PseudoCal: Towards Initialisation-Free Deep Learning-Based Camera-LiDAR Self-Calibration

Mathieu Cocheteux
mathieu.cocheteux@hds.utc.fr
Julien Moreau
julien.moreau@hds.utc.fr
Franck Davoine
franck.davoine@hds.utc.fr

Université de technologie de Compiègne
CNRS, Heudiasyc Laboratory
Compiègne, France

Abstract

Camera-LiDAR extrinsic calibration is a critical task for multi-sensor fusion in autonomous systems, such as self-driving vehicles and mobile robots. Traditional techniques often require manual intervention or specific environments, making them labour-intensive and error-prone. Existing deep learning-based self-calibration methods focus on small realignments and still rely on initial estimates, limiting their practicality. In this paper, we present PseudoCal, a novel self-calibration method that overcomes these limitations by leveraging the pseudo-LiDAR concept and working directly in the 3D space instead of limiting itself to the camera field of view. In typical autonomous vehicle and robotics contexts and conventions, PseudoCal is able to perform one-shot calibration quasi-independently of initial parameter estimates, addressing extreme cases that remain unsolved by existing approaches.

1 Introduction

Camera to LiDAR extrinsic calibration is crucial for enabling seamless sensor fusion and comprehensive environmental understanding in autonomous systems such as self-driving vehicles and mobile robots. The objective of this task is to determine the 6D transformation T between the coordinate systems of a camera and a LiDAR (that is, rotation and translation). Calibration techniques based on traditional vision methods achieved accurate results [1, 2, 9, 11, 20, 24, 30, 31], but often require labour-intensive manual procedures, specific environments, or targets. This leads to potential inaccuracies and inefficiencies. Although deep learning-based calibration methods [6, 12, 13, 16, 17, 25, 29] have emerged as powerful alternatives, they are limited by their dependence on initial parameter knowledge (approximation of T).

To address these challenges, we propose PseudoCal, a novel sensor calibration method that capitalises on the pseudo-LiDAR concept [27]. It enables accurate and efficient calibration independent of initial parameter knowledge. Existing deep learning-based methods rely

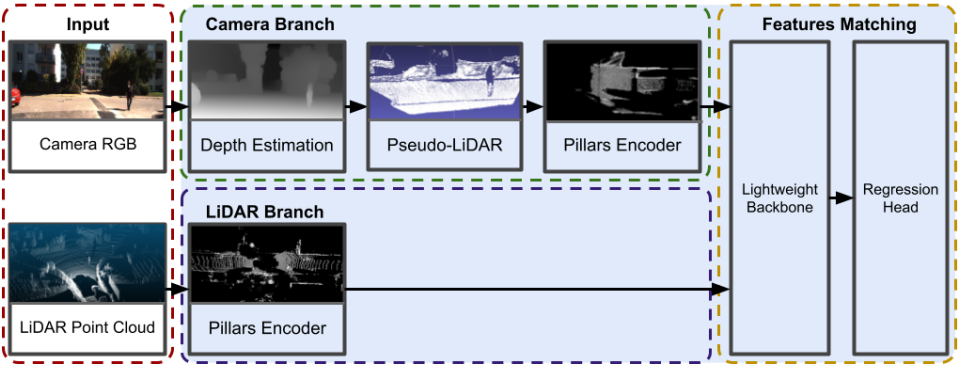


Figure 1: Illustration of the PseudoPillars module, key component of the PseudoCal method allowing for calibration estimation quasi-independently from initial values.

on a LiDAR projection in the camera image, based on an initial approximate knowledge of the extrinsic calibration parameters. Thus, they depend on the availability and accuracy of this initial knowledge, and discard most of the LiDAR point cloud, which is not projected into the camera field of view. By leveraging the pseudo-LiDAR concept into the calibration process, PseudoCal is able to work directly in 3D space, dismissing the reliance on initial parameter knowledge.

In autonomous vehicle and robotics contexts, calibration without an initial estimate is critical. It facilitates on-the-fly recalibration, essential when initial parameters are unavailable or post-mechanical stress like accidents or maintenance, as well as in cases where a robot might be physically unreachable. This adaptability ensures the reliability and precision of the fused sensor data. Furthermore, it accelerates the integration of new sensors, eliminating the need for laborious manual calibration, thus promoting faster adaptation within the autonomous vehicle ecosystem.

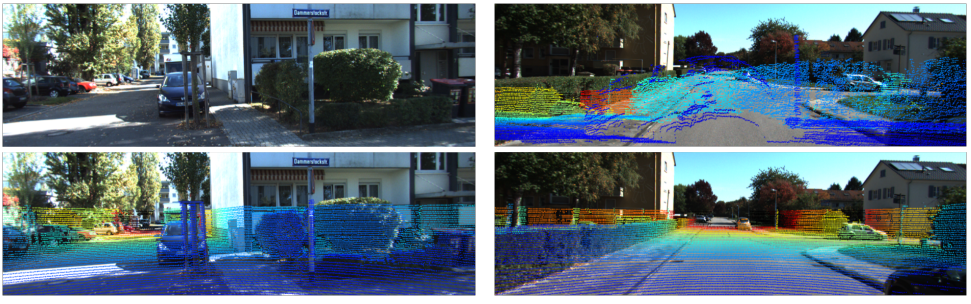
The main contribution of this paper lies in a novel camera-to-LiDAR self-calibration technique that effectively leverages the pseudo-LiDAR concept through our proposed PseudoPillars module, as depicted in Figure 1. This method is able to perform calibration quasi-independently of initial parameter estimates in typical autonomous vehicle and robotic contexts, addressing extreme cases that remain unsolved by existing approaches (refer to Figure 2 for examples).

2 Related Work

In this section, we provide an overview of calibration methods, focusing on deep learning-based techniques, and highlighting the similarities and differences with our approach. We also discuss the development of monocular depth estimation and pseudo-LiDAR, emphasising their relevance to our work.

2.1 Automatic Camera-LiDAR Calibration Methods

Classical computer vision approaches have been used to address this task and have achieved satisfactory accuracy. However, these methods have drawbacks, such as requiring a target [1,



(a) Exemple of a pitch decalibration $> 20°$ (b) Exemple of a yaw decalibration $\approx 90°$

Figure 2: Illustration of the method on driving scenes from KITTI. Points are a depth-coloured LiDAR projection. On top are decalibrated samples in two extreme scenarios where existing deep learning-based methods fail. Bottom visuals represent PseudoCal's correction. (a) shows no LiDAR-camera overlap due to severe pitch axis decalibration, and (b) presents a severe yaw axis decalibration which results in projected points corresponding to objects outside of the camera field of view.

9, 32], specific environment features [30], and for most of them offline and relatively long computation (seconds to minutes)[1, 9, 11, 30]. Deep learning-based calibration methods have emerged as powerful tools for addressing sensor calibration challenges due to their ability to capture complex relationships between sensor modalities while using larger parts of the scene than target-based methods. Examples of such methods include RegNet [25], CalibNet [12], SemAlign [16], LCCNet [17], DXQ-Net [13], and UniCal [6]. Each of these methods has made significant contributions. RegNet [25] was the first work to address this task with a deep learning approach that matches the camera image and projected LiDAR, with parameters refined in a cascaded architecture. DXQ-Net [13] introduced a differentiable pose estimation module and probabilistic modelling of the task, improving accuracy and generalisation. A more recent approach, UniCal [6], leverages a Transformer[26]-based backbone network to bring attention mechanisms to calibration. It achieves state-of-the-art results with a lighter single-branch architecture.

However, these methods share a common drawback. They rely on a good initial guess of extrinsic parameters, which may not always be available. Therefore, we propose with PseudoCal a procedure to get an accurate calibration quasi-independently from the initial parameters. Details on the architecture are given in Section 3.1.

2.2 Exploiting a 2D camera for 3D information

Pseudo-LiDAR The pseudo-LiDAR concept has emerged as a key technique for bridging the performance gap between image-based and LiDAR-based 3D object detection. It consists in using a depth map obtained from 2D sensors to generate a point cloud by projecting its points in a 3D space. Wang *et al.* [27] first demonstrated significant performance improvements on image-based 3D detection by converting stereo-based depth maps into pseudo-LiDAR representations. Weng and Kitani [28] then proposed to generate the depth map from a single camera by monocular depth estimation. Our work capitalises on this approach to generate a point cloud from monocular camera images.

