RC-AutoCalib: An End-to-End Radar-Camera Automatic Calibration Network

Supplementary Material

A. Detail of Feature Extraction

First, we transform point clouds and images into two unified representations: frontal view depth maps (I_R^{FV}, $I_I^{FV} \in \mathbb{R}^{H\times W}$) and BEV images ($I_I^{BEV}$, $I_R^{BEV} \in \mathbb{R}^{H'\times W'}$). These representations allow us to effectively compare and fuse sensor data from different perspectives.

To extract features from radar data (the mis-calibrated images I_R^{FV}, I_R^{BEV}), we employ the first three blocks of ResNet [7] as the network structure. This setup, which includes convolutional and pooling layers, is well-suited for extracting low-level image features such as edges and textures. Additionally, considering the sparse nature of radar data and its distinct characteristics compared to image data, we train the radar-specific ResNet from scratch to effectively capture the relevant features.

For the depth map I_I^{FV} and pseudo-BEV map I_I^{BEV}, which are derived from the input image using a depth estimation network, feature extraction is performed using just two convolutional layers. Given that these image-derived maps are rich in semantic information, this simplified network configuration has proven sufficient for extracting detailed features while avoiding unnecessary complexity

To enhance semantic content in the frontal view, context features are extracted from the original input image using ResNet18's first three blocks with pretrained weights from ImageNet [4]. These blocks excel in capturing rich contextual information, which is integrated with features extracted from depth map I_I^{FV} to produce a comprehensive feature representation enriched with semantic information. This fusion not only enhances semantic details in the frontal view but also improves contrast and consistency across multi-view features.

Finally, we obtain feature sets from different perspectives for radar and camera, represented as F_R^{FV}, $F_I^{FV} \in \mathbb{R}^{H/8\times W/8\times C}$ and F_R^{BEV}, $F_I^{BEV} \in \mathbb{R}^{H'/8\times W'/8\times C}$, where H and W are the dimensions of the frontal view image, and H' and W' are the dimensions of the BEV image.

B. Detail of Multi-Modal Cross-Attention Mechanism

The output $O_{I\leftarrow R}$ of the Multi-Modal Cross-Attention Mechanism, as shown in Eq. (1), is computed by concatenating the image feature f_I, reshaped from F_I to dimensions $(m \times c)$, with the attended feature $m_{I\leftarrow R}$. This concatenated feature is then processed through a feed-forward network (FFN) that employs LayerNorm [1], GELU [8] activation functions, and linear layers, resulting in the output reshaped to $O_{I\leftarrow R} \in \mathbb{R}^{h\times w\times c}$. Similarly, $O_{R\leftarrow I}$ is computed using the same process, as shown in Eq. (2).

$$\begin{aligned} O_{I\leftarrow R} &= \Theta(F_I, m_{I\leftarrow R}) \\ &= \text{reshape}(\text{FFN}(concat[f_I, m_{I\leftarrow R}]), (h, w, c)), \end{aligned} \tag{1}$$

$$\begin{aligned} O_{R\leftarrow I} &= \Theta(F_R, m_{R\leftarrow I}) \\ &= \text{reshape}(\text{FFN}(concat[f_R, m_{R\leftarrow I}]), (h, w, c)), \end{aligned} \tag{2}$$

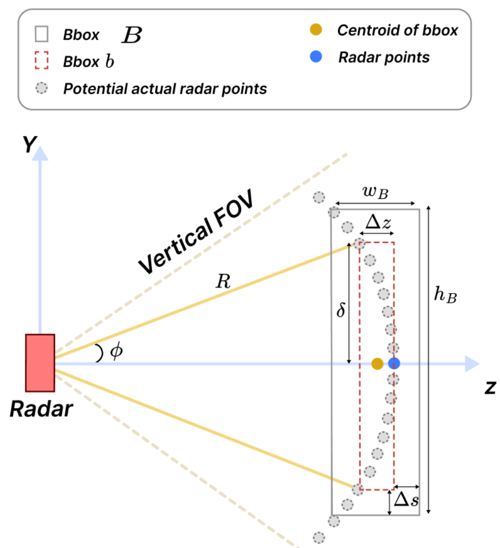

Figure 1. Illustration of bounding box B. Suppose we consider only the y and z axes to calculate w_B based on δ

The cross-attention maps $I_{I\leftarrow R}$, $I_{R\leftarrow I}$ between radar and image features will be computed according to the following equation:

$$I_{I\leftarrow R} = \text{reshape}(\overset{m}{\underset{j}{\max}}(\text{Softmax}(a_{IR})_{ij}), (h, w, 1)) \tag{3}$$

$$I_{R\leftarrow I} = \text{reshape}(\overset{m}{\underset{j}{\max}}(\text{Softmax}(a_{IR}^{\top})_{ij}), (h, w, 1)) \tag{4}$$

C. Detail of Noise-Resistant Matcher

Fig. 1 illustrates the principle of the noise-resistant matcher, which simplifies by removing the x-axis. The radar point cloud is computed based on azimuth angle θ and distance R, all lying on the radar point plane with a constant y-axis value of 0. However, in reality, radar points are reflected from objects at a distance from the radar plane, leading to the appearance of uncertain elevation angle ϕ within the vertical FOV boundary. Therefore, we depict the gray points in figure as potential actual radar points within the FOV, at the same distance R but with varying elevation angles ϕ.

For potential actual radar point, there is an error in both the x and z axes, corresponding to Δx and Δz as defined in [34]. Using these errors, we define a region encompassing neighboring LiDAR points. Essentially, each radar point creates a bounding box b to identify LiDAR neighbors associated with potential actual radar points. This 3D bounding box is fixed with a parameter δ, which is the allowable height error threshold, and the width and depth correspond to Δx and Δz, respectively. In Fig. 1, when the allowable height limit for the potential actual radar point is δ, the maximum allowable z value for the gray point is when it coincides with the radar point (blue) $P_r^c(X_r^c, Y_r^c, Z_r^c)$, and the

Figure 2. Calibration results by projecting radar points onto the FV image

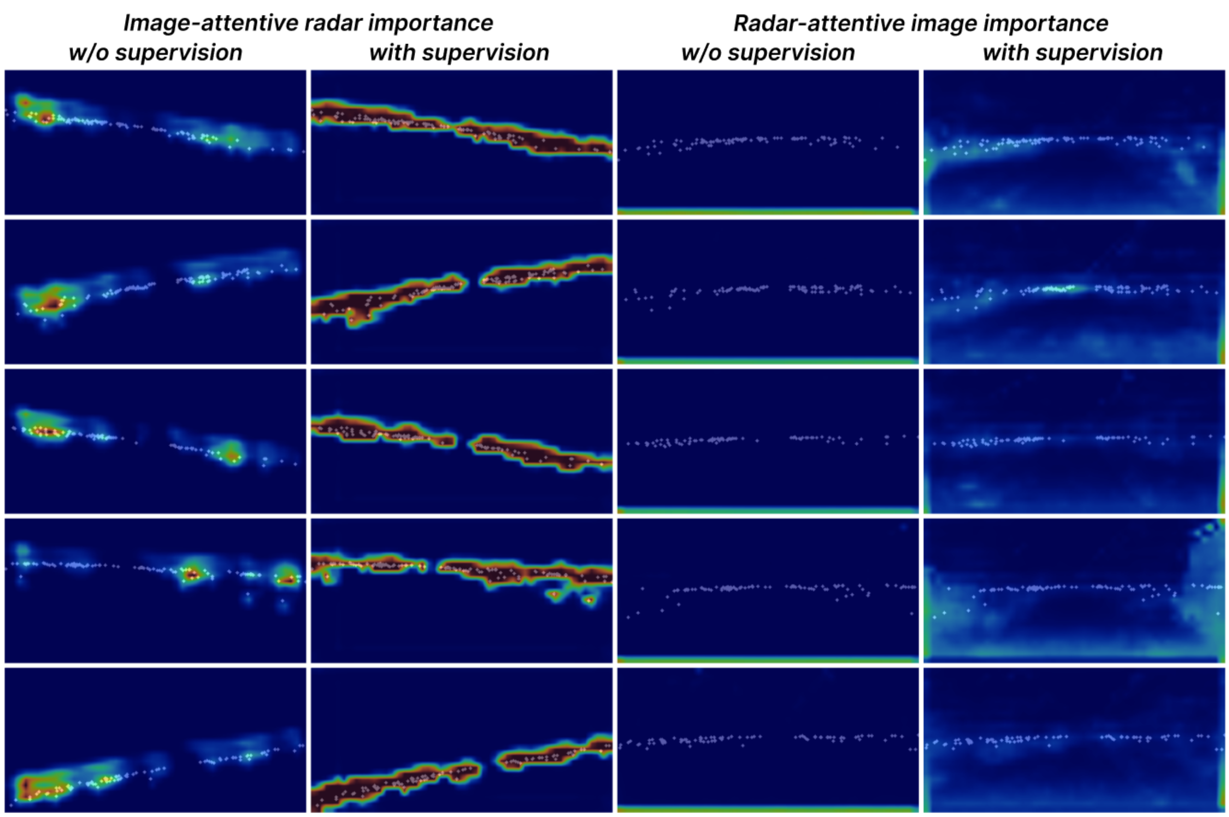

Figure 3. Examples of FV cross-attention maps highlight the important regions the model focuses on.

Scenario	Methods	Rotation (°)			
		Mean	Roll	Pitch	Yaw
Urban	LCCNet-1	2.969	3.123	2.703	3.081
	NetCalib2	2.643	**0.742**	3.221	3.966
	CalibDepth	3.656	2.088	3.913	4.966
	Coarse [31]	4.395	3.148	4.645	5.392
	Fine [31]	4.956	3.152	5.195	6.520
	Ours	**1.875**	0.934	**2.609**	**2.082**
Rain	LCCNet-1	2.394	3.400	**1.929**	**1.853**
	NetCalib2	2.612	0.644	2.781	4.411
	CalibDepth	3.412	1.686	4.299	4.251
	Coarse [31]	4.092	2.060	4.663	5.554
	Fine [31]	4.776	2.050	5.269	7.007
	Ours	**1.922**	**0.622**	3.273	1.870

Table 1. Cross-dataset evaluation on the aiMotive dataset. The nuScenes-trained models are evaluated on aiMotive scenarios (urban and rain) with an initial rotation error range within 10°.

minimum allowable z value is at $(Z_r^c - \Delta z)$, similarly for the x-axis. Therefore, the center of the 3D bounding box b is defined as $(X_r^c - \Delta x/2, y_r^c, Z_r^c - \Delta z/2)$.

Additionally, in reality, LiDAR points will not fit exactly with potential actual radar points due to measurement inaccuracies of both the radar and LiDAR sensors. Therefore, we add an offset to the width, height, and depth by a fixed error Δs, forming the 3D bounding box B. Both parameters δ and Δs are tuned based on the unit meter.

D. Implementation Details

We resized the original 1600x900 images to 400x192 pixels. Training was conducted on an NVIDIA GTX 3090 GPU for 50 epochs using the Adam optimizer with an initial learning rate of 1e-4, halving it every 8 epochs. The loss function weights were set to $\lambda = 0.75$ and $\beta = 0.1$. In the Regression Head, the LSTM module had a fixed iterative step size of 3. In the noise-resistant matcher section, we selected a threshold τ of 3, Δs of 0.5, and δ of 1.

E. Additional Experimental Results

E.1. Cross-dataset evaluation

We compared our RC-AutoCalib with other related methods on the **aiMotive** [26] dataset, as shown in Tab. 1. In this experiment, all models were trained on the nuScenes dataset and directly tested on two scenarios from **aiMotive**. Our method outperformed others in both scenarios, demonstrating the superior gener alization ability of our model.

E.2. Positive-Negative Balance in Feature Matching Supervision Loss

In Tab. 2, we experimented with different values of λ for $L_{matching}$, including 0.9, 0.75, and 0.5. When λ was set to 0.75, both rotation error and translation error reached their lowest values.

λ	Rotation(°)				Translation(cm)			
	Mean	Roll	Pitch	Yaw	Mean	X	Y	Z
0.9	0.460	0.142	0.222	1.017	10.896	12.561	7.503	**12.625**
0.75	**0.427**	**0.130**	**0.199**	**0.953**	**9.498**	12.564	**3.295**	12.635
0.5	0.442	0.153	0.209	0.9634	11.295	**12.547**	8.699	12.638

Table 2. Ablation Study on Positive-Negative Balance in Feature Matching Supervision Loss

E.3. LiDAR-Camera Calibration

To showcase the adaptability of our approach, we extended it to LiDAR-camera calibration. We trained our method on the nuScenes dataset using the same train-test split as reported in the main paper and on the KITTI dataset with 24,000 training samples and 6,000 test samples. These experiments were conducted without the Noise-Resistant Matcher, which is specific to radar data.

As shown in Tab. 3, we compare our method with previous approaches, including CalibNet[11], CalibRCNN[33], CalibDNN[48], CalNet[32], and CalibDepth[49]. The results demonstrate that our method outperforms them, confirming its scalability and robustness.

Range	Methods	Rotation(°)				Translation (cm)			
		Mean	Roll	Pitch	Yaw	Mean	X	Y	Z
KITTI	CalibNet	0.410	0.150	0.900	0.181	7.82	12.10	3.49	7.87
	CalibRCNN	0.428	0.199	0.640	0.446	5.30	6.20	4.30	5.40
	CalibDNN	0.210	0.110	0.350	0.180	5.07	3.80	1.80	9.60
	CalNet	0.200	0.100	0.380	**0.120**	3.03	3.65	1.63	3.80
	Ours	**0.142**	**0.066**	**0.096**	0.268	**1.941**	**2.479**	**0.998**	**2.347**
nuScenes	CalibDepth	0.408	0.215	0.226	0.794	8.33	11.19	4.27	9.53
	Ours	**0.208**	**0.142**	**0.148**	**0.337**	**3.183**	**1.010**	**0.7836**	**7.836**

Table 3. Comparison of the method extension to the LiDAR-Camera auto-calibration task on the nuScenes and KITTI datasets. The methods are compared with mis-calibration ranges R1 ($\pm 10°$, $\pm 0.25m$). Notably, the CalibDepth method was retrained on the nuScenes dataset by us.

E.4. Effects on Downstream tasks

To validate the impact of our method on 3D object detection, we initialized random incorrect extrinsic parameters, corrected the parameters for each image in the scenes test set, and evaluated the pre-trained CRN [16] 3D object detection model. The mAP performance decreased by only **0.27%** compared to using ground-truth calibration, indicating a negligible difference.

E.5. Qualitative Results

Fig. 2 shows additional calibration results, including the results for each iteration. It can be observed that even with a large initial error, our method effectively reduces the error progressively with each iteration. In Fig. 3, we present additional attention maps using heatmaps in the FV, with the projected radar points marked in white to indicate critical regions.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 4

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 7

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[5] Joris Domhof, Julian FP Kooij, and Dariu M Gavrila. An extrinsic calibration tool for radar, camera and lidar. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8107–8113. IEEE, 2019. 1

[6] Ghina El Natour, Omar Ait Aider, Raphael Rouveure, François Berry, and Patrice Faure. Radar and vision sensors calibration for outdoor 3d reconstruction. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2084–2089. IEEE, 2015. 1, 3

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 1

[8] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 1

[9] Markus Hiller, Krista A. Ehinger, and Tom Drummond. Perceiving longer sequences with bi-directional cross-attention transformers. *ArXiv*, abs/2402.12138, 2024. 5

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 6

[11] Ganesh Iyer, R Karnik Ram, J Krishna Murthy, and K Madhava Krishna. Calibnet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1110–1117. IEEE, 2018. 1, 3, 4

[12] 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. IEEE, 2022. 1

[13] Du Yong Kim and Moongu Jeon. Data fusion of radar and image measurements for multi-object tracking via kalman filtering. *Information Sciences*, 278:641–652, 2014. 1, 3

[14] Jihun Kim, Dong Seog Han, and Benaoumeur Senouci. Radar and vision sensor fusion for object detection in autonomous vehicle surroundings. In *2018 tenth international conference on ubiquitous and future networks (ICUFN)*, pages 76–78. IEEE, 2018. 3

[15] Taehwan Kim, Sungho Kim, Eunryung Lee, and Miryong Park. Comparative analysis of radar-ir sensor fusion methods for object detection. In *2017 17th International Conference on Control, Automation and Systems (ICCAS)*, pages 1576–1580. IEEE, 2017. 1, 3

[16] Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. Crn: Camera radar net for accurate, robust, efficient 3d perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17615–17626, 2023. 3

[17] Jesse Levinson and Sebastian Thrun. Automatic online calibration of cameras and lasers. In *Robotics: science and systems*. Citeseer, 2013. 3

[18] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519, 2019. 6

[19] 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. 3

[20] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection, 2022. 1

[21] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 5

[22] Zhijian Liu, Haotian Tang, Sibo Zhu, and Song Han. Semalign: Annotation-free camera-lidar calibration with semantic alignment loss. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8845–8851, 2021. 1, 2

[23] Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, and Praveen Narayanan. Radar-camera pixel depth association for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12507–12516, 2021. 1, 2

[24] Xudong Lv, Boya Wang, Ziwen Dou, Dong Ye, and Shuo Wang. Lccnet: Lidar and camera self-calibration using cost volume network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2894–2901, 2021. 1, 2, 3, 4, 7

[25] Xudong Lv, Shuo Wang, and Dong Ye. Cfnet: Lidar-camera registration using calibration flow network. *Sensors*, 21(23): 8112, 2021. 1

[26] Tamás Matuszka, Iván Barton, Ádám Butykai, Péter Hajas, Dávid Kiss, Domonkos Kovács, Sándor Kunsági-Máté, Péter Lengyel, Gábor Németh, Levente Pető, et al. aimotive dataset: A multimodal dataset for robust autonomous driving with long-range perception. *arXiv preprint arXiv:2211.09445*, 2022. 3

[27] Gaurav Pandey, James McBride, Silvio Savarese, and Ryan Eustice. Automatic targetless extrinsic calibration of a 3d li-

dar and camera by maximizing mutual information. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2053–2059, 2012. 3

[28] Juraj Peršić, Luka Petrović, Ivan Marković, and Ivan Petrović. Online multi-sensor calibration based on moving object tracking. *Advanced Robotics*, 35(3-4):130–140, 2021. 3

[29] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014. 3, 6

[30] 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. 1, 3

[31] Christoph Schöller, Maximilian Schnettler, Annkathrin Krämmer, Gereon Hinz, Maida Bakovic, Müge Güzet, and Alois Knoll. Targetless rotational auto-calibration of radar and camera for intelligent transportation systems. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3934–3941. IEEE, 2019. 1, 3, 7

[32] Hongcheng Shang and Bin-Jie Hu. Calnet: Lidar-camera online calibration with channel attention and liquid time-constant network. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 5147–5154, 2022. 3

[33] Jieying Shi, Ziheng Zhu, Jianhua Zhang, Ruyu Liu, Zhenhua Wang, Shengyong Chen, and Honghai Liu. Calibrcnn: Calibrating camera and lidar by recurrent convolutional neural network and geometric constraints. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10197–10202. IEEE, 2020. 1, 2, 3

[34] Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, and Alex Wong. Depth estimation from camera image and mmwave radar point cloud. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9275–9285, 2023. 6, 1

[35] Shigeki Sugimoto, Hayato Tateda, Hidekazu Takahashi, and Masatoshi Okutomi. Obstacle detection using millimeter-wave radar and its visualization on image sequence. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, pages 342–345. IEEE, 2004. 1, 3

[36] Zachary Taylor and Juan Nieto. Motion-based calibration of multimodal sensor arrays. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4843–4850. IEEE, 2015. 3

[37] Guangming Wang, Jiahao Qiu, Yanfeng Guo, and Hesheng Wang. Fusionnet: Coarse-to-fine extrinsic calibration network of lidar and camera with hierarchical point-pixel fusion. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8964–8970. IEEE, 2022. 1, 2, 4

[38] Tao Wang, Nanning Zheng, Jingmin Xin, and Zheng Ma. Integrating millimeter wave radar with a monocular vision sensor for on-road obstacle detection applications. *Sensors*, 11(9):8992–9008, 2011. 1, 3

[39] Emmett Wise, Juraj Peršić, Christopher Grebe, Ivan Petrović, and Jonathan Kelly. A continuous-time approach for 3d radar-to-camera extrinsic calibration. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13164–13170. IEEE, 2021. 3

[40] Emmett Wise, Qilong Cheng, and Jonathan Kelly. Spatiotemporal calibration of 3-d millimetre-wavelength radar-camera pairs. *IEEE Transactions on Robotics*, 2023. 3

[41] Shan Wu, Amnir Hadachi, Damien Vivet, and Yadu Prabhakar. Netcalib: A novel approach for lidar-camera auto-calibration based on deep learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6648–6655. IEEE, 2021. 1, 2

[42] 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*, 21(24):27779–27788, 2021. 1, 7

[43] Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, and Deng Cai. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5418–5427, 2022. 1

[44] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 5

[45] Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. Radarnet: Exploiting radar for robust perception of dynamic objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 496–512. Springer, 2020. 1

[46] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 4

[47] 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. 1, 2, 3

[48] Ganning Zhao, Jiesi Hu, Suya You, and C-C Jay Kuo. Calibdnn: multimodal sensor calibration for perception using deep neural networks. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXX*, pages 324–335. SPIE, 2021. 2, 4, 3

[49] Jiangtong Zhu, Jianru Xue, and Pu Zhang. Calibdepth: Unifying depth map representation for iterative lidar-camera online calibration. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 726–733. IEEE, 2023. 2, 6, 7, 3

[50] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16280–16290, 2021. 1