

On the Impact of Dataset Quality in Multiview BEV Mapping

Alaa Eddine Ben Zekri , Reza Bahmanyar , and Houda Chaabouni-Chouayakh 

Abstract—Bird’s eye view (BEV) mapping from multicamera systems has emerged as an important form of ground-level remote sensing for constructing geospatial representations in autonomous navigation and urban analytics. Although recent advances in deep learning architectures have expanded the capabilities of BEV mapping, the impact of data-quality factors on mapping performance remains insufficiently explored. In this work, we investigate how fundamental dataset characteristics, such as spatial coverage, ground sampling resolution, annotation quality, and cross-domain differences impact BEV map accuracy, boundary fidelity, and robustness. Using the nuScenes dataset as a primary case study, we systematically vary these data-quality attributes and quantify their effects using geospatial accuracy metrics. To ensure the generality of our findings, all experiments are conducted using two architecturally distinct models, BEVCar and SimpleBEV, and the observed trends are consistent across both. Our results show that reducing BEV map coverage consistently improves accuracy by limiting predictions to visible regions. In addition, we found that resolution and class granularity jointly affect boundary quality; finer annotations at higher resolutions enhance precision for small structures. Once the previous quality factors were evaluated, we created a synthetic dataset under conditions similar to nuScenes to assess generalization under domain shifts and evaluate cross-domain performance between real and synthetic environments. We highlighted that such data augmentation reduces the tendency to hallucinate unseen objects. By quantifying the data-quality drivers of mapping performance, this work offers general insights for the design, curation, and quality assessment of AI-based remote sensing datasets used for geospatial mapping and perception, which are essential for sustainable mobility and resilient urban planning.

Index Terms—Autonomous driving, bird’s eye view (BEV) mapping, domain shift, geospatial data quality, synthetic dataset, urban mobility.

I. INTRODUCTION

ACCURATE extraction of geospatial information has become fundamental to modern remote sensing applications. It underpins tasks, such as spatial segmentation, mapping, scene understanding, and autonomous navigation. As cities grow denser and more dynamic, generating precise maps is essential for sustainable urban development. This enables resilience

planning, infrastructure monitoring, and data-driven decision-making. However, extracting fine-grained spatial information from aerial or satellite imagery alone is challenging due to occlusions, limited vertical resolution, and the inability to capture street-level context, which is critical for many geospatial tasks. In this context, bird’s eye view (BEV) mapping has gained particular attention. A BEV map is a top-down spatial representation synthesized from ground-level sensor observations that unifies ground-level perception with top-down geospatial analysis.

Recent deep learning methods based on convolutional neural network architectures have shown strong potential in learning complex spatial representations directly from sensor data [1]. One technique that is gaining particular attention is BEV map generation from multicamera input. In BEV mapping, a model receives images of the surrounding area captured by multiple cameras mounted on a vehicle and produces a top-down semantic segmentation map of the environment. This map provides a spatially consistent view that captures the layout of the road, lanes, sidewalks, and other static infrastructure, as well as dynamic objects, such as vehicles, from a bird’s-eye perspective. This representation is highly beneficial for ego-localization and broader perception tasks, such as urban planning and navigation [2]. Since these approaches are inherently data-driven, the quality, diversity, and structure of the underlying datasets play a crucial role in shaping model accuracy and robustness. Despite rapid advances in BEV mapping, however, limited attention has been paid to how dataset properties, such as resolution affect model performance, generalization, and real-world applicability.

In this work, we aim to explore and evaluate the impact of various dataset-related factors on BEV map prediction performance. The nuScenes dataset [3] has become the standard benchmark in this field, so we use it as the foundation for our experiments to identify broader insights and challenges that can inform the design and usage of future datasets in model training and evaluation. In addition, since access to large-scale, well-annotated, real-world datasets is limited, we created a synthetic dataset with the same sensor configuration and annotation structure as the nuScenes dataset. This follows the approach introduced in our previous work on synthetic-data-driven aerial image segmentation [4], which demonstrated the complementary value of controlled simulation, but extends it here to a multiview setting tailored for BEV map generation. This dataset enables us to assess the generalization of models from real to simulated environments and to investigate the broader role of simulated data as a complementary alternative for training, augmentation,

Received 20 December 2025; revised 23 April 2026; accepted 8 June 2026.
Date of publication 10 June 2026; date of current version 29 June 2026.
(Corresponding author: Alaa Eddine Ben Zekri.)

Alaa Eddine Ben Zekri and Reza Bahmanyar are with Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: alaa.benzekri@dlr.de; reza.bahmanyar@dlr.de).

Houda Chaabouni-Chouayakh is with Sm@rts laboratory, Digital Research Center of Sfax, Sfax 3021, Tunisia (e-mail: houda.chaabouni@crns.rnrt.tn).

Digital Object Identifier 10.1109/JSTARS.2026.3702298



Fig. 1. Samples from the nuScenes (top) and Synthetic (bottom) datasets: ground view (left) and $100 \times 100 \text{ m}^2$ BEV map (right). In the ground views, the top images show the front of the ego-vehicle and the bottom images show the rear. The color in the synthetic sample represents all kinds of road markings.

and domain transfer, particularly when real data collection is costly, incomplete, or constrained. Fig. 1 shows example images and their corresponding BEV maps from the nuScenes and our synthetic datasets. Specifically, we investigate the following key questions.

- 1) Does reducing spatial coverage improve the accuracy of the BEV mapping model, particularly by avoiding predictions in regions that are not visible to the ground sensors?
- 2) Does higher ground sampling distance (GSD)—spatial resolution representing the size of one pixel on the ground—lead to more precise BEV maps, or does it introduce noise and instability, particularly in boundary-level predictions?
- 3) How do class definitions and annotation quality affect model performance, particularly for fine-grained versus large-scale infrastructure classes?
- 4) Do BEV mapping models generalize reliably in both the real-to-synthetic and synthetic-to-real directions, or do dataset-specific biases hinder their robustness and limit the role of synthetic data as a complementary alternative training source?

The rest of this article is organized as follows. Section II reviews related work on BEV mapping and the existing datasets. Section III describes the datasets used in our study, including nuScenes and the synthetic dataset we generated. Section IV describes the experimental setup. Section V presents the quantitative and qualitative results. Section VI discusses their implications. Section VII outlines limitations and directions for future research. Finally, Section VIII concludes this article.

II. RELATED WORK

In this section, we present key developments that have shaped the BEV mapping task, and identify the different datasets used to train and evaluate said methods.

Initial works in this field started by using a single-view as input to estimate the BEV map, which is the case of Lu et al.'s approach [5] that employed a variational encoder-decoder architecture to map features from front-view image to a semantic map without using any geometric modeling or inferring any depth information. PON [6] is another work that used a single-view as input, however, it uses multiple scales to map features at different distances from the vehicle in the BEV using dense transformer layers. Trying to improve on the latter method, Gosala and Valada [7] separated the front-view into vertical and flat surfaces before projecting them into the BEV plane and processing them together. The use of attention-mechanisms in this task was introduced in TIIM [8] as they developed a model that attends image vertical scanlines to BEV map polar rays treating it as a sequence-to-sequence translation task. While these single-view approaches established the feasibility of learning-based BEV mapping, their reliance on a single viewpoint limits spatial coverage and increases depth ambiguity, motivating the shift toward multiview inputs.

Another direction in the literature uses multiview images as input instead of single-view. VPN's approach [9] consists of getting feature maps from the different views then using a transformation module which extracts the relations between the different views to predict a final BEV map. A popular approach is

that of LSS [10], which lifts each view by estimating a contextual depth distribution for each of its pixels, then joins them together in a single point-cloud using geometric modeling based on camera parameters of the different views, which would be splatted to obtain the BEV map. MapPrior [11] improves on this method by adding a VQGAN to enhance noisy initial map predictions. SimpleBEV [12] takes a contrasting approach by projecting image features directly onto a 3-D voxel grid without explicit depth estimation, demonstrating that a simpler fusion strategy can achieve competitive performance. The authors of CVT [13] aim to avoid such explicit geometric modelling through the use of attention mechanisms, as they use spatial attention across different views and incorporate the camera parameters through the positional encoding. BEVFormer [14] and BEVSegFormer [15] continued in this direction, however, they employed deformable attention to reduce computational overhead, with BEVSegFormer being more flexible as they do not require camera parameters. BEV-Car's [16] method combines some of the aforementioned ideas; it uses the lifting technique however it employs deformable attention to get the image features to the BEV space. In addition, it uses a frozen DINOv2 [17] model with a ViT Adapter [18] to be able to extract rich features from the camera data while limiting the computational needs. Despite their architectural diversity, these methods uniformly treat the training dataset as a fixed given, with little attention paid to how dataset properties, such as spatial coverage, resolution, or annotation quality affect model performance and generalization.

Early works on BEV mapping relied on datasets, such as CityScapes [19] and KITTI [20], which lacked ground-truth BEV semantic maps. To overcome this limitation, approximate labels were generated by projecting front-view semantic annotation combined with disparity or depth maps. However, this indirect supervision limited the accuracy and consistency of the resulting BEV predictions. The release of the nuScenes [3] dataset marked a major step forward as it provides a map annotation along with its rich sensor data. As a result, it has become the standard benchmark for training and evaluating BEV mapping methods. Although other datasets, such as Argoverse [21], offer similar multisensor setup, they suffer from higher GSD and a fewer annotated classes in their provided maps, limiting their utility for detailed semantic mapping. However, the design choices embedded in nuScenes, including its fixed spatial coverage, default GSD, and annotation inconsistencies, have received little systematic attention in terms of their direct influence on model performance.

In this work, we address this gap by conducting a dataset-centric analysis of BEV map prediction. Rather than proposing a new architecture, we systematically investigate how key dataset characteristics, namely spatial coverage, GSD, class configuration and domain composition, affect model accuracy, boundary quality, and generalization. We evaluate these effects across two architecturally distinct models, BEVCar [16] and SimpleBEV [12], to ensure the observed trends are not architecture-specific. To complement the real-world analysis, we also introduce a synthetic dataset and study cross-domain generalization, an aspect that has remained largely unexplored due to the scarcity of compatible public datasets.

III. DATASETS

In this section, we first provide an overview of the nuScenes dataset, followed by a description of the generation process and key statistics of a synthetic dataset we developed using a comparable setup. These two datasets form the foundation of our experimental framework, which we detail in the following sections.

A. nuScenes Dataset

The nuScenes dataset [3] is a widely used benchmark in autonomous driving research. It offers diverse, sensor-rich data for developing and evaluating perception algorithms. Released in 2019, the dataset contains 1000 urban driving scenes from Boston and Singapore. Each scene lasts 20 s and is annotated at 2 Hz. This results in over 28 000 training samples, 6000 validation samples, and 6000 test samples. The dataset features a comprehensive sensor suite, including a 32-beam LiDAR, six RGB cameras, and multiple radars, which provide full 360° coverage. The nuScenes dataset is designed for tasks, such as object detection, tracking, and motion forecasting. It captures a wide range of urban environments, weather conditions, lighting conditions, and traffic scenarios. Its annotated BEV maps include classes, such as *drivable area*, *pedestrian crossing*, *walkway*, *stop line*, *car park*, *Road divider*, and *Lane divider*. Fig. 1 shows an example from the nuScenes dataset including images from the surrounding RGB cameras and the ground truth BEV map covering $100 \times 100 \text{ m}^2$ around the vehicles.

This dataset is a valuable resource for the research and development in the BEV map segmentation and ego-localization domains. However, two key limitations restrict its applicability for high-precision mapping and localization. First, the BEV ground truth maps are provided at a default GSD of 50 cm/pixel, which is too coarse to reliably represent fine-grained objects like road markings and to train models for precise localization or detailed semantic mapping. Second, while initial positions and sensor poses are provided, the lack of accurate GPS coordinates prevents reliable vehicle trajectory reconstruction.

B. Our Synthetic Dataset

For generating our synthetic dataset, we use the CARLA simulator [22] for its realistic visuals and ease of sensor integration.

We chose four predefined towns (Town01, Town02, Town03, and Town10) in CARLA based on their diversity and urban scenery, the layout of which can be observed in Fig. 2. We also considered two weather/time presets as shown in Fig. 3: clear noon and cloudy sunset. These presets cover some of the more likely conditions that would be captured in real life.

Having a set of predefined vehicle spawn-points provided by CARLA, a first intuition to reduce computational load would be to spawn a limited number of vehicles over those spawn-points and simultaneously extract data from them. Each town will thus be subdivided into smaller regions, defined by spawn points' positions. The arrangement of the points within each town is carried out using the k -means clustering algorithm. First, the

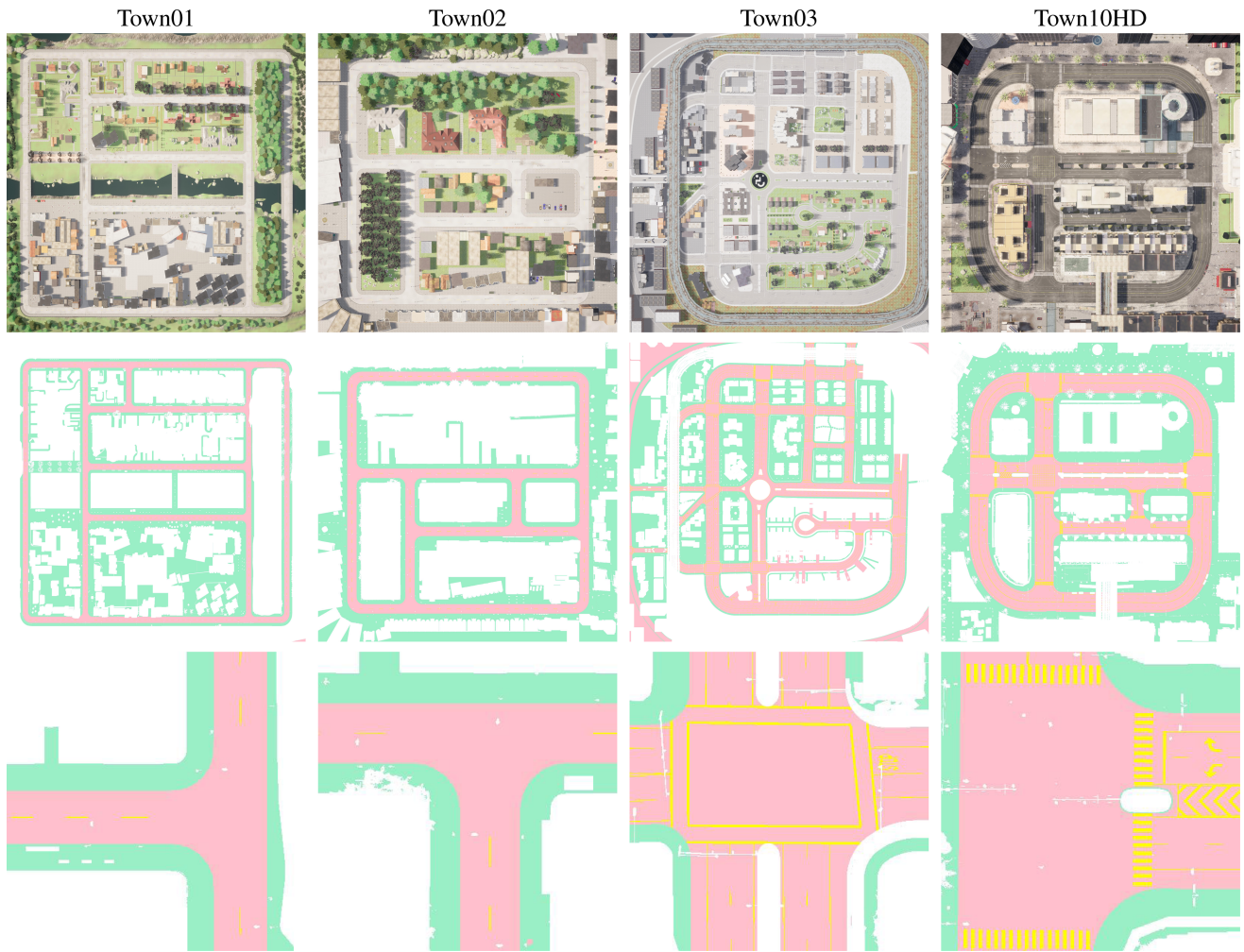


Fig. 2. Aerial view and semantic annotation of the used maps in our dataset generation. First row is an aerial view of the town, second row is its semantic and third row is a zoom-in on a part of the town; Drivable area in ■ and Walkway in ■ and Road Markings in ■.



Fig. 3. