Supplementary for "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

A. Technical Details

A.1. Datasets

KITTI-360 [4]: Sequences are selected and cropped by considering vehicle speed variations to remove time-space compensation issues as described in main article Sec. 4.3. Once sequences are cropped, one out of two frames are kept for all sensors to obtain a total of 40 frames per sequence. This decision was made to match the same length as the NVS benchmark sequences present on the dataset. The details from each sequence are summarized in Tab. 1.

Sequence	KITTI-360 run	Starting frame	Ending frame
1	0009	980	1058
2	0009	2854	2932
3	0010	3390	3468
4	0002	4722	4800
Straight line	0009	220	298

Table 1. Selected frames for each KITTI-360 [4] sequence.

nuScenes [2]: Since nuScnes poses are provided only in $\mathrm{SE}(2)$, they cannot be used directly for our method. Instead, we use KISS-ICP [8] to get a good estimate of the LiDAR poses. Extrinsic calibration provided by the dataset is then used to obtain the poses for all cameras. We select the sequences 916, 410 and 417 for our experiments, as they are more suitable for the calibration (closer structures, more speed variation). All LiDAR scans are used during calibration as the LiDAR is sparser than the one in KITTI-360, while one out of two images is subsampled to reduce training time.

Pandaset [10]: Since extrinsic parameters are not provided by the dataset, they are estimated using the global poses of all sensors at several frames by calculating the transformation between the frames with the same timestamp from each sensor. Sequences 33, 40 and 53 are used for our experiments as they have more close structures. We apply the same subsampling strategy as for nuScenes.

A.2. Architecture and Losses

For our NeRF network architecture, we use the same model as MOISST [3] which is inspired by the `nerfacto` model of Nerfstudio[1] open source project. It uses the combination of two papers. The first one is the proposal network from MipNeRF-360 [1] with two proposal networks for the coarse density estimation and a final NeRF for the radiance and the fine density, improving the geometry of the scene, the rendering quality and reducing the training time. The second one is the hash grid introduced by instant-NGP [5] to replace the deterministic positional encoding, which also accelerates the training. Following the nerfacto implementation, 128 points (instead of 256) per ray are sampled for the first proposal model, 96 points for the second one, and 48 points for the final NeRF model, which outputs our results.

On top of $\mathcal{L}_C$, $\mathcal{L}_{Cam}$ and $\mathcal{L}_D$, two losses for geometric consistency, also used by MOISST, are added: a structural dissimilarity (DSSIM) loss $\mathcal{L}_{SSIM}$ [9], and a depth smoothness loss $\mathcal{L}_{DS}$ from RegNeRF [6].

A.3. Hyperparameters

In Tab. 3 are indicated the hyperparameters used for the training of SOAC, and in Tab. **??** are the NeRF delaying epochs depending on the dataset. Delaying the NeRF proves advantageous in scenarios characterized by a multitude of sensors, some of which exhibit minimal overlap with the reference sensor throughout the sequence. This approach facilitates the accurate propagation of calibration

[1]https://docs.nerf.studio/en/latest/nerfology/methods/nerfacto.html

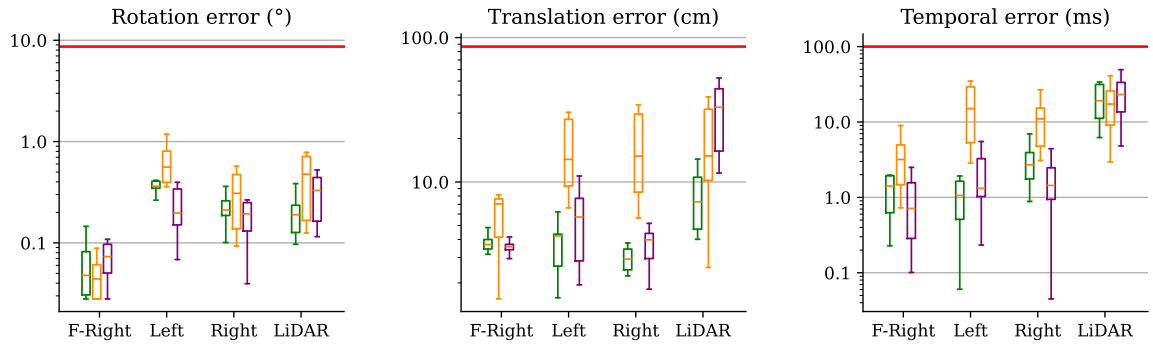

Figure 1. Ablation results on KITTI-360 [4] sequence 4: for SOAC, SOAC w/o Sigmoid and SOAC w/o visibility grid as box plots with log scale, the red lines show the initial error (Best viewed in color).

Hyperparameter	Value
Number of epochs	20
Initial calibration lr	1e-3
Final calibration lr	1e-4
Visibility grid size	20
Batch size	100
Patch size	[15, 15]
$\mathcal{L}_C$ coef	1
$\mathcal{L}_{Cam}$ coef	1
$\mathcal{L}_{SSIM}$ coef	0.1
$\mathcal{L}_D$ coef	1
$\mathcal{L}_{DS}$ coef	1e-4
Translation bounding	2 meters
Temporal bounding	500 ms

Table 2. SOAC hyperparameters used for the training.

Sensor	KITTI-360 [4]	nuScenes [2]	Pandaset [10]
Diagonal cams	-	1	3
Side cams	1	9	-
LiDAR	6	5	8

Table 3. SOAC hyperparameters used for the training.

information during training from sensors presenting significant overlap with the reference sensor to those with lesser or no overlap at all. Basically, larger overlaps and larger quantities of data reduce the number of necessary delay epochs. The number of epochs for training MOISST is reduced to 20, as improvement was not observed with more. The spatial and temporal optimization learning rate is fine-tuned to 5e-4.

B. Additional ablations

Correction bounding. The addition of the sigmoid for bounding the translation and temporal corrections allows better stability and robustness as shown in Fig. 1 on which a huge decrease in calibration accuracy can be noticed when removing the sigmoid.

Dataset	MOISST [3]	SOAC
KITTI-360 [4]	$\sim$ 2 h 30 min	$\sim$ 1 h 30 min
Nuscenes [2] (3 cams)	$\sim$ 2 h 30 min	$\sim$ 1 h
Nuscenes [2] (5 cams)	$\sim$ 4 h 30 min	$\sim$ 2 h 30 min
Pandaset [10]	$\sim$ 1 h 30 min	$\sim$ 1 h 20 min

Table 4. Training time comparison on different sequences.

Downscale factor	Calibration error(°/cm/ms) SOAC	Calibration error(°/cm/ms) MOISST [3]	Training time (min) SOAC	Training time (min) MOISST [3]
1	0.2/4.6/3.9	0.1/5.3/1.3	605	163
2	0.3/4.6/2.5	0.3/24.1/5.3	181	42
4	0.2/4.6/1.7	1.1/41.4/13.1	85	17
8	0.4/12.3/8.2	2.6/56.6/28.6	53	12

Table 5. Training time and calibration accuracy for varying downscale factor on KITTI-360 [4] sequence 1 seed 0

Visibility grid. Removing the visibility grids deteriorates the performance of the LiDAR calibration rotation and translation as shown in Fig. 1.

C. Training time

We report the mean training times with both SOAC and MOISST for the sequences from each tested dataset in Tab. 4. For all the experiments, we used a GPU of similar performance to an RTX 3090. The shown results are with the downscaled images as described in the paper. SOAC is able to provide better calibration than MOISST with shorter training time, even if multiple NeRFs are used, as it can use much smaller images. To measure the impact of the image downscale factor in relation to each method's training time, we train both methods at different downscale factors and report results on Tab. 5. As it can be observed, MOISST accuracy is considerably harmed by using lower-resolution images in comparison to SOAC. Furthermore, SOAC achieves high accuracy even with large downscale factors on the images (i.e. downscaling the image resolu-

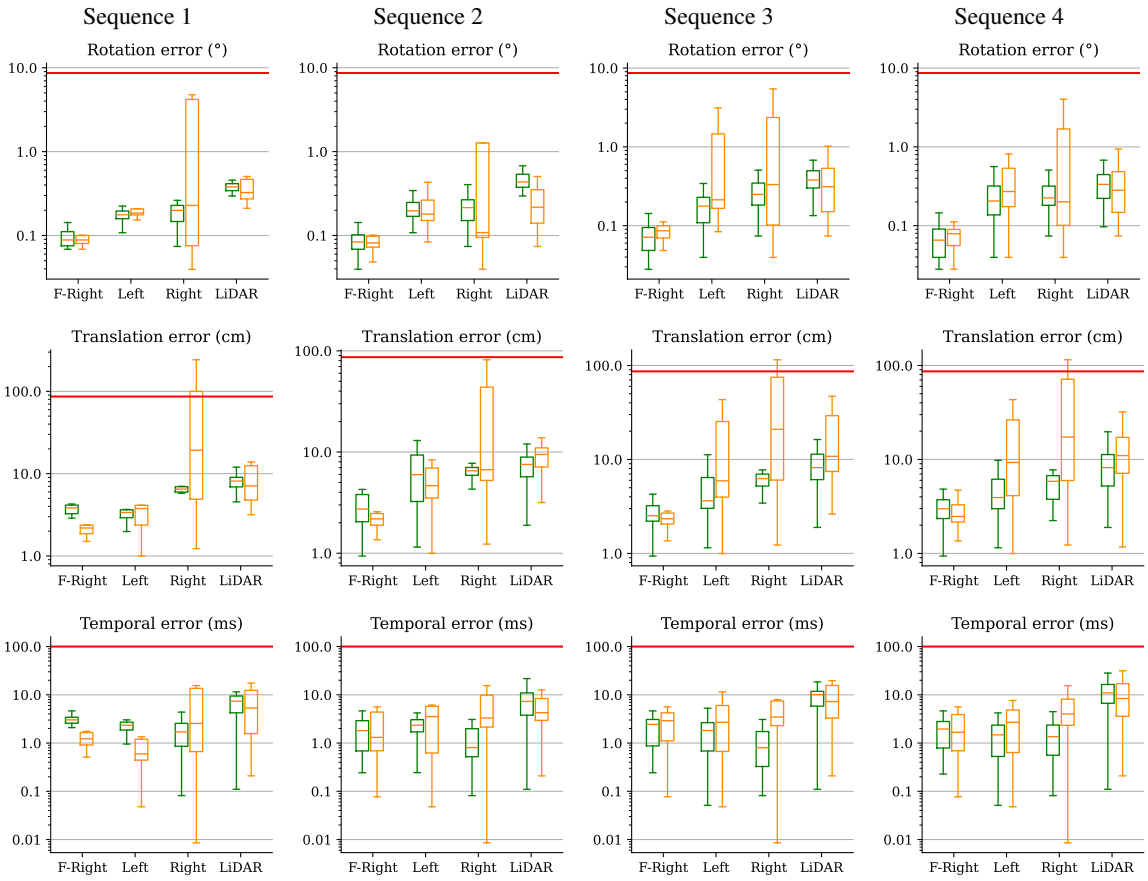

Figure 2. Results for KITTI-360 [4] per sequence for SOAC and MOISST [3] as box plots with log scale. The red lines show the initial error (best viewed in color).

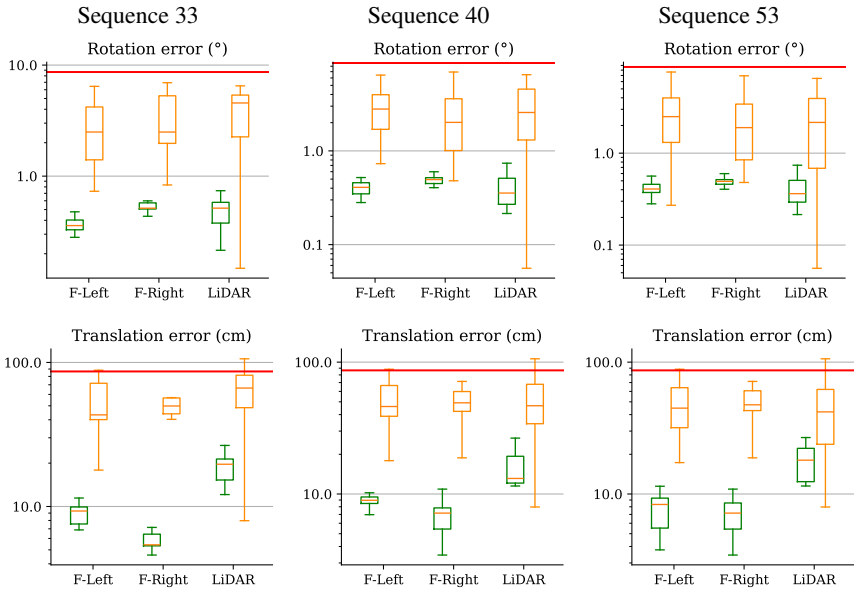

Figure 3. Results for Pandaset [7] per sequence for SOAC and MOISST [3] as box plots with log scale. The red lines show the initial error (best viewed in color).

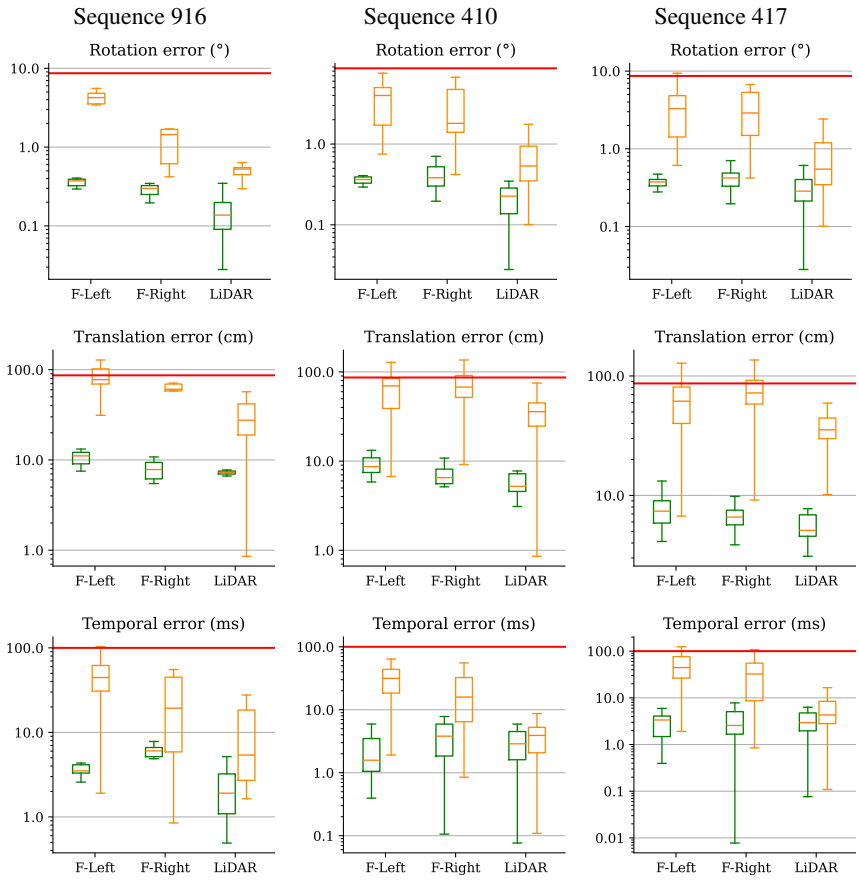

Figure 4. Results for nuScenes [2] per sequence for SOAC and MOISST [3] as box plots with log scale. The red lines show the initial error (best viewed in color).

tion by 4 shows no drop in accuracy for SOAC while being 8 times faster. In comparison, MOISST presents a severe drop in performance when downscaling). This enables SOAC to achieve more efficient training times given its ability to exploit lower-resolution images.

D. Quantitative results

Specific box plot results are provided for each sequence. The results for KITTI-360 are in Fig. 2, the results for Nuscenes in Fig. 4, and the results for Pandaset in Fig. 3.

On KITTI-360, MOISST seems to provide results on par with SOAC on the Front-right camera and the LiDAR. However, on the side cameras, there is a significant difference in the stability of the calibration. On Nuscenes and Pandaset, SOAC is much more precise and stable than MOISST all across the board.

E. Qualitative results

In Fig. 6 and Fig. 5 are shown LiDAR/Camera projection on nuScenes and Pandaset sequences. The calibration optimized by SOAC provides substantially better alignment than the one from MOISST.

Fig. 7 shows the predicted images and masks from each NeRF trained with different cameras. The visibility masks are coherent with the predicted RGB images, allowing correct filtering for SOAC.

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. 1

[2] 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. 1, 2, 4, 5

[3] 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

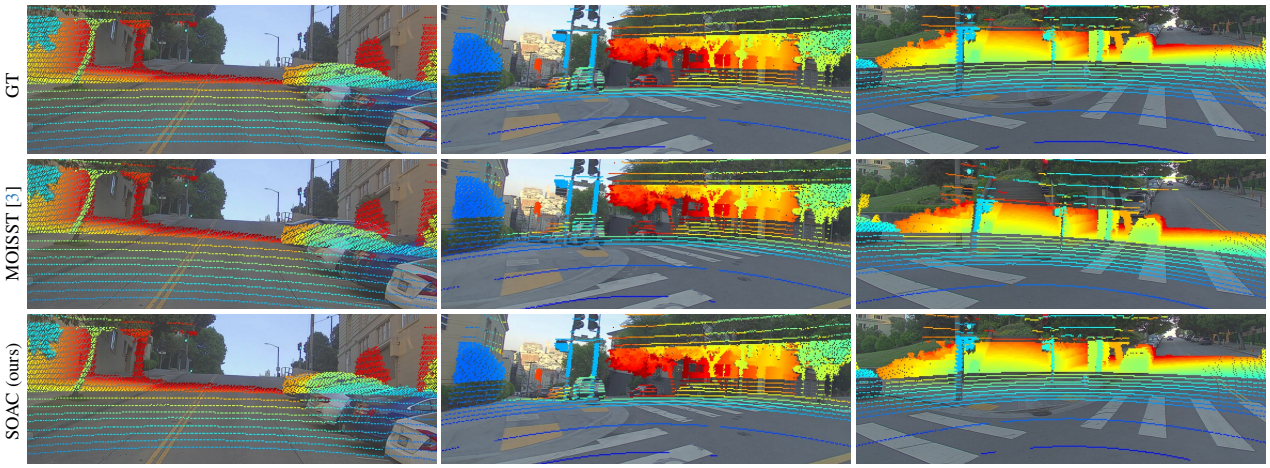

Figure 5. Qualitative LiDAR/Camera reprojection results on Pandaset [10] dataset.

Figure 6. More qualitative LiDAR/Camera reprojection results on nuScenes [2] dataset.

Figure 7. Results using visibility grids on a Pandaset [10] sequence – Prediction from different NeRFs.

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2023. 1, 2, 3, 4, 5

[4] 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. 1, 2, 3

[5] 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. 1

[6] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, pages 5480–5490, 2022. 1

[7] 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. 3

[8] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. Kiss-icp: In defense of point-to-point icp–simple, accurate, and robust registration if done the right way. *IEEE Robotics and Automation Letters (RA-L)*, 8(2):1029–1036, 2023. 1

[9] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 1

[10] 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. 1, 2, 5, 6