ELLIS Summer School
Using Neural Radiance Fields in real-world scenarios

ClearView

Camera Pose Refinement for Improved Radiance Fields

By Jauhri, Liu, Konen, Sabharwal, Di Nucci and Quattrini

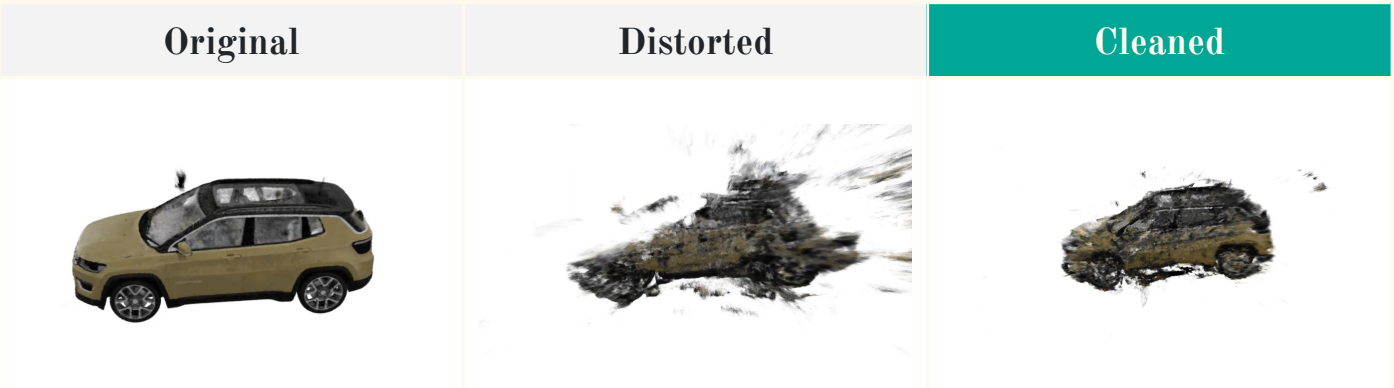

Figure 1: NeRF Reconstructions Comparison - Original poses are displayed on the left, illustrating the standard structure. The middle shows results from distorted poses, revealing apparent anomalies. The right demonstrates our method's refinement of distorted poses, yielding significantly improved coherence and structure.

INTRODUCTION

The introduction of Neural Radiance Fields (NeRFs) has provided a significant push in synthesising novel views of complex scenes, representing them as continuous functions in a multi-dimensional colour-geometry space [1].

This study explores and addresses the inherent challenges Neural Radiance Fields (NeRFs) face in rendering high-fidelity 3D scene reconstructions, explicitly focusing on artefacts induced by noisy camera poses. We propose an approach aimed at pre-processing and refining camera poses before training NeRF models, targeting the mitigation of such artefacts and enhancing the quality of reconstructions. A recent method, SPARF [4], refines poses by projecting key points in 3D and reprojecting them in 2D at train time. Our methodology circumvents this by adding an optimisation process during the pre-processing stage, allowing for universal applicability across various NeRF architectures. The anticipated outcome is a reduction in artefacts and inconsistencies in the reconstructed scenes, enabling more accurate and reliable renderings. This project, while minimally put together as a preliminary study, holds the potential to contribute to the development of more precise 3D modelling and rendering techniques.

METHOD

In an ideal world, we work under the assumption that we can generate sets of corresponding pixels when provided with an image pair. According to the fundamentals of multi-view geometry, it is a requirement

that both of these pixels must map back to the identical 3D point within the world coordinate system (Figure 2, left). When we introduce noise on the poses, this assumption becomes invalid, creating inconsistencies in the projection into the real world (Figure 2, centre). In this scenario, our method is composed of four different steps.

Keypoint estimation: To perform the 3D projection, it is first necessary to obtain correspondences between images, and this can be achieved by relying on a pre-trained dense correspondence network [6] to extract matches between the training views.

Depth estimation: A depth estimation network is vital in determining the depth ***Dkv*** of each keypoint ***KPv*** of every view ***V*** and extracting their respective 3D coordinates. An affordable choice for this task is DINOV2, which has demonstrated state-of-the-art performance in monocular depth estimation.

3D projection: Knowing the ***RTv*** matrix for a view and the intrinsic values ***K*** of the camera. For each keypoint identified within a view, it is projected into 3D space using the following formula:

$$3Dkp = (KPv * K^{-1}) * Dkv * RTv$$

2D reprojection: Acknowledging the potential imperfections in-depth estimation and the difficulty for the architecture to precisely locate the correct points, our proposal involves reprojecting keypoints from their 3D positions into another view. Subsequently, we compute the distance between the reprojected 2D values. In the context of two image pairs, ***V1*** and ***V2***, the keypoints identified in these images are subsequently reprojected into the 2D view of the other image(Figure 2, right), thereby determining a distance that serves as a valuable loss metric.

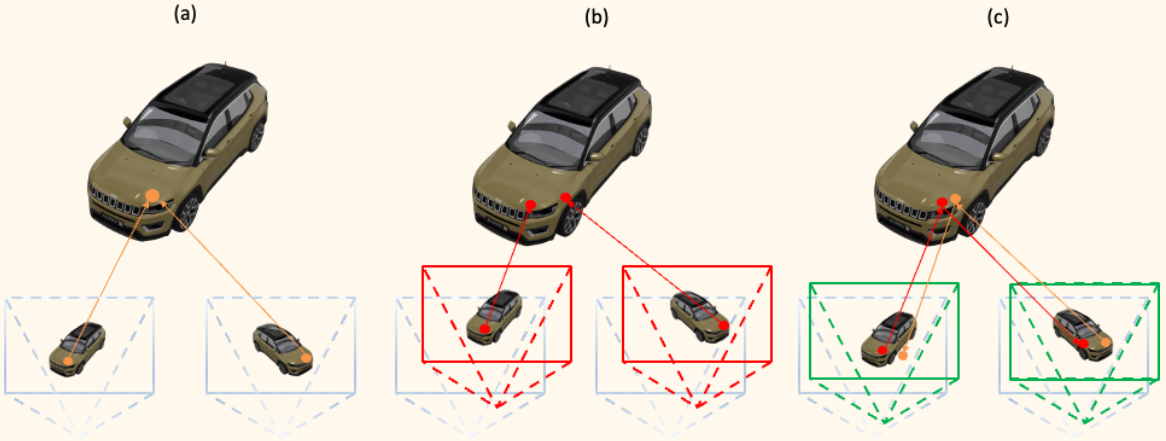

Figure 2: (a) Keypoint 3D projection with ground truth camera poses, (b) 3D projection with noisy camera poses, (c) 3D projection with optimised camera poses and reprojection to 2D.

EXPERIMENTAL RESULTS

We employ the CarPatch dataset [3] for our experiments—a synthetic benchmark specialising in vehicles that include images, intrinsic and extrinsic camera parameters, and depth maps. The average rotation and translation errors between the optimised poses and the ground truth are reported to underline the effectiveness of pose correction in our approach. Additionally, a qualitative comparison is performed for view synthesis.

Our initial findings, derived from the ground truth keypoints and depth maps, aim to elucidate the potential ramifications on subsequent NeRF reconstructions. Instant-NGP [2] models are trained utilising nerfstudio [5] over 30k iterations, a process completing in approximately 30 minutes on an NVIDIA RTX6000. For the sake of experimentation, intentional distortions were introduced; an average noise of 0.25m in translation and 5 degrees in rotation. The results of the pose optimization are depicted in Table 1, illustrating the efficacy of our pre-processing method in rectifying the noisy poses.

	Translation Error	Rotation Error
Noisy poses	0.24m	4.53°
Cleaned poses	0.21m	4.11°

Table 1: Comparative Analysis of Average Errors

To quantify the influence of the refined poses on NeRF representation, an ablation study was conducted, with initial training based on clean poses serving as the baseline. Subsequent introduction of random noise to the camera's extrinsic parameters led to a discernible decline in performance, rectified by the application of our refined methodology, resulting in enhanced NeRF output reconstruction.

Figure 1 illustrates the disparities in novel view rendering between models trained with noisy and refined poses. The former struggles with learning accurate geometries, yielding degenerate forms, while the latter, fortified with refined poses, exhibits robustness and precision in rendering accurate views.

FINAL THOUGHTS

We introduce ClearView, a pre-processing methodology designed to rectify inaccuracies in camera estimates by enforcing multi-view consistency through key points and depth maps. This results in refined camera poses and improved NeRF reconstructions. Unlike methods such as SPARF, which optimise poses during training, our refinement process during data pre-processing is universally applicable to any existing NeRF implementation, eliminating the need for additional code adaptations. The initial results demonstrate promise, and subsequent studies will explore important metrics including PSNR and SSIM to quantify improvements.

Bibliography

1. 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. Communications of the ACM,
2. Thomas Müller, Alex Evans, Christoph Schied, Alexander Keller, "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding", SIGGRAPH, July 2022
3. Di Nucci, Davide, et al. "CarPatch: A Synthetic Benchmark for Radiance Field Evaluation on Vehicle Components." *ICIAP*. Cham: Springer Nature Switzerland, 2023.
4. Truong, Prune, et al. "Sparf: Neural radiance fields from sparse and noisy poses." CVPR. 2023.
5. Tancik, M., et al. "Nerfstudio: A modular framework for neural radiance field development.
6. Truong, Prune, et al. "Learning accurate dense correspondences and when to trust them." *CVPR*. 2021.
7. Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision."