane class the improvement is similarly significant, but absolute performance degenerates a bit sooner. This stems from the smaller object size and the related faster occlusion-to-object increase when increasing image occlusions. For the chair class, the improvement is less significant but the structural similarity is preserved for much longer. For the plane and car class the reconstruction quality is resilient to information loss (right) up to ∼50%, where the decrease gains momentum. With the low amounts of image occlusion, the improvement is not significant for the chair class (consistent with left and middle).

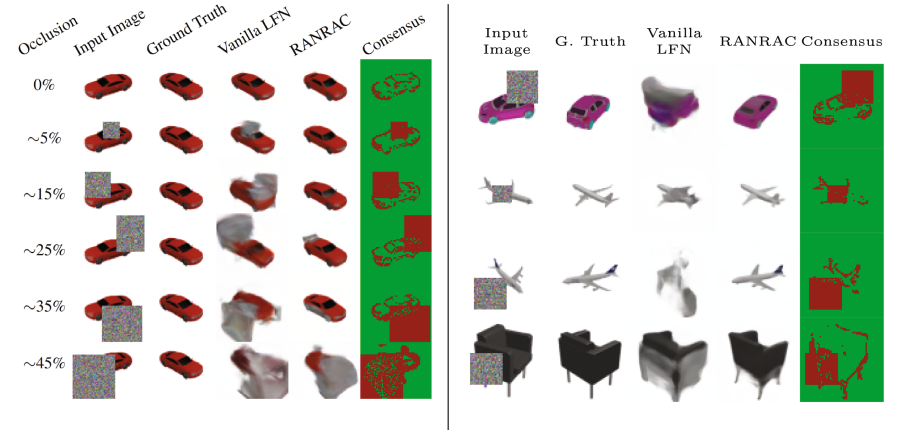

Fig. 4. On the left, we show the qualitative effect of increasing occlusion on the same observation for the reconstruction of a novel view. Reconstructions of LFNs break early globally whereas RANRAC still provides a very decent reconstruction, only slowly introducing minor local (and natural/comprehensible) artifacts for completely hidden object parts. We further show the obtained consensus set, used for the final reconstruction (green inliers, red outliers). On the right, we show more qualitative results for novel view synthesis on different classes and the corresponding consensus sets. (Color figure online)

Table 2. RANRAC outperforms the state of the art RobustNeRF [38] and the baseline NeRF (without a method of robustness) for scenes contaminated with occlusions, blur and noisy camera parameters. We report PSNR↑ averaged across perspectives and the 5th percentile (Avg.|P_5) as artifacts introduced by inconsistencies only contaminate some views. We compare on the captured watering pot dataset with milder (10%) and heavier (17.5%) amounts of occluded perspectives and off-the-shelf datasets [29] with blurred perspectives and additive Gaussian noise $\mathcal{N}(5°, 1°)$ on the camera parameters of 10% of the perspectives. All three variants are built on top of instant-nsr for an isolated fair comparison of the robustness method. Note that both robust approaches struggle separating the strong view-dependent effects of the ship scene from blur, leading to the exclusion of some perspectives (lower P_5 PSNR), while RANRAC still improves the overall reconstruction. For all other scenes and inconsistencies, RANRAC reliably separates inconsistent observations from clean ones.

Inconsistency	Mild Occ.	Strong Occ.	Blurred Perspectives				
Dataset	Watering Pot		Lego	Ship	Chair	Ficus	Mic
RANRAC	**27.11**|**25.99**	**26.12**|**24.94**	**34.79**|**29.91**	**29.76**|16.14	**35.25**|**31.26**	**31.36**|**28.54**	**35.78**|**32.99**
RobustNeRF	26.83|25.58	25.93|24.79	29.14|23.69	23.31|20.19	31.11|27.12	24.57|23.35	30.21|26.85
NeRF	26.65|22.61	25.36|18.47	31.15|19.00	28.48|**20.35**	33.21|22.10	29.04|20.12	31.91|18.94

Inconsistency	Noisy Camera Parameters						
Dataset	Lego	Drums	Mic	Ship	Ficus	Hotdog	Materials
RANRAC	**34.95**|**31.22**	**25.88**|**22.77**	**35.85**|**33.55**	**30.77**|**22.05**	**31.41**|**28.82**	**37.16**|**30.53**	**28.93**|**25.39**
RobustNeRF	29.83|26.82	23.91|22.10	29.36|27.55	24.21|19.98	23.17|21.85	32.65|25.16	24.90|22.30
NeRF	24.67|15.95	23.22|17.96	28.82|22.15	23.68|14.80	28.12|22.77	28.16|18.37	24.73|19.89

Whereas our robust LFN approach led to significant improvements for single-shot reconstruction, our concept might also be applied to other conditioned neural fields, based on a smart choice of the sampling domain, targeting photo-realism. One could also use importance sampling based on a prior, instead of a uniform sampling, leveraging the unevenly distributed information entropy. Neural sampling priors [6], semantic segmentation and more advanced schemes (e.g. LO-RANSAC [11], DSAC [5]) might prove useful here.

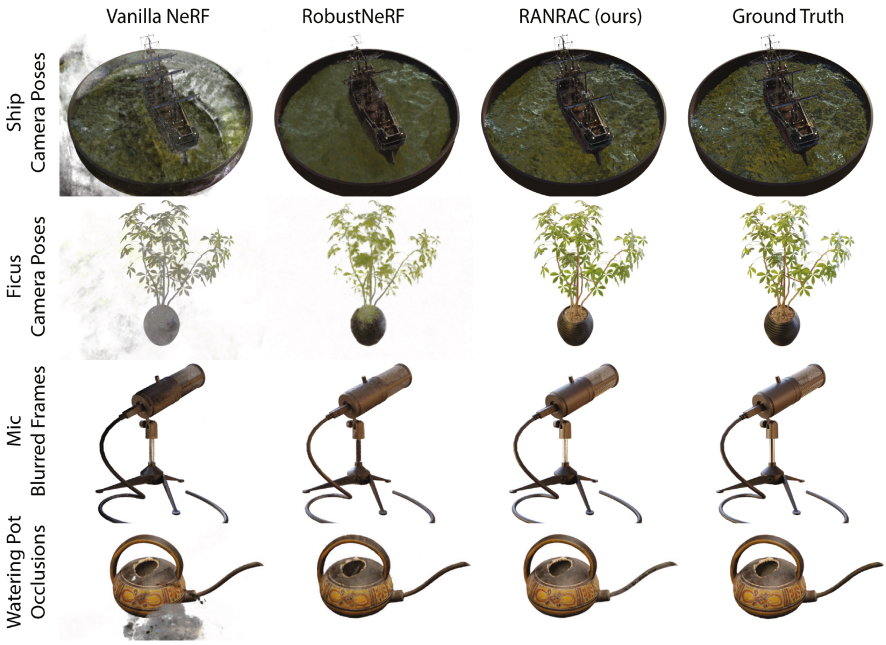

Fig. 5. The occlusions lead to well-visible artifacts in the reconstructions of NeRF, these artifacts are completely removed by RANRAC. While RobustNeRF struggles with view-dependent and high-frequency details, RANRAC reliably reconstructs them.

7 Conclusion

We introduced a novel approach to increase the robustness of neural fields, inspired by the RANSAC paradigm. Following this concept, we introduced a novel robust approach for single-shot reconstruction from occluded views based on LFNs which achieves a significant improvement in reconstruction quality for distracted and occluded scenarios, even for extreme cases. Furthermore, we introduced a respective RANRAC-based NeRF variant that allows robust photo-realistic reconstruction from multiple views with typical inconsistencies such as occlusions, noisy camera parameters, or blurred images – resulting in significant improvements compared to state-of-the-art methods – without relying on assumptions about the distractions.

Acknowledgements. Benno Buschmann was at FAU funded through a gift by Mitsubishi Electric Research Laboratories (MERL). Andreea Dogaru was funded by the German Federal Ministry of Education and Research (BMBF), FKZ: 01IS22082 (IRRW). The authors are responsible for the content of this publication.

References

1. Barron, J.T.: A general and adaptive robust loss function. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4331–4339 (2019)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields. In: ICCV (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-NeRF 360: unbounded anti-aliased neural radiance fields. In: CVPR (2022)
4. Bian, W., Wang, Z., Li, K., Bian, J.W., Prisacariu, V.A.: Nope-NeRF: optimising neural radiance field with no pose prior. In: CVPR (2023)
5. Brachmann, E., et al.: DSAC-differentiable RANSAC for camera localization. In: CVPR (2017)
6. Brachmann, E., Rother, C.: Neural-guided RANSAC: learning where to sample model hypotheses. In: ICCV (2019)
7. Brown, M., Lowe, D.G.: Automatic panoramic image stitching using invariant features. Int. J. Comput. Vision **74**, 59–73 (2007)
8. Chang, A.X., et al.: ShapeNet: an information-rich 3D model repository. Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015)
9. Chen, X., et al.: Hallucinated neural radiance fields in the wild. In: CVPR (2022)
10. Choi, S., Kim, T., Yu, W.: Performance evaluation of RANSAC family. J. Comput. Vis. **24**(3), 271–300 (1997)
11. Chum, O., Matas, J., Kittler, J.: Locally optimized RANSAC. In: Michaelis, B., Krell, G. (eds.) DAGM 2003. LNCS, vol. 2781, pp. 236–243. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45243-0_31
12. Derpanis, K.G.: Overview of the RANSAC algorithm. Image Rochester NY **4**(1), 2–3 (2010)
13. Egger, B., Schneider, A., Blumer, C., Forster, A., Schönborn, S., Vetter, T.: Occlusion-aware 3D morphable face models. In: BMVC, vol. 2, p. 4 (2016)
14. Egger, B., et al.: Occlusion-aware 3D morphable models and an illumination prior for face image analysis. Int. J. Comput. Vision **126**, 1269–1287 (2018)
15. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
16. Goli, L., Reading, C., Sellán, S., Jacobson, A., Tagliasacchi, A.: Bayes' Rays: uncertainty quantification in neural radiance fields. arXiv (2023)
17. Guangcong, Chen, Z., Loy, C.C., Liu, Z.: SparseNeRF: distilling depth ranking for few-shot novel view synthesis. In: ICCV (2023)
18. Guo, Y.C.: Instant neural surface reconstruction (2022). https://github.com/bennyguo/instant-nsr-pl
19. Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., Park, J.: Self-calibrating neural radiance fields. In: ICCV (2021)
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian splatting for real-time radiance field rendering (2023)
21. Kim, M., Seo, S., Han, B.: InfoNeRF: ray entropy minimization for few-shot neural volume rendering. In: CVPR (2022)
22. Kirillov, A., et al.: Segment anything. arXiv:2304.02643 (2023)

23. Lee, J., Kim, I., Heo, H., Kim, H.J.: Semantic-aware occlusion filtering neural radiance fields in the wild. arXiv preprint arXiv:2303.03966 (2023)
24. Li, R., Gao, H., Tancik, M., Kanazawa, A.: NerfAcc: efficient sampling accelerates NerFS. arXiv preprint arXiv:2305.04966 (2023)
25. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: BARF: bundle-adjusting neural radiance fields. In: ICCV (2021)
26. Liu, Y.L., et al.: Robust dynamic radiance fields. In: CVPR (2023)
27. Liu, Y., et al.: Neural rays for occlusion-aware image-based rendering. In: CVPR (2022)
28. Martin-Brualla, R., Radwan, N., Sajjadi, M.S.M., Barron, J.T., Dosovitskiy, A., Duckworth, D.: NeRF in the wild: neural radiance fields for unconstrained photo collections. In: CVPR (2021)
29. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 405–421. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_24
30. Müller, T.: Tiny-cuda-nn (2021). https://github.com/NVlabs/tiny-cuda-nn
31. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. In: SIGGRAPH (2022)
32. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S.M., Geiger, A., Radwan, N.: RegNeRF: regularizing neural radiance fields for view synthesis from sparse inputs. In: CVPR (2022)
33. Philip, J., Deschaintre, V.: Floaters no more: radiance field gradient scaling for improved near-camera training. In: Eurographics Symposium on Rendering. The Eurographics Association (2023)
34. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: neural radiance fields for dynamic scenes. In: CVPR (2021)
35. Raguram, R., Frahm, J.-M., Pollefeys, M.: A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 500–513. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88688-4_37
36. Rebain, D., Matthews, M., Yi, K.M., Lagun, D., Tagliasacchi, A.: LOLNeRF: learn from one look. In: CVPR (2022)
37. Rematas, K., et al.: Urban radiance fields. In: CVPR (2022)
38. Sabour, S., Vora, S., Duckworth, D., Krasin, I., Fleet, D.J., Tagliasacchi, A.: RobustNeRF: ignoring distractors with robust losses. In: CVPR (2023)
39. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
40. Sitzmann, V., Rezchikov, S., Freeman, W.T., Tenenbaum, J.B., Durand, F.: Light field networks: neural scene representations with single-evaluation rendering. In: NeurIPS (2021)
41. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: continuous 3D-structure-aware neural scene representations. In: NeurIPS (2019)
42. Sun, J., et al.: Neural 3D reconstruction in the wild. In: SIGGRAPH (2022)
43. Tancik, M., et al.: Block-NeRF: scalable large scene neural view synthesis. In: CVPR (2022)
44. Tewari, A., Thies, J., Mildenhall, B., et al.: Advances in neural rendering. Comput. Graph. Forum. **41**, 703–735 (2022)
45. Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: SPARF: neural radiance fields from sparse and noisy poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4190–4200 (2023)

46. Turki, H., Ramanan, D., Satyanarayanan, M.: Mega-NeRF: scalable construction of large-scale nerfs for virtual fly-throughs. In: CVPR (2022)
47. Warburg, F., Weber, E., Tancik, M., Hołyński, A., Kanazawa, A.: Nerfbusters: removing ghostly artifacts from casually captured NerFS. In: ICCV (2023)
48. Wirth, T., Rak, A., Knauthe, V., Fellner, D.W.: A post processing technique to automatically remove floater artifacts in neural radiance fields. In: Computer Graphics Forum (2023)
49. Wu, T., Zhong, F., Tagliasacchi, A., Cole, F., Oztireli, C.: D2NeRF: self-supervised decoupling of dynamic and static objects from a monocular video. In: NeurIPS (2022)
50. Xiangli, Y., et al.: BungeeNeRF: progressive neural radiance field for extreme multi-scale scene rendering. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV (2022). https://doi.org/10.1007/978-3-031-19824-3_7
51. Xie, Y., et al.: Neural fields in visual computing and beyond. Comput. Graphics Forum **41**, 641–676 (2021)
52. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: neural radiance fields from one or few images. In: CVPR (2021)
53. Zhang, J.Y., Ramanan, D., Tulsiani, S.: RelPose: predicting probabilistic relative rotation for single objects in the wild. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) ECCV (2022). https://doi.org/10.1007/978-3-031-19821-2_34
54. Zhu, C., Wan, R., Tang, Y., Shi, B.: Occlusion-free scene recovery via neural radiance fields. In: CVPR (2023)