2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

# BEVCar: Camera-Radar Fusion for BEV Map and Object Segmentation

Jonas Schramm[1*], Niclas Vödisch[1*], Kürsat Petek[1*], B Ravi Kiran[2], Senthil Yogamani[3], Wolfram Burgard[4], and Abhinav Valada[1]

***Abstract*— Semantic scene segmentation from a bird's-eye-view (BEV) perspective plays a crucial role in facilitating planning and decision-making for mobile robots. Although recent vision-only methods have demonstrated notable advancements in performance, they often struggle under adverse illumination conditions such as rain or nighttime. While active sensors offer a solution to this challenge, the prohibitively high cost of LiDARs remains a limiting factor. Fusing camera data with automotive radars poses a more inexpensive alternative but has received less attention in prior research. In this work, we aim to advance this promising avenue by introducing BEVCar, a novel approach for joint BEV object and map segmentation. The core novelty of our approach lies in first learning a point-based encoding of raw radar data, which is then leveraged to efficiently initialize the lifting of image features into the BEV space. We perform extensive experiments on the nuScenes dataset and demonstrate that BEVCar outperforms the current state of the art. Moreover, we show that incorporating radar information significantly enhances robustness in challenging environmental conditions and improves segmentation performance for distant objects. To foster future research, we provide the weather split of the nuScenes dataset used in our experiments, along with our code and trained models at `http://bevcar.cs.uni-freiburg.de`.**

## I. Introduction

Mobile robots such as autonomous vehicles heavily rely on accurate and robust perception of their environment. Therefore, robotic platforms are typically equipped with a variety of sensors [1]–[3], each providing complementary information. For instance, surround-view cameras offer dense RGB images, while LiDAR or radar systems provide sparse depth measurements. However, fusing data from these different modalities poses a significant challenge due to inherently different data structures. A common approach to address this challenge is to employ a bird's-eye-view (BEV) representation as a shared reference frame [4]–[9].

While both LiDAR and radar data can be directly transformed into BEV space, camera-based information requires conversion from the image plane to a top-down view. Consequently, various lifting strategies have been proposed [4], [10], [11] resulting in tremendous performance improvements of vision-only approaches, some of which have been extended to incorporate LiDAR data [5], [7]. Despite the ability of LiDARs to yield highly accurate 3D point clouds, their suitability for large-scale deployment remains controversial due to their substantially higher costs compared to automotive radars. Nonetheless, camera-radar fusion has received considerably less attention from the research community, often only explored in addition to LiDAR input [8], [12]. In contrast, radar has been criticized as being too sparse to be effectively utilized in isolation [12].

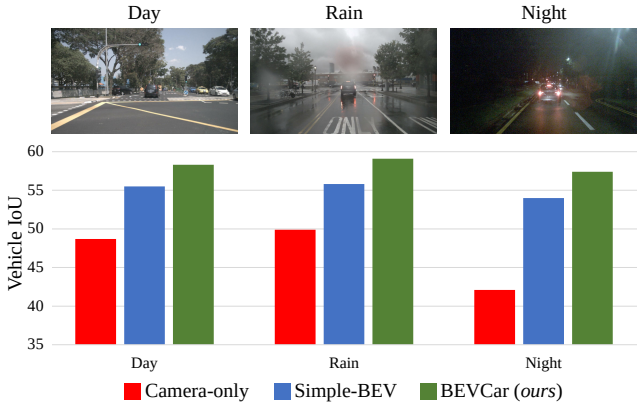


Fig. 1. We propose a novel method for BEV Camera-radar fusion (BEVCar) for map and object segmentation. We demonstrate that BEVCar yields more accurate predictions under adverse weather conditions than camera-only baselines while outperforming prior camera-radar works [6].

In this work, we underscore the pivotal role of radar in advancing robust robotic perception. Specifically, we focus on BEV object and map segmentation, highlighting the distinct advantage of radar in vision-impaired environmental conditions. While previous research has explored camera-radar fusion for BEV segmentation, some approaches necessitate additional LiDAR supervision during training [9] or rely on specific radar metadata [6], [8], which may not be accessible across models from different manufacturers. To address these limitations, we propose a novel method that operates independently of such constraints. Our proposed BEVCar architecture comprises two sensor-specific encoders and two attention-based modules for image lifting and BEV camera-radar fusion, respectively. Subsequently, we feed the fused features through a multi-task head to generate both map and object segmentation maps. We extensively evaluate our approach on the nuScenes [2] dataset and demonstrate that it achieves state-of-the-art performance for camera-radar fusion while being robust in challenging illumination conditions.

The main contributions are as follows:

1) We introduce the novel BEVCar for BEV map and object segmentation from camera and radar data.
2) We propose a new attention-based image lifting scheme that exploits sparse radar points for query initialization.

* Equal contribution.
[1] Department of Computer Science, University of Freiburg, Germany.
[2] Qualcomm SARL France.
[3] QT Technologies Ireland Limited.
[4] Department of Eng., University of Technology Nuremberg, Germany.
This work was funded by Qualcomm Technologies Inc., the German Research Foundation (DFG) Emmy Noether Program grant No 468878300, and an academic grant from NVIDIA.

3) We show that learning-based radar encoding outperforms the usage of raw metadata.
4) We extensively compare BEVCar with previous baselines under challenging environmental conditions and demonstrate the advantage of utilizing radar measurements.
5) We make the used day/night/rain splits on nuScenes [2] publicly available and release our code and trained models at http://bevcar.cs.uni-freiburg.de.

## II. Related Work

In this section, we present an overview of vision-only methods operating in the bird's-eye-view (BEV) and review previous approaches for radar-based perception.

*Camera-Based BEV Perception:* Current research in the field of camera-based BEV perception aims to handle the view discrepancy between the image space and the BEV space. Existing approaches typically employ an encoder-decoder architecture, incorporating a distinctive view transformation module to address spatial variations between the image and BEV planes. Early works leverage variational autoencoders to decode features directly into a 2D top-view Cartesian coordinate system [13]. In contrast, VPN [14] utilizes a multilayer perceptron (MLP) to model dependencies across spatial locations in the image and BEV feature maps, ensuring global coverage in the view transformation. Roddick *et al.* [15] improve upon these works by introducing a more explicit geometry modeling. In particular, they propose a pyramid occupancy network with a per-scale dense transformer module to learn the mapping between a column in the image view and a ray in the BEV map. PoBEV [16] extends this concept by processing flat and vertical features separately with distinct transformer modules resulting in further performance improvement.

Recent methods can be categorized into lifting-based and attention-based mechanisms. Lifting-based approaches incorporate either an implicit depth distribution module [10] to project features to a latent space or an explicit depth estimation module to generate an intermediate 3D output, e.g., for the tasks of object detection [17] or scene completion [18]. Attention-based approaches formulate view transformation as a sequence-to-sequence translation from the image space to BEV. TIIM [19] applies inter-plane attention between a polar ray in the BEV space and a vertical column in the image combined with self-attention across each respective polar ray with significant performance improvement with respect to depth-based approaches such as LSS [10].

Recent advancements include full-surround view BEV perception approaches, such as CVT [20] that uses a cross-view transformer with learned positional embeddings to avoid explicit geometric modeling and exploiting this BEV representation for policy learning [21]. In contrast, BEVFormer [4] and BEVSegFormer [22] model geometry explicitly using camera calibration parameters and propose a deformable attention-based [23] spatial cross-attention module for view unprojection. BEVFormer [4] additionally employs a temporal attention module for aggregating BEV maps over time using vehicle ego-motion, which represents the state of the art in 3D object detection. Temporal aggregation is also employed in BEVerse [11], which extends existing approaches with a motion prediction head and demonstrates that the proposed multi-task network outperforms single-task networks indicating a positive transfer among the tasks. The aforementioned approaches are often combined with novel data augmentation techniques [24], which address the view discrepancy between the image and BEV by maintaining spatial consistency across each intermediate embedding. Finally, SkyEye [25] proposes a less-constrained method that learns semantic BEV maps from labeled frontal view images by reconstructing semantic images over time. Our work leverages recent progress in monocular BEV perception and takes advantage of the radar modality for a more geometrically feasible view projection. This is achieved through a novel attention-based image lifting scheme using radar queries. Additionally, we propose to exploit existing image backbones that are pre-trained with contrastive learning to further regularize the modality-specific branches.

*Radar-Based Perception:* Radars measure the distance to a target based on the time difference between emitting a radio wave and receiving its reflection. Published datasets for robotic applications include different types of radars such as spinning radars [1], automotive radars [2], or 4D imaging radars [3]. In this work, we focus on automotive radars. As radar poses a comparably inexpensive technology to measure distance directly, it has been leveraged to improve vision-based 3D object detection. While ClusterFusion [28] merges radar and camera data only in the image space, SparseFusion3D [29] performs sensor fusion both in the image and BEV space.

In segmentation, initial works investigated semantic segmentation of radar point clouds [30] without complementary vision input. More recently, research towards multi-modal BEV map and object segmentation has received growing attention. The authors of the pioneering work FISHING Net [12] propose an MLP-based lifting strategy for camera features. To combine these features with radar data, which are encoded by a UNet-like network, FISHING Net performs class-based priority pooling. In contrast, Simple-BEV [6] processes the raw radar data in a rasterized BEV format and concatenates these with image features that are lifted via bilinear sampling. Although Simple-BEV targets object-agnostic vehicle segmentation, the training relies on additional instance information for object center and pixel offset prediction. Since purely concatenation-based fusion might suffer from spatial misalignment, CRN [9] employs deformable attention [23] to aggregate image and radar features. However, the method uses LSS [10] for lifting the image features and requires LiDAR during training to supervise the depth distribution network. Finally, BEVGuide [8] does not exploit further knowledge other than available during deployment. Using homography-based projection, features from the EfficientNet [31] image backbone are transformed into a scale-ambiguous top-down representation. The radar data is converted to BEV space and then encoded by two convolutional layers. In contrast to prior works, BEVGuide proposes a bottom-up lifting approach by

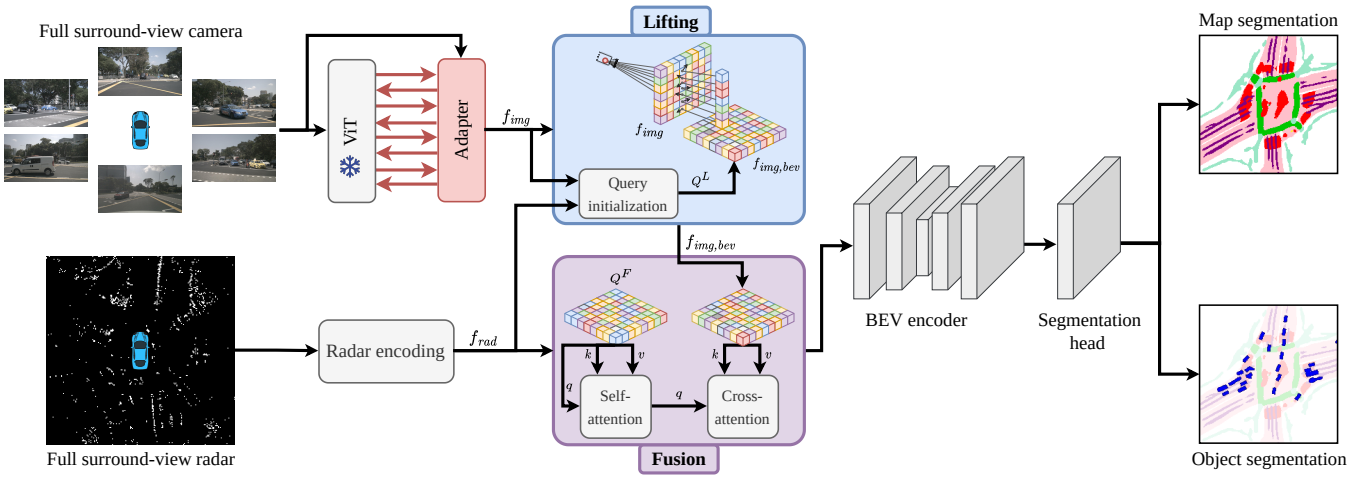


Fig. 2. Overview of our proposed BEVCar approach for camera-radar fusion for BEV map and object segmentation. We utilize a frozen DINOv2 [26] with a learnable adapter to encode the surround-view images. Inspired by LiDAR-based perception [27], we employ a learnable radar encoding instead of processing the raw metadata. We then lift the image features to the BEV space via deformable attention including the novel radar-driven query initialization scheme. Finally, we fuse the lifted image representation with the learned radar features in an attention-based manner and perform multi-class BEV segmentation for both vehicles and the map categories.

querying the sensor features from a unified BEV space to obtain sensor-specific embeddings that are then concatenated. In this work, we further advance these ideas and utilize a more refined radar encoder that is inspired by LiDAR processing [27]. Moreover, we propose a novel lifting scheme that explicitly leverages radar points as a strong prior.

## III. Technical Approach

In this section, we present our proposed BEVCar approach for BEV object and map segmentation from surround-view cameras and automotive radar. As illustrated in Fig. 2, BEVCar comprises two sensor-specific encoders for image and radar data, respectively. We lift the image features to the BEV space via deformable attention, where we utilize radar data to initialize the queries. Following an intermediate fusion strategy, we then combine the lifted image representation with the learned radar features using a cross-attention module. Finally, we reduce the spatial resolution in a bottleneck operation and perform BEV segmentation for both vehicles and the map with a single multi-class head. We provide further details of each step in the following subsections.

### A. Sensor Data Encoding

As depicted in Fig. 2, we process the raw data of both modalities in two separate encoders.

*Camera:* For encoding the camera data, we employ a frozen DINOv2 ViT-B/14 [26], whose image representation captures more semantic information than ResNet-based backbones [32]. Following the common approach [33], [34], we utilize a ViT adapter [35] with learnable weights. To cover the surround-view vision, we concatenate the images from $N$ cameras at each timestamp resulting in an input dimension of $N \times H \times W$, where $H$ and $W$ denote the image height and width, respectively. For downstream processing, the ViT adapter outputs multi-scale feature maps with $F$ channels that correspond to scales ¼, ⅛, 1/16, and 1/32 of the image size.

*Radar:* The radar data is represented by a point cloud with various features available for each point. Unlike prior works [6], [8], we emphasize that relying on the built-in post-processing from a specific radar model makes a method less versatile. Hence, similar to SparseFusion3D [29], we utilize only $D$ basic characteristics of a radar point: the 3D position $(x, y, z)$, uncompensated velocities $(v_x, v_y)$, and the radar cross-section *RCS*, which captures the detectability of a surface. Instead of utilizing the raw data [6], we propose to learn a radar representation inspired by encoding LiDAR point clouds [27]. First, we group the radar points based on their spatial position in a voxel grid of size $X \times Y \times Z$ that corresponds to the resolution of the BEV space and a discretization in height. To restrict memory requirements and alleviate bias towards high-density voxels, we employ random sampling in those voxels that contain more than $P$ radar points. Each point including its metadata is then fed through the point feature encoding as illustrated in Fig. 3, where FCN refers to fully connected layers. Note that the point feature encoding does not accumulate information from multiple voxels. Subsequently, we employ max pooling for each voxel to obtain a single feature vector of size $F$. Finally, we feed the voxel features through a CNN-based voxel space encoder to compress the features along the height dimension, resulting in the overall radar BEV encoding $f_{rad}$.

### B. Image Feature Lifting

We follow a learning-based approach to lift the encoded vision features from the 2D image plane to the BEV space. Inspired by BEVFormer [4], we utilize deformable attention [23] but propose a novel query initialization scheme that exploits sparse radar points.

*Query Initialization:* The core motivation of our proposed query initialization scheme is to leverage 3D information from radar measurements for an initial lifting step of the 2D image features to the BEV space. As visualized in Fig. 4, we first create a voxel space of size $X \times Y \times Z$ that is

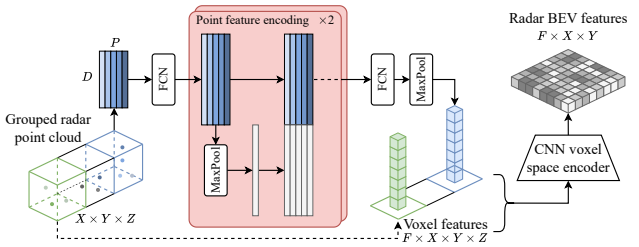


Fig. 3. Inspired by LiDAR processing, we encode the radar data with fully connected layers (FCN) in a point-wise manner and combine point features within a voxel with max pooling. Subsequently, we employ a CNN-based height compression to obtain the overall radar features in the BEV space.

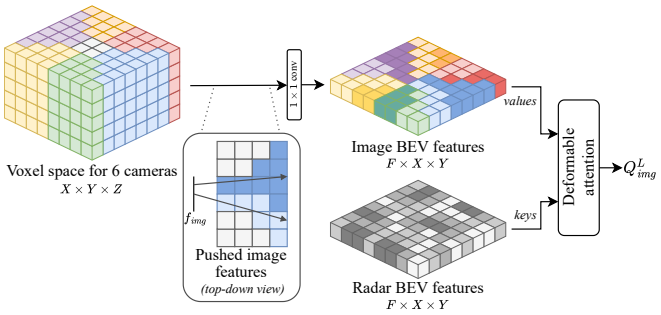


Fig. 4. Our data-driven query initialization scheme leverages 3D radar information to guide lifting the 2D image features to the BEV space. While the image BEV features are only obtained from uniform assignment along camera rays, the final $Q^L_{img}$ considers depth from radar via deformable attention.

defined by the BEV resolution $X \times Y$, an additional height discretization $Z$, and centered at the forward-facing camera. Second, we assign each voxel to one or two cameras based on their fields of view. Third, we push the vision features from the 2D image plane to the 3D voxel space via ray projection, i.e., each voxel within a frustum along a ray contains the same image features. In particular, we utilize the image features from scale ⅛. If the fields of view of two cameras overlap, we average the features in the affected voxels. Subsequently, we employ a $1 \times 1$ convolutional layer to remove the height component resulting in an $X \times Y$ voxel grid with $F$ feature channels. Note that at this stage, the image features are still uniformly distributed without a notion of depth. Therefore, we use deformable attention [23] guided by the sparse radar point cloud to filter the feature map resulting in the initialized query $Q^L_{img}$ of size $F \times X \times Y$.

*Lifting:* In the next step, we combine our data-driven initial queries $Q^L_{img}$ with a learnable position embedding $Q^L_{pos}$ to achieve permutation invariance and learnable BEV queries $Q^L_{bev}$ [4], [6]:

$$Q^L = Q^L_{img} + Q^L_{pos} + Q^L_{bev} \tag{1}$$

Employing deformable attention [23] again, we construct a 3D voxel space of size $X \times Y \times Z$ to pull the vision encoding from the images. In contrast to the query initialization, we now sample offsets on the image planes instead of the BEV space. After six cascaded transformer blocks, we obtain the final feature map $f_{img,bev}$ that has the same dimension as the encoded radar data, i.e., $F \times X \times Y$.

### C. Sensor Fusion

For fusing the lifted image features with the encoded radar data, we follow a scheme comparable to the lifting step. Inspired by TransFusion [7], which fuses camera and LiDAR for 3D object detection, we query image features in the surroundings of the radar points and extract the values via deformable attention [23]. Similar to Eq. (1), we form the initial query by summing the encoded radar data $f_{rad}$, a learnable position encoding $Q^F_{pos}$, and learnable BEV queries $Q^F_{bev}$:

$$Q^F = f_{rad} + Q^F_{pos} + Q^F_{bev} \tag{2}$$

Importantly, the lifted image features only serve as keys and values during the cross-attention step. In total, we utilize a cascade of six transformer blocks. Finally, to encode the features of both modalities in a joint manner, we feed the output of the last block through a ResNet-18 [32] bottleneck, referred to as BEV encoder in Fig. 2.

### D. Segmentation Head

We employ a single head for multi-class BEV semantic segmentation. In detail, we utilize two convolutional layers with ReLU activations followed by a final $1 \times 1$ convolutional layer to output one object class and $M$ map classes. Given the BEV space resolution, the segmentation head produces an output of size $(M+1) \times X \times Y$. Thus, a pixel can not only capture both a vehicle and a map class prediction but can also be assigned to multiple map categories.

*Object Segmentation:* For segmenting objects, we consider all vehicle-like entities, e.g., passenger cars and trucks. Unlike prior works [6], we emphasize that object-agnostic segmentation should not rely on instance-aware information during training time as this renders the application of a method less flexible due to requiring additional annotations. Therefore, we supervise the object channel of the segmentation head solely via the binary cross-entropy loss:

$$\mathcal{L}_{BCE} = \frac{-1}{N} \sum_{i=1}^{N} \log(p_{i,t}) \,, \tag{3}$$

where $p_{i,t}$ is defined per pixel $i \in [1, N]$ as:

$$p_{i,t} = \begin{cases} p_i & \text{if } y_i = 1 \\ 1 - p_i & \text{otherwise.} \end{cases} \tag{4}$$

The binary ground truth label $y_i \in \{0, 1\}$ specifies whether the pixel $i$ belongs to the vehicle class. The corresponding predicted probability for $y_i = 1$ is denoted by $p_i$.

*Map Segmentation:* While most previous methods [4], [8], [9] predict only the road and occasionally also lane dividers, we include further map classes such as pedestrian crossings and walkways. For an exhaustive list, please refer to Sec. IV-A. To supervise the map channels of the segmentation head during training, we employ a multi-class variant of the $\alpha$-balanced focal loss [36]:

$$\mathcal{L}_{FOC} = \sum_{c=1}^{C} \frac{-1}{N} \sum_{i=1}^{N} \alpha_{i,t} \, (1 - p_{i,t})^{\gamma} \log(p_{i,t}) \,, \tag{5}$$

TABLE I

BASELINE COMPARISON ON THE NUSCENES DATASET

| Method | Modalities | Image Backbone | Vehicle | Driv. Area | Lane | Map | mIoU |
|---|---|---|---|---|---|---|---|
| CVT [20] | C | EfficientNet | 36.0 | 74.3 | – | – | 55.2 |
| BEVFormer-S [4] | C | ResNet-101 | 43.2 | 80.7 | 21.3 | – | 62.0 |
| Simple-BEV [6] | C | ResNet-101 | 47.4 | – | – | – | – |
| Simple-BEV [6] | C+R | ResNet-101 | 55.7 | – | – | – | – |
| Simple-BEV++ | C+R | ResNet-101 | 52.7 | 77.7 | 35.8 | 46.1 | 65.2 |
| Simple-BEV++ | C+R | ViT-B/14 | 54.5 | 81.2 | 40.4 | 50.4 | 67.9 |
| BEVGuide [8] | C+R | EfficientNet | **59.2** | 76.7 | <u>44.2</u> | – | 68.0 |
| CRN [9] | C+R (+L) | ResNet-50 | <u>58.8</u> | <u>82.1</u> | – | – | <u>70.5</u> |
| BEVCar (*camera*) | C | ViT-B/14 | 48.8 | 81.1 | 40.6 | 50.5 | 65.0 |
| BEVCar (*ResNet*) | C+R | ResNet-101 | 57.3 | 81.8 | 43.8 | <u>53.0</u> | 69.6 |
| BEVCar (*ours*) | C+R | ViT-B/14 | 58.4 | **83.3** | **45.3** | **54.3** | **70.9** |

We compare BEVCar with both camera-only (C) and camera-radar (C+R) BEV segmentation methods on the nuScenes [2] validation split. Simple-BEV++ is a customized Simple-BEV [6] without instance-aware training but with the same radar metadata and map segmentation head as our method. Note that CRN [9] uses LiDAR during training. The "map" metric averages the IoU of all nuScenes map classes. Previous works report predictions for fewer classes, indicated by "–". To compare BEVCar with these methods, we provide the mean of "vehicle" and "drivable area" classes as "mIoU". Bold and underlined values denote the best and second-best metrics per column, respectively.

where $c \in [1, C]$ refers to the semantic classes and $\gamma$ is a focusing parameter to differentiate between easy/hard examples. Additionally, $\alpha_{i,t}$ is defined analogously to Eq. (4):

$$\alpha_{i,t} = \begin{cases} \alpha & \text{if } y_i = 1 \\ 1 - \alpha & \text{otherwise,} \end{cases} \tag{6}$$

with tunable parameter $\alpha$ to address the foreground-background imbalance.

## IV. EXPERIMENTAL EVALUATION

In this section, we outline the experimental setup and compare our BEVCar approach to various baselines. We further analyze the impact of the components of our method and demonstrate the advantage of radar measurements over vision-only methods under adverse conditions.

### A. Experimental Settings

We introduce the utilized dataset and metrics for evaluation and provide further implementation details.

*Dataset and Metrics:* We evaluate our BEVCar approach on the nuScenes dataset [2] for automated urban driving in Singapore and Boston, MA, being the only publicly available dataset that provides the required sensor data and ground truth map annotations. The nuScenes dataset comprises surround-view vision from six RGB cameras and five automotive radars and provides BEV map information. For training and evaluation, we use the official training/validation split, containing 28,130 and 6,019 samples, respectively. We further categorize the validation scenes into day (4,449 samples), rain (968 samples), and night (602 samples) scenes and include this split in our code release. For object segmentation, we combine all subclasses of the "vehicle" category. For map segmentation, we consider all available classes, i.e., "drivable area", "carpark area", "pedestrian crossing", "walkway", "stop line", "road divider", and "lane divider". We report individual intersection over union (IoU) metrics [37] for those classes that have been addressed by prior works and refer to the mean IoU of all map classes by "map". To compare BEVCar with previous baselines that predict fewer classes, we report the average of "vehicle" and "drivable area" as "mIoU".

*Implementation Details:* Similar to related work [6], [8], [9], our BEV grid covers an area of $100\,\mathrm{m} \times 100\,\mathrm{m}$ centered at the ego vehicle and is discretized at a resolution of $200 \times 200$ cells. We further construct an up/down span from the ground to $10\,\mathrm{m}$ in height and discretize it into eight bins. The resulting 3D tensor is oriented with respect to the forward-facing camera serving as the reference coordinate system. For both training and inference, we resize the images of the six cameras to $448 \times 896$ pixels adapting the results of an analysis from Harley *et al.* [6] to the requirements of the employed ViT adapter. In accordance with the released code of the same study, we aggregate five radar sweeps as input. During training, we set the parameters of the focal loss (see Eq. (5)) to $\alpha = 0.25$ and $\gamma = 3$.

### B. Quantitative Results

We compare BEVCar to various baseline works in Tab. I, including the camera-radar fusion methods Simple-BEV [6], BEVGuide [8], and CRN [9], which leverages depth from LiDAR during training. At the time of submission, only the authors of Simple-BEV released their code. We utilize this for an extended version Simple-BEV++ by adding the BEV map segmentation task, removing additional radar metadata (see Sec. III-A), and disregarding the instance-aware losses (see Sec. III-D). To demonstrate the advantage of radar measurements, we further compare BEVCar to the vision-only baselines CVT [20], BEVFormer [4], and variations of both Simple-BEV [6] and our proposed BEVCar.

Concerning the latter, our camera-only version of BEVCar yields a small increase of performance over the Simple-BEV (C) baseline for the "vehicle" class (+1.4 IoU) and over the static version of BEVFormer for the "drivable area" class (+0.4 IoU). We primarily attribute the improvement within the vision-only regime to the semantically rich image representation of the DINOv2 [26] backbone. Integrating radar data via our proposed methodology results in substantially enhanced vehicle predictions (+9.6 IoU) and

TABLE II

COMPONENTS ANALYSIS

| Method | Vehicle | Map |
|---|---|---|
| *Radar encoding* | | |
| No radar encoding [6] | 57.8 | 53.4 |
| BEVCar (*ours*) | 58.4 (*+0.6*) | 54.3 (*+0.9*) |
| *Lifting and fusion* | | |
| Parameter-free [6] | 56.6 | 50.1 |
| BEVCar (*ours*) | 58.4 (*+1.8*) | 54.3 (*+4.2*) |

We demonstrate the efficacy of our employed radar encoding and our attention-based lifting and fusion scheme compared to simpler approaches. The "map" metric denotes the mean IoU of all map classes.

noteworthy improvements in map segmentation (+3.8 mIoU). We thus infer that utilizing radar for robotic perception promises significantly better performance and further analyze this claim in Sec. IV-C under various aspects.

For the vehicle segmentation task, BEVCar outperforms Simple-BEV (+2.7 IoU) and achieves comparable performance to BEVGuide (−0.8 IoU) and CRN (−0.4 IoU). Concerning CRN, it is important to consider that this method relies on LiDAR during the training phase to learn metric depth. For map segmentation, BEVCar improves upon all baselines while providing information for more semantic classes. With respect to the combined evaluation for both tasks, BEVCar achieves the highest performance across the board with +2.9 mIoU versus BEVGuide and +0.4 mIoU versus CRN. We further compare BEVCar to the aforementioned Simple-BEV++. To eliminate the impact of different backbones, we integrate both ResNet-101 [32] and DINOv2 ViT-B/14 [26] in either method. Note that the multi-task training of Simple-BEV++ leads to reduced performance for vehicle segmentation over the Simple-BEV baseline. Although we observe that the DINOv2 backbone also improves the results of Simple-BEV++, our BEVCar approach still outperforms Simple-BEV++ with both image backbones ResNet-101 (+4.4 mIoU) and ViT-B/14 (+3.0 mIoU), demonstrating the novelty of our method.

In Fig. 5, we underline this observation by visualizing the improvements and errors of BEVCar compared to Simple-BEV++. We further show the ground truth BEV object and map segmentation and provide visual predictions from the camera-only baseline, Simple-BEV++, and our BEVCar approach. For a detailed analysis of the different weather and illumination conditions, please refer to the next section.

### C. Ablations and Analysis

To further analyze our proposed BEVCar approach, we provide ablations for its components and evaluate its performance under challenging conditions.

*Components Analysis:* We evaluate the impact of two key components of BEVCar, i.e., the proposed radar point encoding and the new radar-driven image feature lifting, and report the improvements over baselines inspired by Simple-BEV [6] in Tab. II. First, compared to utilizing the raw radar data without a learning-based encoding, our approach yields +0.6 IoU and +0.9 mIoU for the vehicle and map segmentation tasks, respectively. Second, while the baseline uses a parameter-free lifting of the image features to the BEV space, our attention-based scheme leverages radar information already during the lifting stage. In comparison, this results in an increase of +1.8 IoU for vehicle segmentation and +4.2 mIoU for map segmentation.

TABLE III

VEHICLE PERCEPTION RANGE

| Method | Modalities | 0-50 m | Range Intervals 0-20 m | 20-35 m | 35-50 m |
|---|---|---|---|---|---|
| BEVCar (*camera*) | C | 48.8 | 68.7 | 45.8 | 29.1 |
| Simple-BEV [6] | C+R | 55.5 | 70.1 | 53.6 | 39.1 |
| Simple-BEV++ | C+R | 54.5 | 71.4 | 52.4 | 36.6 |
| BEVCar (*ours*) | C+R | **58.4** | **74.0** | **56.5** | **41.0** |

The modalities denote camera (C) and radar (R) input. Simple-BEV++ and BEVCar use a ViT-B/14 image backbone. For Simple-BEV, we utilize the model trained by the authors.

TABLE IV

WEATHER AND ILLUMINATION ANALYSIS

| Method | Day Vehicle | Day Map | Rain Vehicle | Rain Map | Night Vehicle | Night Map |
|---|---|---|---|---|---|---|
| BEVCar (*camera*) | 48.7 | 53.3 | 49.9 | 45.1 | 42.1 | 39.2 |
| Simple-BEV [6] | 55.5 | – | 55.8 | – | 54.0 | – |
| Simple-BEV++ | 54.4 | 53.1 | 54.8 | 45.6 | 52.4 | 37.8 |
| BEVCar (*ours*) | **58.3** | **57.3** | **59.1** | **48.8** | **57.4** | **42.4** |

"Map" denotes the mean IoU of all map classes. Simple-BEV++ and BEVCar use a ViT-B/14 image backbone. For Simple-BEV, we utilize the model trained by the authors, which does not perform map segmentation.

*Distance-Based Object Segmentation:* In Tab. III, we analyze the vehicle segmentation quality of BEVCar, its camera-only variant, Simple-BEV [6], and Simple-BEV++ for three different range intervals including 0-20 m, 20-35 m, and 35-50 m. Note that the overall performance of Simple-BEV is slightly lower than reported in Tab. I due to rerunning the authors' code to enable the range-based evaluation. Generally, we observe that the results of the camera-only baseline significantly differ between the evaluation criteria. While the IoU in the 0-20 m range is comparable to Simple-BEV, for the 35-50 m range it achieves only half of the initial performance. Although the general trend is similar among all camera-radar methods, the effect is the least severe for BEVCar. Our experiment demonstrates the advantage of utilizing radar measurements to maintain object segmentation performance also at larger distances.

*Robustness to Weather and Illumination:* Besides providing complementary information, i.e., dense RGB data versus sparse distance and velocity measurements, a core difference between cameras and radars is the source of energy utilized by the respective sensor. While passive sensors such as cameras rely on an external source like the sun, active sensors such as radars provide their own energy. Therefore, passive sensors suffer from challenging illumination conditions, e.g., faced during rain or at night. We thus emphasize that evaluating automotive perception systems specifically in these situations is imperative to understand their performance fully.

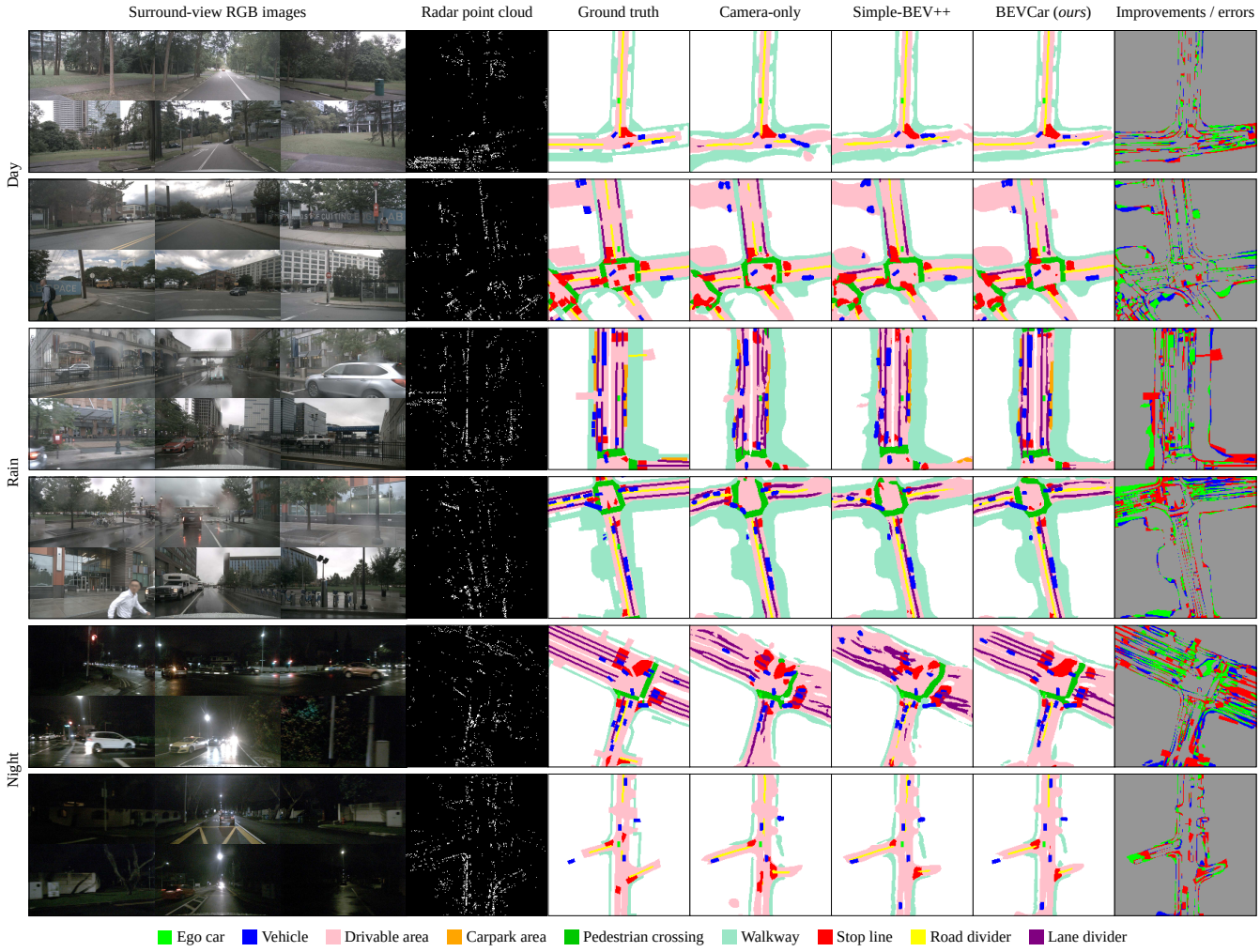


Fig. 5. Qualitative results of our proposed BEVCar, the camera-only baseline, and Simple-BEV++ (ViT-B/14), for which we also show the improvement/error map. Pixels misclassified by Simple-BEV++ and correctly predicted by BEVCar are shown in green, pixels misclassified by BEVCar and correctly predicted by Simple-BEV++ in blue, and pixels misclassified by both models in red.

TABLE V
RUNTIME ANALYSIS

| Method | Image Backbone | Deformable Attention | Runtime ms | Runtime FPS |
|---|---|---|---|---|
| Simple-BEV [6] | ResNet-101 | | 18 | 54.9 |
| BEVCar | ResNet-101 | ✓ | 137 | 7.3 |
| BEVCar (*camera*) | ViT-B/14 | ✓ | 352 | 2.8 |
| Simple-BEV++ | ViT-B/14 | ✓ | 277 | 3.6 |
| BEVCar (*ours*) | ViT-B/14 | ✓ | 382 | 2.6 |

Runtime of a forward pass measured on an Nvidia A100 GPU.

In Tab. IV, Fig. 1, and Fig. 5, we separate the previously reported metrics for BEVCar and the same baselines as in the study on the perception range into *day*, *rain*, and *night*. We observe that the vehicle segmentation IoU of the camera-only baseline is subject to substantial degradation during the night. In contrast, all camera-radar methods can maintain their performance, whereas BEVCar achieves the highest performance. On the other hand, the map segmentation mIoU decreases during rain and even further at night, which holds for all investigated methods. The results indicate that radar is most beneficial for object detection and less relevant for BEV mapping, which is expected as depth information is less important for learning a mapping of the planar map classes from the 2D image space to the BEV space than mapping objects with defined height, width, and depth parameters.

*Runtime Analysis:* In Tab. V, we report the runtime of our proposed BEVCar and multiple baseline methods, measured on an Nvidia A100 GPU and averaged over the validation split. In contrast to Harley *et al.* [6], we only consider the forward pass without data loading and loss calculation. Most notable is the slow-down caused by deformable attention [23] employed in our proposed lifting module as well as in the ViT adapter. Importantly, in comparison to the vision-only baseline, including the radar information does not result in a significantly higher runtime. Note that the frequency of the synchronized keyframes in the nuScenes [2] dataset is 2 Hz.

## V. CONCLUSION

In this work, we introduced BEVCar addressing camera-radar fusion for BEV map and object segmentation. BEVCar comprises a new learning-based radar point encoding and leverages radar information early during the lifting step of

the vision features from the image plane to the BEV space. We demonstrated that BEVCar outperforms previous camera-radar approaches when jointly considering map and object segmentation. We extensively evaluated the performance in challenging weather and illumination conditions and analyzed the robustness for various perception ranges. Our results clearly demonstrate the benefit of utilizing automotive radar in addition to surround-view vision. To facilitate further research in this direction, we include our day/rain/night split of the nuScenes [2] validation data in the public release of our code. In the future, we will address robustness in case of partial or complete sensor failure, e.g., by leveraging cross-modality distillation during training of the network.

## REFERENCES


[1] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The Oxford Radar RobotCar dataset: A radar extension to the Oxford RobotCar dataset," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 6433–6438.

[2] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 618–11 628.

[3] L. Zheng, Z. Ma, X. Zhu, B. Tan, S. Li, K. Long, W. Sun, S. Chen, L. Zhang, M. Wan, L. Huang, and J. Bai, "TJ4DRadSet: A 4D radar dataset for autonomous driving," in *IEEE International Conference on Intelligent Transportation Systems*, 2022, pp. 493–498.

[4] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European Conference on Computer Vision*, 2022.

[5] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "BEVFusion: A simple and robust LiDAR-camera fusion framework," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 10 421–10 434.

[6] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, "SimpleBEV: What really matters for multi-sensor BEV perception?" in *IEEE International Conference on Robotics and Automation*, 2023, pp. 2759–2765.

[7] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1080–1089.

[8] Y. Man, L.-Y. Gui, and Y.-X. Wang, "BEV-guided multi-modality fusion for driving perception," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 960–21 969.

[9] Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi, and D. Kum, "CRN: Camera radar net for accurate, robust, efficient 3D perception," in *International Conference on Computer Vision*, 2023, pp. 17 569–17 580.

[10] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *European Conference on Computer Vision*, 2020, pp. 194–210.

[11] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "BEVerse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.

[12] N. Hendy, C. Sloan, F. Tian, P. Duan, N. Charchut, Y. Xie, C. Wang, and J. Philbin, "FISHING Net: Future inference of semantic heatmaps in grids," *arXiv preprint arXiv:2006.09917*, 2020.

[13] C. Lu, M. J. G. van de Molengraft, and G. Dubbelman, "Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 445–452, 2019.

[14] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

[15] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 135–11 144.

[16] N. Gosala and A. Valada, "Bird's-eye-view panoptic segmentation using monocular frontal view images," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1968–1975, 2022.

[17] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving," in *International Conference on Learning Representations*, 2020.

[18] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9087–9098.

[19] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *IEEE International Conference on Robotics and Automation*, 2022, pp. 9200–9206.

[20] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 750–13 759.

[21] R. Trumpp, M. Büchner, A. Valada, and M. Caccamo, "Efficient learning of urban driving policies using bird's-eye-view state representations," in *IEEE International Conference on Intelligent Transportation Systems*, 2023, pp. 4181–4186.

[22] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "BEVSegFormer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5924–5932.

[23] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," *International Conference on Learning Representations*, 2021.

[24] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance multi-camera 3D object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[25] N. Gosala, K. Petek, P. L. J. Drews-Jr, W. Burgard, and A. Valada, "SkyEye: Self-supervised bird's-eye-view semantic mapping using monocular frontal view images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 901–14 910.

[26] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, *et al.*, "DINOv2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[27] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.

[28] I. T. Kurniawan and B. R. Trilaksono, "ClusterFusion: Leveraging radar spatial features for radar-camera 3D object detection in autonomous vehicles," *IEEE Access*, vol. 11, pp. 121 511–121 528, 2023.

[29] Z. Yu, W. Wan, M. Ren, X. Zheng, and Z. Fang, "SparseFusion3D: Sparse sensor fusion for 3D object detection by radar and camera in environmental perception," *IEEE Intelligent Vehicles Symposium*, 2023.

[30] O. Schumann, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic segmentation on radar point clouds," in *International Conference on Information Fusion*, 2018, pp. 2179–2186.

[31] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019, pp. 6105–6114.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[33] M. Käppeler, K. Petek, N. Vödisch, W. Burgard, and A. Valada, "Few-shot panoptic segmentation with foundation models," in *IEEE International Conference on Robotics and Automation*, 2024, pp. 7718–7724.

[34] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "SAN: Side adapter network for open-vocabulary semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 546–15 561, 2023.

[35] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," in *International Conference on Learning Representations*, 2023.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[37] J. V. Hurtado and A. Valada, "Semantic scene segmentation for robotics," in *Deep Learning for Robot Perception and Cognition*, 2022, pp. 279–311.