SOAC: Spatio-Temporal Overlap-Aware Multi-Sensor Calibration using Neural Radiance Fields

Quentin Herau[1,2] Nathan Piasco[1] Moussab Bennehar[1] Luis Roldão[1]
Dzmitry Tsishkou[1] Cyrille Migniot[3] Pascal Vasseur[4] Cédric Demonceaux[2]
[1]Noah's Ark, Huawei Paris Research Center [2]ICB UMR CNRS 6303, Université de Bourgogne
[3]ImViA UR 7535, Université de Bourgogne [4]MIS UR 4290, Université de Picardie Jules Verne
{Quentin.Herau, Nathan.Piasco, Moussab.Bennehar, Luis.Roldao, Dzmitry.Tsishkou}@huawei.com
{Quentin.Herau@etu., Cyrille.Migniot@, Cedric.Demonceaux@}u-bourgogne.fr
Pascal.Vasseur@u-picardie.fr

Abstract

In rapidly-evolving domains such as autonomous driving, the use of multiple sensors with different modalities is crucial to ensure high operational precision and stability. To correctly exploit the provided information by each sensor in a single common frame, it is essential for these sensors to be accurately calibrated. In this paper, we leverage the ability of Neural Radiance Fields (NeRF) to represent different sensors modalities in a common volumetric representation to achieve robust and accurate spatio-temporal sensor calibration. By designing a partitioning approach based on the visible part of the scene for each sensor, we formulate the calibration problem using only the overlapping areas. This strategy results in a more robust and accurate calibration that is less prone to failure. We demonstrate that our approach works on outdoor urban scenes by validating it on multiple established driving datasets. Results show that our method is able to get better accuracy and robustness compared to existing methods.

1. Introduction

Multi-sensor calibration plays a key role in autonomous systems as it ensures accuracy, reliability, and robustness in safety-critical tasks such as localization [6] and perception [22] in self-driving. In typical multi-sensor setups, the sensors are attached to a common rigid body where the spatial relationship between them can be obtained through a rigid transformation matrix. It is therefore important to identify the exact values of those matrices to correctly exploit and merge the data provided by the sensors. The process of finding these spatial transformations is called extrinsic calibration, which is a topic that has been and is still being heavily studied thanks to the increasing popularity of multi-sensor algorithms. In addition to spatial calibration, without an external synchronization system, it is also necessary to perform temporal calibration. Using temporally miscalibrated sensors, performance on different tasks can be severely hindered. Although certain approaches in the literature address temporal misalignment [12, 29, 36], the prevailing assumption among these methods is the presence of properly synchronized sensors. Due to the importance of sensor calibration, a multitude of calibration solutions exist in the literature, as highlighted in the review from Li et al. [17] and summarized in Tab. 1. They can be classified into two main categories: target-based and targetless methods.

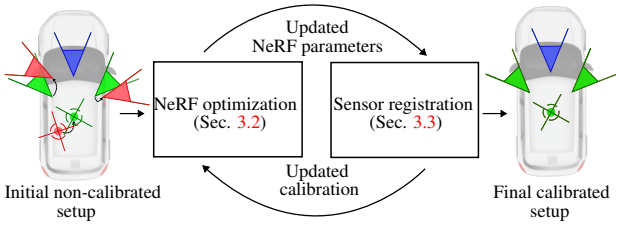


Figure 1. **Method overview.** SOAC is a novel multimodal spatio-temporal calibration method for cameras and LiDAR in the context of autonomous driving. By alternating the training of multiple implicit scenes (Sec. 3.2) and sensors co-registration from these representations (Sec. 3.3), SOAC achieves precise self-supervised calibration from raw data acquired in unconstrained urban environments.

Target-based calibration methods rely on one or more elements of known dimensions and features purposefully placed in the scene. The most classic target is a checkerboard [9, 46], but custom-made planar targets [11] or boxes [30] have also been proposed. These methods usually offer precise and robust calibration compared to targetless approaches. However, requiring hand-placed targets prevents them from being deployed on a large scale and

does not enable on-the-fly re-calibration if needed. Thus, a more suitable method for mass-produced autonomous driving cars would be targetless.

Targetless methods do not require manually placed targets and thus can be used on sequences captured without user intervention. This makes them more suitable for large-scale deployment. These approaches usually rely on shared information (i.e. overlap) between the different sensors, which can be of different modalities. Wang et al. [38] and Pandey et al. [28] propose a correspondence between the reflectivity of the LiDAR scans and the grayscale intensity of the camera images. Other methods propose to find matches of specific features, like edges [44] or semantic classes [16].

Following the development of deep learning, methods relying on deep models were introduced to calibrate RGB images and LiDAR scans. These methods have the advantage of being fast and precise, enabling reliable online calibration. Deep learning techniques can leverage regression [13, 20, 33], flow [15], keypoints [42] or convolutional features [7] to supervise or regularize the training. However, as they are supervised methods, they need an accurately calibrated training dataset to be optimized and have issues with cross-domain data due to overfitting to a specific dataset or sensor layout.

Recently, with the arrival of Neural Radiance Fields (NeRF) [24] for implicit representation of 3D scenes, some works [12, 40, 47] propose to take advantage of the fully differentiable structure of the model to achieve self-supervised targetless calibration. Using a NeRF as the common frame for the sensors, these methods are able to densely correlate the captured observation from different sensors in an implicit volumetric space. Yet, by simultaneously learning the information from multiple sensors, the NeRF might overfit regions of the scene only visible from a single sensor without enforcing consistency on the overlapping regions. This causes the calibration to easily get stuck in a local minimum.

We take inspiration from the aforementioned works by exploiting the fully differentiable properties of the implicit scene representation to achieve spatial and temporal calibration. Different from existing methods [12, 40, 47], we propose to represent the scene by using multiple NeRFs akin to their corresponding sensor and advocate to alternate the optimization target between NeRF training and sensor calibration (i.e. Fig. 1). Our method avoids overfitting the pose optimization to partial regions of the scene, resulting in a more robust and accurate calibration.

2. Related Work

With NeRF and the papers improving upon it [1, 26], the main focus was on the quality of novel view synthesis in addition to training and rendering speeds. However, since these approaches often validate their claims on carefully curated datasets, it is often assumed that the input poses corresponding to the data are already available and are accurate. However, in real-world situations, some or all the captured frames might be unposed or suffer from inaccuracies, hence, significantly impacting the quality of the final reconstruction result [19]. Therefore, several works later on attempted to tackle this issue through different formulations and adaptations of the overall optimization problem.

		Targetless	Cam/Cam	Cam/LiDAR	Temporal	Self-supervised
Target-based	Zhang et al. [46]	X	X	✓	X	-
	Geiger et al. [9]	X	X	✓	X	-
Feature-based	Pandey et al. [28]	✓	X	✓	X	-
	Park et al. [29]	✓	X	✓	✓	-
Deep-learning	RegNet [33]	✓	X	✓	X	X
	LCCNet [20]	✓	X	✓	X	X
NeRF-based	INF [47]	✓	X	✓	X	✓
	MOISST [12]	✓	✓	✓	✓	✓
	SOAC (ours)	✓	✓	✓	✓	✓

Table 1. Comparison of calibration methods.

NeRF-based Image Registration. To register an image with incorrect or no pose, iNeRF [43] proposes to use an already trained NeRF. It finds the pose that minimizes the photometric difference between the captured image and the rendered result from the model. By focusing on regions of interest, it is able to register unseen images with high precision. Using this idea as a basis, Loc-NeRF [21] combines Monte Carlo localization method [5] with the use of a pre-trained NeRF as a map, to build a real-time global localization method. CROSSFIRE [25] takes advantage of the NeRF model's flexibility to learn not only the radiance and density information of the map, but also a descriptor field. During the localization process, by iteratively matching the descriptors from the query image and the information given by the NeRF model, this method is able to provide high-precision localization. Nevertheless, all these methods require training a NeRF from precise camera poses first before being able to localize new query images.

NeRF-based Pose Optimization. The first method to leverage the fully differentiable nature of NeRF to optimize the input poses through backpropagation is NeRF-- [39]. It proposes to optimize both the NeRF and the input poses by representing them as embeddings and show higher novel view synthesis quality when trained from noisy poses. BARF [19] improves upon this idea by adding a coarse-to-fine component to this method. It progressively liberates the frequencies of the input positional encoding to prevent the optimization from getting stuck in a local minimum. SCNeRF [14] adds camera distortion estimation and

uses a different 6-vector rotation formulation in the optimization, while SPARF [37] achieves pose optimization with sparse input views by relying on pixel matching and depth consistency. While the aforementioned methods need an initial estimate of the camera poses, some recent methods completely remove the need for prior poses. NoPe-NeRF [2] uses an off-the-shelf monocular depth estimator (i.e. DPT [31]) to regularize relative poses between successive images. GNeRF [23] relies on adversarial learning to coarsely estimate the initial poses before refining them in a second phase. IR-NeRF [45] improves upon GNeRF by regularizing the implicit pose estimator with the unposed real images, increasing its robustness. Although the NeRF-based pose optimization methods achieve reasonable scene reconstruction by recovering accurate camera poses, they are not suited for autonomous driving data as they do not handle multi-modal observations nor take into account the rigidity constraint between multiple sensors mounted on a vehicle.

NeRF-based Sensor Calibration. NeRF-based calibration methods [12, 40, 47] take advantage of the rigid constraint between the sensors and the differentiable nature of NeRF to efficiently solve this challenging task. These methods have the advantage of being targetless and self-supervised, as they do not rely on an annotated training dataset. The idea is to use the NeRF as a common scene representation. Each sensor provides its observations (RGB images, depth measurement, or point clouds), to both train the NeRF to represent the scene and to optimize its own extrinsic calibration parameters to fit the NeRF representation. In INF [47], the goal is to find the extrinsic transformation between a 360° camera and a LiDAR. First, the density network of NeRF is trained using the LiDAR depth data. Then, the whole scene's radiance is trained using images, while simultaneously calibrating the camera. This method is limited to the calibration of a single 360° camera and a LiDAR, whereas autonomous driving systems rely on multiple cameras with narrower fields of view. AsyncNeRF [40] calibrates a pair of camera and depth sensors. It takes into account the temporal miscalibration between the sensors, by building a trajectory function. Nevertheless, the time offset is provided as input and not determined through optimization, which limits its utilization for spatio-temporal calibration. MOISST [12] proposes to accomplish temporal calibration in addition to extrinsic calibration, and to do so with any number of LiDARs and cameras, by training the NeRF with all the data, while also optimizing the prior extrinsic transformations and time offsets. By using a single NeRF to fuse the information from all the sensors, we cannot prevent degenerate cases where the estimation of the extrinsic parameters of one sensor diverges and causes the NeRF to learn a wrong scene geometry without correlating multi-sensor observations. Our method, SOAC, aims to achieve better robustness and calibration performance by leveraging the use of multiple NeRFs to counterbalance such limitations.

3. Method

Our multi-sensor calibration problem is formulated as follows: given a vehicle trajectory and initial priors of sensor poses mounted on the vehicle, we aim to recover the exact spatio-temporal calibration of the sensors on the vehicle. Our method is composed of two optimization steps that are performed sequentially all along the training (cf. Fig. 1). The first step consists of training multiple implicit scene representations (NeRFs), one by camera, using only the observations from the dedicated sensor. During the second optimization step, we refine the extrinsic and temporal parameters of each sensor using the trained NeRF of all the other sensors in a round-robin manner. The motivation behind this design is to prevent over-fitting, calibration divergence, or implicit model convergence to a poor local minimum when all the observations are fused within the same implicit representation, as in MOISST [12].

3.1. Notations and Background

Without loss of generality, we consider the trajectory of camera r (our reference sensor) as the known trajectory of the vehicle. We use the same notations introduced in MOISST [12] to describe our method:

- $S = \{C, L\}$: the set of sensors composed of at least one or more cameras C and, optionally, one or more LiDARs L,
- $\{F_i\}$: the set of frames captured by the sensor $i \in S$,
- $t^{n_i} \in \mathbf{R}^+$: the timestamp of frame $n_i \in F_i$ relative to the sensor $i \in S$,
- $\delta_i \in \mathbf{R}$: the time offset between the reference camera and the sensor $i \in S$ ($\delta_r = 0$),
- ${}_wT^i(t) \in \mathbf{R}^{4\times4}$: the pose of sensor $i \in S$ at time t (the time is relative to sensor i's own clock) in the world reference frame,
- ${}_jT^i \in \mathbf{R}^{4\times4}$: the transformation matrix from sensor i to sensor j.

Our goal is to find the optimal transformations ${}_r\hat{T}^i$ and time offsets $\hat{\delta}_i$ of the different sensors with respect to the reference camera. The poses of the reference camera r can be obtained by relying on IMU, SLAM [27], or Structure-from-Motion [34]. Similar to MOISST, we build a continuous trajectory of the reference sensor r, $\mathcal{T}_r$, from the discrete poses of r using linear interpolation for the pose translation and spherical linear interpolation (SLERP [35]) for the rotation. This trajectory is expressed as a function of time, that returns the pose of the reference camera r for any given time t: ${}_wT^r(t) = \mathcal{T}_r(t)$. Using the extrinsic transformations and the time offsets between the other sensors

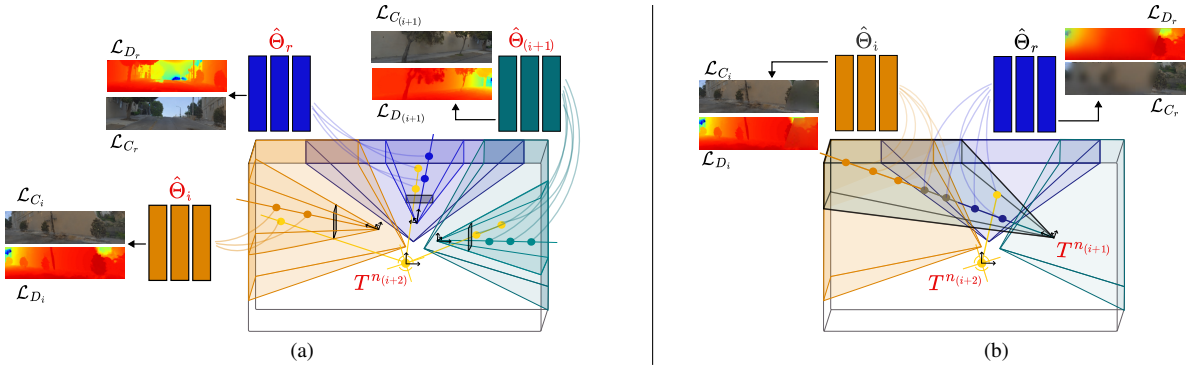


Figure 2. **SOAC training strategy.** (a) Scene representation training (Sec. 3.2): The parameters $\hat{\Theta}$ of each NeRF are trained with the images from their associated cameras and the LiDAR scans. The LiDAR calibration is also optimized through $T^{n_{(i+2)}}$. (b) Extrinsic and temporal optimization (Sec. 3.3): The real frame from the sensor is compared to the predicted frame on the other NeRFs to calculate the losses. The calibration is then optimized with backpropagation through the poses $T^{n_{(i+1)}}$ and $T^{n_{(i+2)}}$.

and the reference camera r, we can compute the absolute pose of sensor i at specific timestamps with the following equation:

$$ {}_{w}T^{i}(t^{n_i} + \delta_i) = \mathcal{T}_r(t^{n_i} + \delta_i)\, {}_{r}T^{i}. \tag{1} $$

In order to simplify the equations, we designate the absolute pose of sensor i computed from its extrinsic as $T^{n_i} = {}_{w}T^{i}(t^{n_i} + \delta_i)$.

NeRF model. NeRF is a function of parameters Θ that takes as input rays obtained from a sensor's intrinsic parameters and pose, and generates for each ray color and density information via volumetric rendering. This information can be combined into a color image $\mathcal{R}_I\left(T^{n_i} \mid \Theta\right)$ and a depth scan $\mathcal{R}_D\left(T^{n_i} \mid \Theta\right)$ of frame n_i for sensor i.

3.2. Scene Representation Training

For each camera sensor, a dedicated NeRF with parameters Θ_i is trained using rays that are generated exclusively from camera i. Each NeRF model with parameters Θ_i will only learn the part of the scene that is observed by its respective camera sensor i (cf. Fig. 2a). The color loss for training the scene representation is:

$$ \mathcal{L}_C = \sum_{i \in C} \sum_{n_i \in F_i} \left\| \mathcal{R}_I\left(T^{n_i} \mid \Theta_i\right) - I^{n_i} \right\|_2^2, \tag{2} $$

with I^{n_i} the color image n_i of camera i. The training objective is to estimate the optimal parameters $\hat{\Theta}_i$ for the NeRF models such as:

$$ \left\{\hat{\Theta}_i\right\}_{i \in C} = \underset{\{\Theta_i\}_{i \in C}}{\operatorname{argmin}} \left(\mathcal{L}_C\right). \tag{3} $$

3.3. Extrinsic and Temporal Optimization

During the calibration step, our objective is to optimize the extrinsic transformation matrix ${}_{r}T^{i}$ and temporal parameters δ_i by optimizing the poses of camera i using all the NeRFs, except the NeRF of parameters Θ_i associated to the current camera being calibrated (cf. Fig. 2b). Using this optimization formulation, we enforce the images captured by each camera to be coherent with the NeRF trained by the other cameras. The camera calibration loss can be written as:

$$ \mathcal{L}_{Cam} = \sum_{j \in C} \sum_{\substack{i \in C \\ i \neq j}} \sum_{n_i \in F_i} \left\| \mathcal{R}_I\left(T^{n_i} \mid \Theta_j\right) - I^{n_i} \right\|_2^2, \tag{4} $$

and by considering Eq. 1, the optimization objective during the spatio-temporal optimization step is:

$$ \left\{ {}_{r}\hat{T}^{i}, \hat{\delta}_i \right\}_{i \in C} = \underset{\{{}_{r}T^{i}, \delta_i\}_{i \in C}}{\operatorname{argmin}} \left(\mathcal{L}_{Cam}\right). \tag{5} $$

3.4. LiDAR Calibration

As LiDARs only provide geometric information, we cannot register an RGB image to a NeRF which was only trained on LiDAR scans. This means that the registration step (cf. Sec. 3.3) could not be accomplished on a LiDAR-trained NeRF. Instead of dedicating a NeRF for each LiDAR, we simultaneously train the camera NeRFs with all the LiDAR scans, and calibrate the LiDARs against all NeRFs (cf. Fig. 2). Thus, we have for both the NeRF training step and calibration step:

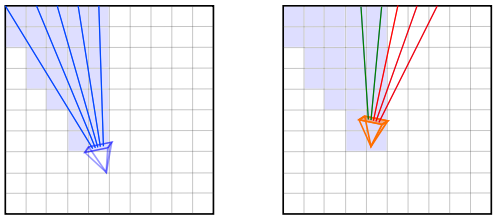
$$ \mathcal{L}_D = \sum_{j \in C} \sum_{i \in L} \sum_{n_i \in F_i} \left| \mathcal{R}_D\left(T^{n_i} \mid \Theta_j\right) - D^{n_i} \right|, \tag{6} $$

with D^{n_i} the point cloud scan n_i of LiDAR i. When adding the LiDAR loss in the objective Eq. 3, it becomes:

$$ \left\{\hat{\Theta}_i\right\}_{i \in C}, \left\{ {}_{r}\hat{T}^{j}, \hat{\delta}_i \right\}_{j \in L} = \underset{\{\Theta_i\}, \{{}_{r}T^{j}, \delta_j\}}{\operatorname{argmin}} \left(\mathcal{L}_C + \mathcal{L}_D\right). \tag{7} $$

and Eq. 5 becomes:

$$ \left\{ {}_{r}\hat{T}^{i}, \hat{\delta}_i \right\}_{i \in S} = \underset{\{{}_{r}T^{i}, \delta_i\}}{\operatorname{argmin}} \left(\mathcal{L}_{Cam} + \mathcal{L}_D\right). \tag{8} $$



(a) (b)

Figure 3. **SOAC's visibility grid** (Sec. 3.5). (a) Grid filling: Rays from camera C_i fill the visibility grid linked to Nerf Θ_i. (b) Ray filtering: For cameras $C_{j,\forall j \neq i}$, rays are kept or filtered according to visibility from (a).

3.5. Visibility Grid

In a multi-sensor setup, the NeRF representation exploits the overlap between sensors w.r.t. the whole sequence rather than a particular frame as for traditional targetless methods. However, the portions of the scene observed from the different sensors might not entirely overlap. This can lead to noisy reconstruction in the NeRF model if inference is performed at the unobserved regions (cf. Fig. 4). To overcome this problem, NeRF2NeRF [10] performs pairwise registration of two NeRF models produced from different viewpoints by aligning the partially overlapping geometry of the two models. In a similar sense, we aim to consider the overlapping geometry from our different NeRFs that have been learned separately from each camera.

To achieve this, a boolean visibility grid for each NeRF model is reconstructed by considering the rays belonging to its akin sensor (see Fig. 3a) during the scene representation step (Sec. 3.2). During the calibration step (Sec. 3.3) we exploit this visibility grid to only consider rays that overlap with trained regions on each NeRF used for registration (cf. Fig. 3b). The grids are reinitialized every few epochs to account for the new poses resulting from the calibration refinements.

3.6. Optimization Details

Overall, the training process can be summarized as follows: during each training step, a mini-batch of rays is first used in the scene representation training step (Sec. 2a). Rays of each camera train their specific NeRF and fill the respective visibility grids (Sec. 3.5). The LiDAR rays train all the NeRFs and are used to optimize the LiDAR calibration parameters after being filtered by the visibility grids. In a subsequent step, the same mini-batch is passed to the extrinsic and temporal optimization. Rays are filtered through the visibility grids before being fed to the NeRFs as explained in Sec. 3.3. Calibration losses (Eq. 8) are computed and the gradient is backpropagated to optimize the calibration parameters. Once this is done, we continue the training with the next mini-batch.

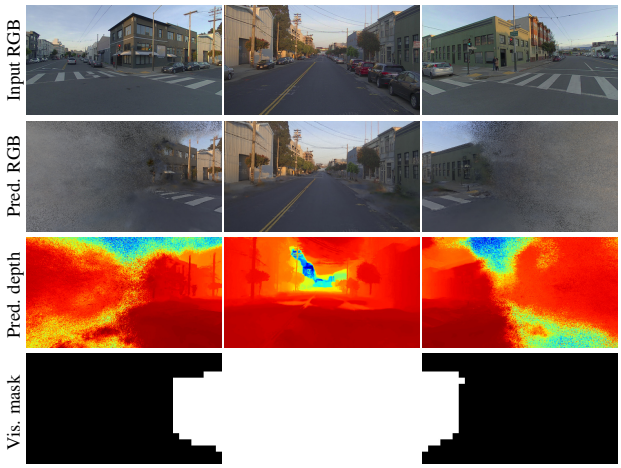


Figure 4. **Visualization of the visibility grid** (Sec. 3.5). Predictions done with the NeRF trained by front camera on a Pandaset [41] sequence.

NeRF delaying. In our system, all the sensors, except the reference camera, have incorrect calibration. As such, the NeRF trained with the reference camera is the most adequate for calibration at the beginning. That is why we introduce a delaying schedule for the other NeRFs based on the overlap with the reference camera; more details about this policy are provided in the supplementary materials.

Correction bounding. As we consider the extrinsic and temporal calibration on a car, we can suppose that the translation error should not be off by more than the car's size. We can also consider that the sensors should not have a time offset too high, even without the help of an external synchronizing system. Thus, by using an offset and scaled sigmoid function on the output of the embeddings for the translation and temporal correction, we can confine the learned correction, avoiding divergence and increasing the stability and robustness of the calibration.

4. Experiments

4.1. Setup

Datasets. We perform experiments on three popular autonomous driving datasets: KITTI-360 [18], nuScenes [3] and Pandaset [41]. For KITTI-360, we use the two front cameras, the two side cameras and the Velodyne LiDAR for our experiments. For nuScenes and Pandaset, we use the front camera, the two front diagonal cameras, and the LiDAR. Undistorted LiDAR scans are considered for all datasets. We assign the front-left camera of KITTI-360, and the front cameras of nuScenes and Pandaset to be the reference sensor. More details on selected sequences and dataset parameters are provided in the supplementary materials.

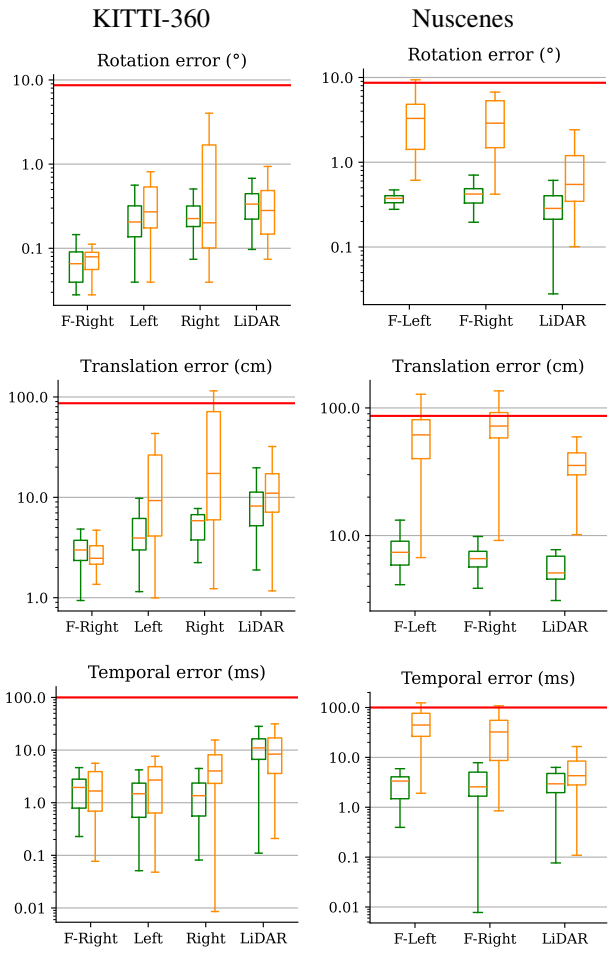


Figure 5. Results for SOAC and MOISST [12] as box plots with log scale on KITTI-360 [18] and Nuscenes [3] sequences. The red lines show the initial error (best viewed in color).

Baseline. We select MOISST [12] as our baseline, as it also aims to solve targetless, multi-modal, and spatiotemporal calibration. We refer to the supplementary for details about the re-implementation of the method. For the LiDAR/Camera calibration task, we compare against LCCNet [20] by using the code and the pre-trained weights from the official repository[1], and with Pandey et al. [28] using the official implementation provided by authors[2]. For SOAC, We replicate MOISST NeRF architecture and apply the same supervision and regularization losses. We refer to the supplementary for more implementation details.

4.2. Results

Spatial and temporal calibration. We run both SOAC and MOISST on 4 KITTI-360 sequences and 3 nuScenes sequences. For SOAC, KITTI-360 images are downscaled by 4, while a downscale factor of 6 is applied for nuScenes.

[1]https://github.com/IIPCVLAB/LCCNet

[2]https://robots.engin.umich.edu/SoftwareData/InfoExtrinsicCalib

	KITTI-360 [18]		Pandaset [41]	
	Rotation (°)	Translation (cm)	Rotation (°)	Translation (cm)
Pandey et al. [28]	11.8 ± 5.4	143 ± 109	15.4 ± 0.8	139 ± 17.5
LCCNet [20]	1.9 ± 0.1	95.8 ± 7.7	14.3 ± 3.4	370 ± 11.6
MOISST [12]	$\mathbf{0.2 \pm 0.1}$	10.0 ± 9.8	2.8 ± 2.3	56.4 ± 17.2
SOAC (ours)	0.3 ± 0.2	$\mathbf{7.8 \pm 3.5}$	$\mathbf{1.3 \pm 0.8}$	$\mathbf{29.4 \pm 13.6}$

Table 2. LiDAR/Camera calibration results.

For MOISST, we do not downscale the KITTI-360 images and apply a downscale factor of 2 for nuScenes as we found that the method performs better with high-resolution images. Each test is run with an initial noise of 50 cm translation error and 5° rotation error on each axis, as well as 100 ms of time offset. We use 10 different seeds to randomly sign the error noises applied and compute the statistics over these 10 runs. Following common practices [32, 36], we show results on Fig. 5 by employing box plots[3] As can be seen, SOAC achieves better calibration results on KITTI-360 with an overall error (average over median for each sensor) of 0.21°, 5.24 cm and 3.95 ms for rotation, translation and time offset, respectively. In contrast, MOISST obtains errors about 10 times higher (i.e. 2.24°, 56.34 cm and 27.07 ms) for the same setup. Detailed quantitative results by sequence are given in the supplementary materials.

LiDAR/Camera calibration. For the task of LiDAR/Camera calibration, the same initial rotation and translation error setup from previous experiments is applied, but without considering any temporal error. We compare our method against LCCNet [20] and Pandey et al. [28]. The provided weights for LCCNet were pretrained on the KITTI odometry dataset [8]. For KITTI-360, We predict the calibration between the front-left camera and the LiDAR. For Pandaset, we predict the calibration between the front camera and the 360° LiDAR. Results are shown in Tab. 2. The performance of LCCNet, is very poor in comparison to SOAC, especially for the translation (results per sequence are provided in the supplementary). As LLCNet is a supervised method, we observe that it is setup-specific, and a slight change in the LiDAR/Camera configuration greatly reduces the performance. This was also highlighted by Fu et al. [7] when using the front-right camera for calibration on the KITTI odometry dataset. For Pandey et al. [28], we were unable to obtain convincing calibration results on the sequences. We argue that feature-based targetless methods are not designed for "in-the-wild" calibration, and sequences need to be acquired in a specific manner to obtain proper results (i.e. indoor, structured environment, dense LiDAR).

[3]The boxes show the first quartile Q_1, median, third quartile Q_3. The whiskers use 1.5 IQR (Interquartile range) above and below the box and stop at a value within the results.

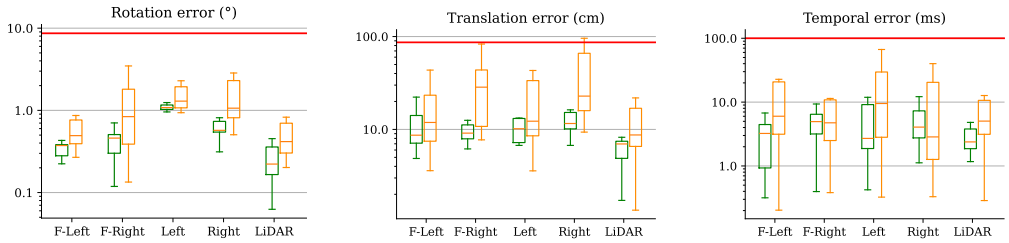


Figure 6. Results on nuScenes [3] with 5 cameras for SOAC and SOAC w/o NeRF delaying as box plots with log scale, the red lines show the initial error (best viewed in color).

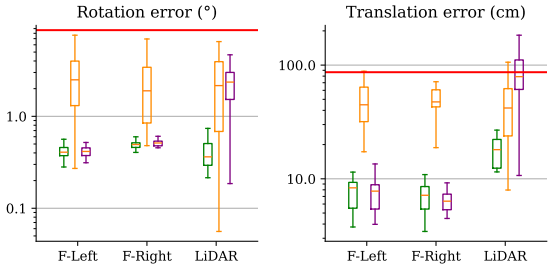


Figure 7. Results on Pandaset [41] for SOAC, MOISST [12] and SOAC w/o semantic filtering as box plots with log scale, the red lines show the initial error (best viewed in color).

Calibration in dynamic environments. For the evaluation in dynamic environments, we select 3 Pandaset sequences with the presence of dynamic elements (e.g. cars, pedestrians). When calibrating on dynamic scenes, the moving elements are not handled by the NeRF model. Therefore, a simple and efficient way of removing these elements is to filter the dynamic classes with semantic segmentation. This results in losing some useful information for calibration (i.e. parked vehicles). Nevertheless, if the rest of the scene provides sufficient overlap, proper calibration can be obtained. We apply an analogous setup to KITTI-360 and nuScenes, except for the removal of temporal calibration and the initial time offset (cf. Sec. 4.3 on Time-space compensation). We downscale the image by a factor of 4 for SOAC and 2 for MOISST. We use semantic segmentation computed by Mask2Former [4] to remove all classes that can be considered dynamic for both methods and test SOAC with and w/out semantic filtering. Results are shown in Fig. 7. It can be observed that by applying semantic filtering, calibration performance on SOAC can be greatly improved on the LiDAR with a median error of 0.41° and 7.79 cm on rotation and translation, respectively, in comparison to results w/out filtering (2.36° / 79.17 cm). It can be also seen that SOAC performs much better than MOISST on the overall calibration of all the sensors with a mean error an of 0.42° / 11.18 cm vs. 2.18° / 44.73 cm for MOISST.

Complete camera rig calibration. To evaluate SOAC performances with a nearly complete 360° camera rig, we add two additional side cameras on the nuScenes sequences. We run both with and without the NeRFs delaying scheduling as explained in Sec. 3.6. In Fig. 6 we can see the impact of not delaying the NeRFs, as the accuracy and stability of the calibration plummet.

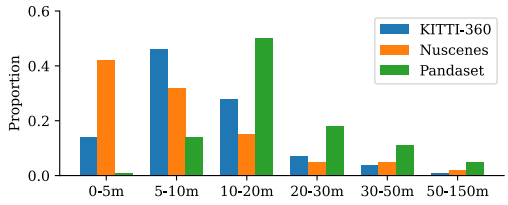


Figure 8. LiDAR ray length distribution of the sequences used in our calibration experiments.

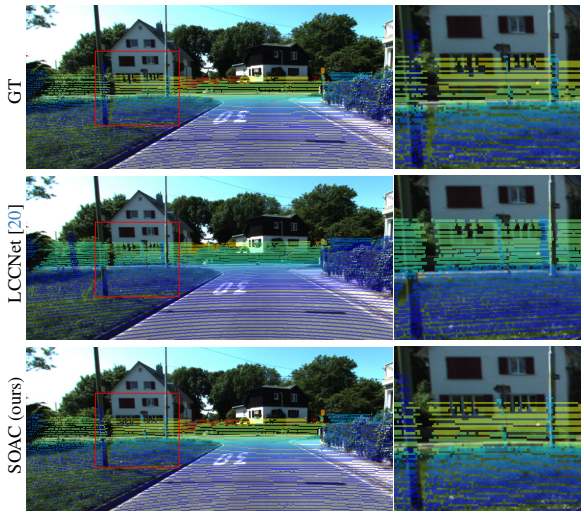


Figure 9. Qualitative LiDAR/front camera calibration results on KITTI-360 [18] dataset.

Qualitative results. We show the reprojection of the LiDAR on the images using the calibration obtained from different methods. On KITTI-360 (cf. Fig. 9) we can see that LCCNet does not provide a satisfying result and that SOAC is able to provide a visually comparable alignment to the ground truth calibration. On nuScenes (cf. Fig. 10), the calibration from SOAC provides a better alignment than MOISST, assessing the quantitative results of Fig. 5. More

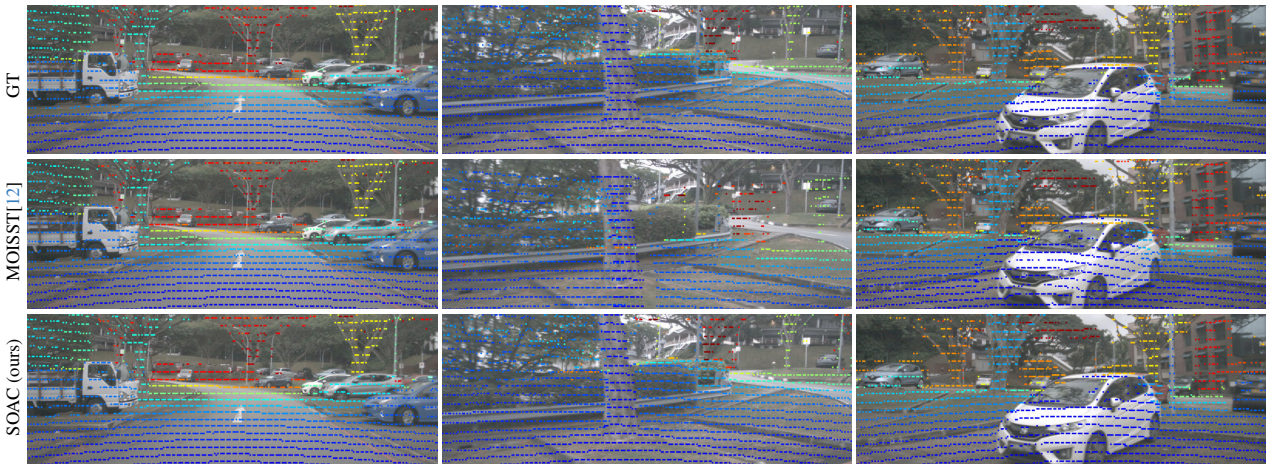


Figure 10. Qualitative LiDAR/Camera reprojection results on nuScenes [3] dataset.

Errors		Cam Front	Cam Left	Cam Right	LiDAR
Translation (cm)	Extrinsic	47.9	67.8	70.4	50.9
	Poses	3.5	2.6	26.2	19.1
Rotation (°)	Extrinsic	0.13	0.18	0.23	0.58
	Poses	1.63	1.12	1.35	0.60
Time offset (ms)		39.18	58.16	40.74	39.38

Table 3. **SOAC space-time compensation on a sequence from KITTI-360 [18].** Mean absolute poses of sensors are correct whereas the spatio-temporal calibration computed by the method is erroneous.

qualitative results are given in the supplementary, along with ablation studies on visibility grids (cf. Sec. 3.5) and correction bounding (cf. Sec. 3.6).

4.3. Limitations

Time-space compensation. When simultaneously calibrating spatially and temporally, there are cases where the disentanglement is impossible. In a sequence where the vehicle is driving in a straight line at a constant speed, there is an infinite number of solutions that can provide the correct poses. In Tab. 3, we show the calibration results on a straight line with constant speed from KITTI-360. We can see the pose error is fairly low, but the extrinsic and temporal calibration is incorrect. This means that there is a need to select scenes with speed variation in order to reduce to a single possible solution. As most Pandaset sequences are in a straight line at a constant speed, we decided to not do temporal calibration on them.

Scene structure. When the scenes are more open and/or larger, the projected rays will have to travel a longer distance before reaching the scene's structures. Considering LiDAR to camera calibration, the rotation error has a linearly increasing impact according to the ray distance when reprojected to the camera frame, while the translation error's impact is independent of the ray distance. Thus, we tend to lose precision on the translation as the ray gets longer. When analyzing the LiDAR rays length distribution of the datasets in Fig. 8, we observe that the LiDAR rays on Pandaset are longer, meaning that the scenes are larger and open, and the structures are farther than on KITTI-360 and nuScenes. This explains most likely the decrease in calibration performances for the LiDAR extrinsic translation parameters on Pandaset (median error of 18.1 cm) compared to KITTI-360 (median error of 8.2 cm) or nuScenes (median error of 5.1 cm).

Training time. As we train one NeRF per camera, and register all the other sensors on each NeRF, the training time increases exponentially with the number of cameras. For instance, on nuScenes one epoch takes approximately 1 minute 45 seconds for 3 cameras and 8 minutes for 5 cameras using the same GPU. This reduces the scalability of our method, but this phenomenon is mitigated by the fact that we use much smaller images than MOISST to reach better performance. We refer to the supplementary for more in-depth details on the efficiency of our method wrt. image size compared to MOISST.

5. Conclusion

In this paper, we presented SOAC, a targetless and self-supervised method for spatial and temporal calibration. This approach is able to simultaneously calibrate multiple sensors of different modalities, by leveraging the use of multiple camera-specific implicit scene representations, and taking into account the overlap between the sensors. Our approach is fully automatic by relying on gradient descent for the optimization process, and surpasses similar methods previously introduced. The reliance on a reference sensor with known trajectory, and the need of near structures for a precise calibration, are restrictions that could open to future research to alleviate them.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, pages 5470–5479, 2022. 2

[2] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *CVPR*, pages 4160–4169, 2023. 3

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 5, 6, 7, 8

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 7

[5] Frank Dellaert, Dieter Fox, Wolfram Burgard, and Sebastian Thrun. Monte carlo localization for mobile robots. In *IEEE international conference on robotics and automation (ICRA)*, pages 1322–1328, 1999. 2

[6] Jamil Fayyad, Mohammad A Jaradat, Dominique Gruyer, and Homayoun Najjaran. Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors*, 20(15):4220, 2020. 1

[7] Lanke Frank Tarimo Fu and Maurice Fallon. Batch Differentiable Pose Refinement for In-The-wild Camera/LiDAR Extrinsic Calibration. In *Conference on Robot Learning (CoRL)*, 2023. 2, 6

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 6

[9] Andreas Geiger, Frank Moosmann, Ömer Car, and Bernhard Schuster. Automatic camera and range sensor calibration using a single shot. In *IEEE international conference on robotics and automation (RA-L)*, pages 3936–3943, 2012. 1, 2

[10] Lily Goli, Daniel Rebain, Sara Sabour, Animesh Garg, and Andrea Tagliasacchi. nerf2nerf: Pairwise registration of neural radiance fields. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9354–9361, 2023. 5

[11] Carlos Guindel, Jorge Beltrán, David Martín, and Fernando García. Automatic extrinsic calibration for lidar-stereo vehicle sensor setups. In *IEEE international conference on intelligent transportation systems (ITSC)*, pages 1–6, 2017. 1

[12] Quentin Herau, Nathan Piasco, Moussab Bennehar, Luis Roldão, Dzmitry Tsishkou, Cyrille Migniot, Pascal Vasseur, and Cédric Demonceaux. MOISST: Multimodal Optimization of Implicit Scene for SpatioTemporal calibration. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. 1, 2, 3, 6, 7, 8

[13] Ganesh Iyer, R Karnik Ram, J Krishna Murthy, and K Madhava Krishna. CalibNet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1110–1117, 2018. 2

[14] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *ICCV*, pages 5741–5751, 2021. 2

[15] Xin Jing, Xiaqing Ding, Rong Xiong, Huanjun Deng, and Yue Wang. DXQ-Net: differentiable lidar-camera extrinsic calibration using quality-aware flow. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6235–6241, 2022. 2

[16] Akio Kodaira, Yiyang Zhou, Pengwei Zang, Wei Zhan, and Masayoshi Tomizuka. SST-Calib: Simultaneous Spatial-Temporal Parameter Calibration between LIDAR and Camera. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 2896–2902, 2022. 2

[17] Xingchen Li, Yuxuan Xiao, Beibei Wang, Haojie Ren, Yanyong Zhang, and Jianmin Ji. Automatic targetless LiDAR–camera calibration: a survey. *Artificial Intelligence Review*, 56(9):9949–9987, 2023. 1

[18] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE TPAMI*, 45(3):3292–3310, 2022. 5, 6, 7, 8

[19] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, 2021. 2

[20] Xudong Lv, Boya Wang, Ziwen Dou, Dong Ye, and Shuo Wang. LCCNet: LiDAR and camera self-calibration using cost volume network. In *CVPRW*, pages 2894–2901, 2021. 2, 6, 7

[21] Dominic Maggio, Marcus Abate, Jingnan Shi, Courtney Mario, and Luca Carlone. Loc-nerf: Monte carlo localization using neural radiance fields. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4018–4025, 2023. 2

[22] Enrique Marti, Miguel Angel De Miguel, Fernando Garcia, and Joshue Perez. A review of sensor technologies for perception in automated driving. *IEEE Intelligent Transportation Systems Magazine (ITSM)*, 11(4):94–108, 2019. 1

[23] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *ICCV*, pages 6351–6361, 2021. 3

[24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[25] Arthur Moreau, Nathan Piasco, Moussab Bennehar, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. CROSSFIRE: Camera Relocalization On Self-Supervised Features from an Implicit Representation. *ICCV*, 2023. 2

[26] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4):1–15, 2022. 2

[27] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics (T-RO)*, 31 (5):1147–1163, 2015. 3

[28] Gaurav Pandey, James McBride, Silvio Savarese, and Ryan Eustice. Automatic targetless extrinsic calibration of a 3d lidar and camera by maximizing mutual information. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2012. 2, 6

[29] Chanoh Park, Peyman Moghadam, Soohwan Kim, Sridha Sridharan, and Clinton Fookes. Spatiotemporal camera-LiDAR calibration: A targetless and structureless approach. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):1556–1563, 2020. 1, 2

[30] Zoltan Pusztai and Levente Hajder. Accurate calibration of LiDAR-camera systems using ordinary boxes. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 394–402, 2017. 1

[31] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021. 3

[32] Joern Rehder, Paul Beardsley, Roland Siegwart, and Paul Furgale. Spatio-temporal laser to visual/inertial calibration with applications to hand-held, large scale scanning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 459–465, 2014. 6

[33] Nick Schneider, Florian Piewak, Christoph Stiller, and Uwe Franke. RegNet: Multimodal sensor registration using deep neural networks. In *IEEE intelligent vehicles symposium (IV)*, pages 1803–1810, 2017. 2

[34] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, pages 4104–4113, 2016. 3

[35] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985. 3

[36] Zachary Taylor and Juan Nieto. Motion-based calibration of multimodal sensor extrinsics and timing offset estimation. *IEEE Transactions on Robotics (T-RO)*, 32(5):1215–1229, 2016. 1, 6

[37] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *CVPR*, pages 4190–4200, 2023. 3

[38] Ruisheng Wang, Frank P Ferrie, and Jane Macfarlane. Automatic registration of mobile LiDAR and spherical panoramas. In *CVPRW*, pages 33–40, 2012. 2

[39] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[40] Zirui Wu, Yuantao Chen, Runyi Yang, Zhenxin Zhu, Chao Hou, Yongliang Shi, Hao Zhao, and Guyue Zhou. Async-NeRF: Learning Large-scale Radiance Fields from Asynchronous rgb-d Sequences with time-pose function. *arXiv preprint arXiv:2211.07459*, 2022. 2, 3

[41] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3095–3101, 2021. 5, 6, 7

[42] Chao Ye, Huihui Pan, and Huijun Gao. Keypoint-based LiDAR-camera online calibration with robust geometric network. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2021. 2

[43] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting Neural Radiance Fields for Pose Estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330, 2021. 2

[44] Chongjian Yuan, Xiyuan Liu, Xiaoping Hong, and Fu Zhang. Pixel-level extrinsic self calibration of high resolution lidar and camera in targetless environments. *IEEE Robotics and Automation Letters (RA-L)*, 6(4):7517–7524, 2021. 2

[45] Jiahui Zhang, Fangneng Zhan, Yingchen Yu, Kunhao Liu, Rongliang Wu, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Pose-Free Neural Radiance Fields via Implicit Pose Regularization. In *ICCV*, 2023. 3

[46] Qilong Zhang and Robert Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2301–2306, 2004. 1, 2

[47] Shuyi Zhou, Shuxiang Xie, Ryoichi Ishikawa, Ken Sakurada, Masaki Onishi, and Takeshi Oishi. INF: Implicit Neural Fusion for LiDAR and Camera. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. 2, 3