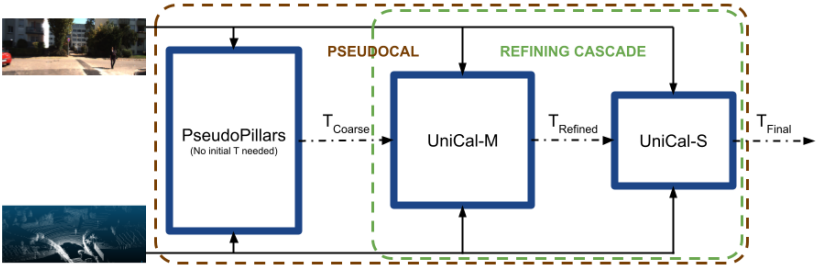


Figure 3: Overview of the PseudoCal architecture and its three main components: PseudoPillars, UniCal-M, and UniCal-S. PseudoPillars provides an initial estimation to the cascaded UniCal modules for refinement. The 6D transformation (calibration parameters) is noted T.

In network feature extraction, point cloud representation varies. Some, like [21, 27], use unordered points, while others prefer to voxelise it [22]. We adopt the approach of Lang *et al*. [15], generating Pillars features from voxels, depicted as a 2D pseudo-image. This method is memory efficient, computationally effective, and suited for sparse or semi-dense point clouds, such as those from LiDARs or pseudo-LiDARs.

Monocular Depth Estimation Monocular depth estimation is a broadly researched computer vision task. State-of-the-art methods such as [10, 14, 19, 23] now demonstrate reliable and accurate results. These approaches take advantage of advanced techniques and architectures, such as Vision Transformers [23], to capture rich contextual information and model complex scene structures effectively. Moreover, some methods, such as the one proposed by Godard *et al*. [10], explore unsupervised learning strategies, reducing the need for large-scale annotated depth datasets and further broadening the applicability of monocular depth estimation. These methods have become more reliable and can now be used as critical components for complex vision pipelines. We use a recent model, Global-Local Path Networks (GPLN) [14], as a base component of our PseudoPillars module to provide depth estimation for generating a pseudo-LiDAR projection. GPLN relies on a Transformer-based architecture to capture the global context of the image while using a novel decoder to consider local connectivity.

3 Method

In this section, we present the methodology of the proposed PseudoCal approach, which consists of our novel PseudoPillars network, followed by a cascade of two UniCal [6] networks.

3.1 Architecture

3.1.1 Cascaded Architecture Rationale

The PseudoCal architecture adopts a cascaded structure illustrated in Figure 3, comprising the proposed PseudoPillars module followed by two UniCal [6] modules trained on decreasing decalibration ranges as described in Table 1.The PseudoPillars module performs a coarse estimation of the calibration, while the UniCal modules sequentially refine the estimation.

Module	Training decalibration range	
	Rotation (°)	Translation (cm)
PseudoPillars	30, 30, 180	150
UniCal-M	10, 10, 10	100
UniCal-S	1, 1, 1	10

Table 1: Training decalibration range for each cascaded module in PseudoCal. Translation range is the same on all axes, while rotation is respectively for roll, pitch, and yaw axes.

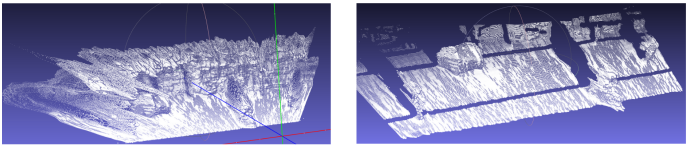


Figure 4: Kitti scene without (left) and with (right) edge removal.

This design choice is motivated by the success of cascaded architectures in various computer vision tasks, including object detection [3], pose estimation [5], and even camera to LiDAR calibration [25]. The cascading approach enables PseudoCal to achieve state-of-the-art calibration results quasi-independently from initial parameters.

3.1.2 The PseudoPillars module

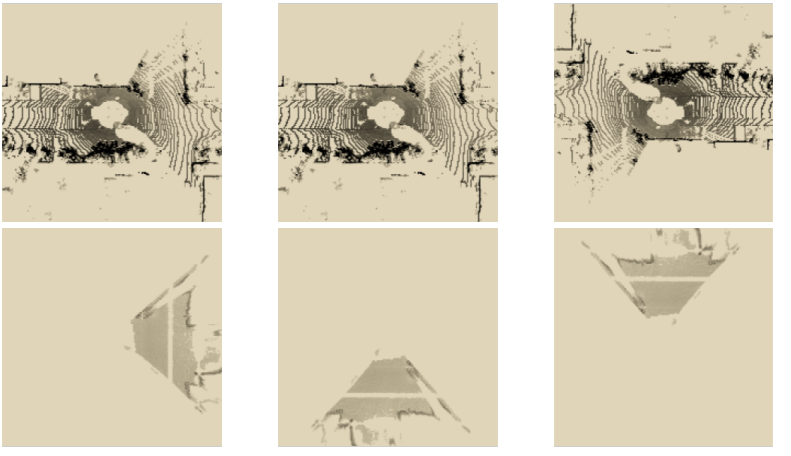
The PseudoPillars module described in Figure 1 is the core component of our PseudoCal architecture. It comprises a depth estimation network, a pseudo-LiDAR projection, a Pillars [15] encoder, a matching lightweight backbone (here MobileViT), and a regression head similar to the one used in [6].

The depth estimation network is based on GLPN [14], which has shown state-of-the-art performance in monocular depth estimation. The estimated depth map is then converted into a pseudo-LiDAR point cloud. This is done by projecting the depth-encoded pixels from the depth map to the 3D space, considering the camera intrinsics. The 3D points coordinates (x,y,z) are given by:

$$z = D(u,v), \quad x = \frac{(u-c_U)\times z}{f_U}, \quad y = \frac{(v-c_V)\times z}{f_V} \tag{1}$$

where D is the depth map, (u,v) the pixel coordinates, (c_U,c_V) the camera center, and(f_U,f_V) the focal length along U and V axes respectively.

This 3D back-projection results in unwanted artefacts on objects' edges, affecting the output of our network. More specifically, trails of points making the junction between 3D objects on different depth, which have already been observed in [27]. Noticing that they are caused by gradient irregularities (inherent to its 2D representation) on the edges in the depth map, we found a simple yet efficient way to filter them. We use the Canny edge detection [4] algorithm to find and remove edges in the depth map before the 3D back-projection, resulting in a clean point cloud, as illustrated in Figure 4. Other methods such as statistical 3D neighbourhood filtering have been considered but have produced lesser results. Other filters could be considered in future research. To efficiently match information between pseudo-LiDAR and LiDAR, a representation robust to point cloud density is needed. Thus, both



(a) Calibrated (b) Decalibrated (c) Decalibrated and augmented

Figure 5: Visualization of the Pillars representation of a KITTI scene for the LiDAR (top) and the camera (bottom). (a) is calibrated, (b) is decalibrated by 90° on yaw for the illustration, (c) has the same decalibration and a 180° yaw augmentation.

pseudo-LiDAR and LiDAR point clouds are passed through Pillars encoders, a technique inspired by the Pillar Feature Network used in LiDAR-based 3D object detection [15]. The Pillars encoder generates a compact and efficient feature representation of the point cloud, of which sampling and density can be tuned. Figure 5 illustrates the Pillars representation obtained from both modalities and highlights their common information.

Both are then jointly passed through a MobileViT [18] backbone, which is responsible for extracting useful features for the regression head. MobileViT is a variant of the Vision Transformer [7] architecture. It has shown impressive performance in various computer vision tasks [18], including camera to LiDAR calibration [6]. It has the advantage of being lightweight, fast, and to leverage Transformers and convolutional operations, which allows it to perceive global information in the image, while also considering local features. The regression head then estimates the extrinsic parameters. It is composed of a common dense layer which then forks to two other dense layers to output rotation and translation parameters.

3.1.3 Refinement cascade

Following the PseudoPillars module, we cascade two UniCal [6] modules to refine the calibration estimation, as illustrated in Figure 3. UniCal is a model which calibrates a camera-LiDAR pair given an approximate initial calibration, specialized on correcting small decalibrations. Unlike other methods, it fuses camera and LiDAR data early in the process, aggregating image channels and LiDAR mappings into a unified representation for joint feature extraction. This approach results in state-of-the-art performance while offering a lightweight solution ideal for resource-constrained applications. In PseudoCal, the UniCal modules are trained on two different ranges of decalibrations on which they will specialize. These are described in Table 1 and noted respectively UniCal-M for the medium range ($\pm 10^\circ$ and $\pm 100cm$), and UniCal-S for the smaller range ($\pm 1^\circ$ and $\pm 10cm$). As motivated

in Section 3.1.1, by organising the UniCal modules in a cascade, we can refine the coarse output from PseudoPillars to achieve state-of-the-art results reported in Table 2 while being quasi-independent from initial parameters. Ablation studies presented in Section 4.2.3 and Table 3 support the choice of using exactly two refining UniCal modules.

3.2 Training Strategy

3.2.1 Loss Function

Our training strategy involves a combination of translation, rotation, and spatial losses to effectively train the PseudoCal network. Translation and rotation losses $\mathcal{L}_t$ and $\mathcal{L}_r$ ensure that the estimated extrinsic parameters are accurate, while spatial losses, inspired by NetCalib [29] and CalibNet [12], take into account geometric information to achieve better learning convergence.

For translation and rotation losses, we use the mean squared error (MSE) between the predicted and ground-truth values. These two losses are balanced with appropriate weighting factors to account for the difference in their magnitudes. The first spatial loss, $\mathcal{L}_{pcl}$, measures the average distance between corresponding points in the predicted and ground-truth point clouds, which is correlated to the rotation error. The second spatial loss, $\mathcal{L}_C$, is the distance between the centroids of the predicted and ground-truth point clouds, which is correlated with the translation error. The global loss is defined in Equation 2 where weights are defined as $\alpha = 1.3,\ \beta = 1.3,\ \gamma = 1,\ \delta = 1.75$.

$$\mathcal{L} = \alpha \times \mathcal{L}_t + \beta \times \mathcal{L}_r + \gamma \times \mathcal{L}_{pcl} + \delta \times \mathcal{L}_C \tag{2}$$

3.2.2 Samples Generation

During training, two complementary processes are applied to prevent the model from learning biases. First, artificial decalibration changes the initial calibration parameters (relative 3D transformation). It is used to generate different samples to train the network (within the specified ranges). It helps to generalize, as the number of real setups available is limited (KITTI setup in our case). On the other side, augmentation is an absolute 3D transformation applied on both sensors (no change of target calibration parameters). Augmentation is needed with PseudoPillars to generalise to any orientation of the sensors in its internal 3D space (Pillars representation).

Artificial Decalibration To train the model, we need a large number of ground-truth values. It is impossible to get enough values naturally, since one value corresponds to a camera-LiDAR pair; thus we have to adopt an approach similar to [6, 25] to generate artificial decalibrations on rotation and translation parameters, which we illustrate in Figure 5b. As we want our network to be able to calibrate a LiDAR and a camera whatever their location on the vehicle or robot, we need to choose decalibration ranges capable of simulating virtually any pairing on the vehicle. Considering a car equipped with a rotating lidar and a camera facing any direction around the car (e.g. front, back, side, etc.), we choose a decalibration of $\pm 180°$ on the yaw axis, $\pm 30°$ on the other rotation axes, and $\pm 150cm$ along each translation axis. Within this setup, the whole yaw rotation is covered. The decalibration range on the other rotation axes is more than any previous works, such as [6, 12, 13, 16, 17, 25, 29]. It is amply sufficient to cover most situations : with a $30°$ rotation on the pitch axis, there is

not any overlap left between Lidar and camera field of view, as shown in Figure 2a. Similarly, 150cm for each translation axis is enough to cover most cases, with, for example, the maximal distance between the LiDAR and a camera in KITTI being 60cm.

Data Augmentation Data augmentation is applied to further improve the generalisation capabilities of PseudoCal. Specifically, we apply a same rotation and translation on point clouds from both sensors, which is illustrated in Figure 5c. This does not affect the calibration parameters (the transformation T between these two sensors), but alters the input for the backbone. It helps ensure that the model does not overfit to a specific sensor configuration or a particular pattern of decalibration.

Training Details The loss weights, learning rate ($3e^{-5}$), and batch size (8) were chosen empirically by doing a sweep across a set of values. The different modules were trained independently from scratch and the weights frozen. Training was done on Nvidia V100 GPUs.

4 Experiments

In this section, we present a comprehensive set of experiments which primary goals are threefold: (i) to validate the performance of PseudoCal in terms of calibration accuracy, (ii) to evaluate the design choices and the contribution of each module, and (iii) to compare PseudoCal's performance with existing state-of-the-art methods.

4.1 Dataset

To evaluate the performance of PseudoCal and compare it to existing works, we employ the KITTI dataset [8]. It provides accurate ground-truth values for extrinsic calibration parameters, making it a reference benchmark for this task. The KITTI dataset comprises real-world data collected by a vehicle equipped with a Velodyne HDL-64E LiDAR sensor and front cameras. We use the same split for training and testing data as the one used in [25].

4.2 Results

4.2.1 Qualitative Results

Figure 6 illustrates the effectiveness of PseudoCal at different stages of the calibration process. The first row shows the initial decalibration, while the second row presents a marked improvement achieved by the PseudoPillars module. This sets the stage for the final row, where the refinement cascade fine-tunes the calibration to a level visually indistinguishable from the ground truth. This demonstrates the efficacy of our two-step approach—coarse adjustment followed by fine-tuning—in achieving accurate calibration, even in extreme cases.

4.2.2 Comparison to the State of the Art

Table 2 presents the quantitative results of the proposed PseudoCal method compared to state-of-the-art deep learning-based self-calibration techniques. We report the Mean Absolute Error (MAE) rotation and translation estimates, as well as the acceptable range of

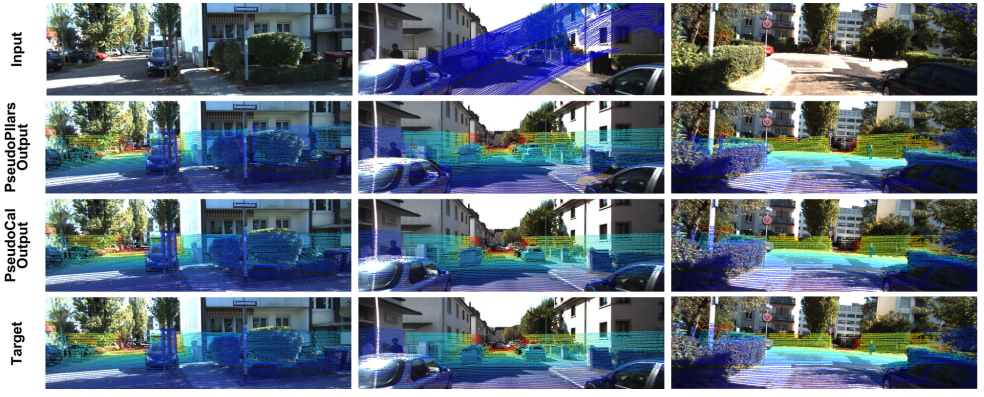


Figure 6: Qualitative evaluation of PseudoCal on KITTI (best viewed on screen). The point clouds are color-coded according to depth. Rows represent the initial decalibrations, PseudoPillars' coarse adjustments, PseudoCal's final refined calibrations, and groundtruth.

Model	Mean Absolute Error		Decalibration Range	
	Rotation (°)	Translation (cm)	Rotation (°)	Translation (cm)
RegNet [25]	0.28	6	20, 20, 20	150
CalibNet [12]	0.41	4.34	10, 10, 10	20
LCCNet [17]	**0.03**	**0.36**	20, 20, 20	150
DXQ-Net [13]	0.04	0.77	5, 5, 5	10
UniCal [6]	0.04	0.89	1, 1, 1	10
PseudoCal (ours)	0.05	1.18	**30, 30, 180**	**150**

Table 2: Comparison on Mean Absolute Error (MAE) with deep learning-based methods from the state of the art. Rotation decalibration values correspond to the roll, pitch, and yaw axes. Translation decalibration has the same range on all axes. Evaluations are made on different subsets of KITTI. PseudoCal is evaluated on the same set as RegNet, which [6] demonstrate to be the most challenging.

decalibration for each method. From these results, we can observe that PseudoCal achieves competitive performance compared to existing state-of-the-art methods [6, 12, 13, 17, 25], while being able to deal with the strongest decalibration ranges of all. The unmatched range of decalibration (up to 180 degrees on the yaw axis) used in our experiments highlights PseudoCal's ability to perform calibration for any camera location without initial information, within usual robotics and autonomous vehicles contexts. Thus, it succeeds in extreme cases where the other compared methods would inherently fail, as illustrated in Figure 2.

In summary, PseudoCal is to our knowledge the first deep-learning based extrinsic calibration method that does not focus only on parameters refinement. This makes PseudoCal a step forward for camera-LiDAR calibration in autonomous systems.

4.2.3 Ablation Study

Results of the conducted experiments are reported in Table 3. Experiment 1 evaluates the performance of the PseudoPillars module alone. It shows, as expected, a higher MAE compared to the complete PseudoCal method. Nevertheless, this is a reliable coarse estimate for

Experiment		Mean Average Error	
		Rotation (°)	Translation (cm)
1	PseudoPillars	3.09	19.9
2	PseudoPillars without Canny-based Noise Removal	4.18	27.93
3	PseudoPillars + UniCal-M	0.90	1.37
4	PseudoPillars + UniCal-M + UniCal-α	0.15	1.25
5	PseudoPillars + UniCal-M + UniCal-α+Unical-S	0.05	1.19
6	**PseudoPillars + UniCal-M + UniCal-S**	**0.05**	**1.18**

Table 3: Ablation experiments results. UniCal-α is trained on a decalibration range of $\pm 3°$ for rotation axes and $\pm 25cm$ for translation axes.

the refining cascade.

Experiment 2 shows the efficiency of our Canny-based noise removal (illustrated in Figure 4) in improving accuracy, as not using it leads to an increase of the average error of about 35% on rotation and 40% on translation.

In Experiment 3, we incorporate one UniCal-M module after the PseudoPillars module from Experiment 1. The addition of a single UniCal-M module dramatically improves the calibration performance, with the rotation MAE reduced to $0.90°$ and the translation MAE at $1.37cm$. This highlights the effectiveness of the cascading architecture in refining the calibration. This also confirms that the training range for UniCal-M is sufficiently large to accommodate the outputs of PseudoPillars. In Experiment 4, we tried adding a UniCal-α, trained on a decalibration range of $\pm 3°$ for rotation axes and $\pm 25cm$ for translation axes, which could be a good intermediate range between UniCal-M and UniCal-S. We then added a final UniCal-S in Experiment 5. Finally, we compared it to the actual PseudoCal architecture in Experiment 6, which requires only two refining modules.

Higher MAE in Experiment 4 compared to Experiment 6 suggest that a module trained on a smaller range, such as UniCal-S, is required to achieve state-of-the-art accuracy. Moreover, similar MAE in Experiments 5 and 6 demonstrate that UniCal-M and UniCal-S are sufficient to correct all samples, as an additional intermediate network does not improve the final accuracy. These results thus demonstrate Experiment 6 as the most suitable architecture.

5 Conclusion

We have introduced PseudoCal, a novel sensor calibration method that exploits the potential of pseudo-LiDAR through the PseudoPillars module, coupled with a cascaded architecture. This technique enables accurate calibration quasi-independently from any initial knowledge of extrinsic parameters, representing a significant breakthrough in camera-LiDAR calibration. PseudoCal has proven its efficacy on the KITTI dataset, demonstrating its robustness in autonomous system applications. Its architecture limits the number of cascaded modules compared to previous methods [17, 25], and rely on a light refining model [6], making it an appealing choice for embedded applications.

Future work will investigate extending our approach to other sensor modalities, incorporating additional pseudo-LiDAR representations, and further refining the network architecture and training strategies. We could also consider processing successive frames as sequences to improve the results. PseudoCal, with its novel approach to sensor calibration, not only advances the field, but also lays a robust foundation for future research.

Acknowledgements

This work was granted access to the HPC resources on the supercomputer Jean Zay of IDRIS under the allocation 2023-AD011014065 made by GENCI.

This work has been carried out within SIVALab, joint laboratory between Renault and Heudiasyc (CNRS / Université de technologie de Compiègne).

References

[1] Jorge Beltrán, Carlos Guindel, Arturo de la Escalera, and Fernando García. Automatic Extrinsic Calibration Method for LiDAR and Camera Sensor Setups. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):17677–17689, October 2022.

[2] Stanley Bileschi. Fully automatic calibration of LIDAR and video streams from a vehicle. *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1457–1464, September 2009.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2021.

[4] John Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, November 1986.

[5] Wentao Cheng, Weisi Lin, Kan Chen, and Xinfeng Zhang. Cascaded parallel filtering for memory-efficient image-based localization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1032–1041, 2019.

[6] Mathieu Cocheteux, Aaron Low, and Marius Bruehlmeier. Unical: a single-branch transformer-based model for camera-to-lidar calibration and validation. *arXiv preprint arXiv:2304.09715*, 2023.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[8] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, September 2013.

[9] Andreas Geiger, Frank Moosmann, Omer Car, and Bernhard Schuster. Automatic camera and range sensor calibration using a single shot. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3936–3943, May 2012.

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611, 2017.

[11] Ryoichi Ishikawa, Takeshi Oishi, and Katsushi Ikeuchi. LiDAR and Camera Calibration Using Motions Estimated by Sensor Fusion Odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7342–7349, October 2018.

[12] Ganesh Iyer, R. Karnik Ram., J. Krishna Murthy, and K. Madhava Krishna. CalibNet: Geometrically Supervised Extrinsic Calibration using 3D Spatial Transformer Networks. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1110–1117, October 2018.

[13] Xin Jing, Xiaqing Ding, Rong Xiong, Huanjun Deng, and Yue Wang. DXQ-Net: Differentiable LiDAR-Camera Extrinsic Calibration Using Quality-aware Flow. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6235–6241, October 2022.

[14] Doyeon Kim, Woonghyun Ka, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022.

[15] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast Encoders for Object Detection From Point Clouds. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12689–12697. IEEE, June 2019.

[16] Zhijian Liu, Haotian Tang, Sibo Zhu, and Song Han. SemAlign: Annotation-Free Camera-LiDAR Calibration with Semantic Alignment Loss. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8845–8851, September 2021.

[17] Xudong Lv, Boya Wang, Ziwen Dou, Dong Ye, and Shuo Wang. Lccnet: Lidar and camera self-calibration using cost volume network. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2888–2895, 2021.

[18] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations (ICLR)*, 2022.

[19] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9685–9694, 2021.

[20] Gaurav Pandey, James McBride, Silvio Savarese, and Ryan Eustice. Automatic targetless extrinsic calibration of a 3d lidar and camera by maximizing mutual information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 2053–2059, 2012.

[21] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017.

[22] Rui Qian, Divyansh Garg, Yan Wang, Yurong You, Serge Belongie, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. End-to-end pseudo-lidar for image-based 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5881–5890, 2020.

[23] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021.

[24] Sergio A. Rodriguez F., Vincent Fremont, and Philippe Bonnifait. Extrinsic calibration between a multi-layer lidar and a camera. In *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 214–219. IEEE, August 2008.

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

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30, 2017.

[27] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8445–8453, 2019.

[28] Xinshuo Weng and Kris Kitani. Monocular 3d object detection with pseudo-lidar point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[29] Shan Wu, Amnir Hadachi, Damien Vivet, and Yadu Prabhakar. This is The Way: Sensors Auto-calibration Approach Based on Deep Learning for Self-driving Cars. *IEEE Sensors Journal*, pages 1–1, 2021.

[30] 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*, 6(4):7517–7524, 2021.

[31] Qilong Zhang and R. Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2301–2306 vol.3, 2004.

[32] Lipu Zhou, Zimo Li, and Michael Kaess. Automatic Extrinsic Calibration of a Camera and a 3D LiDAR Using Line and Plane Correspondences. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5562–5569, 2018.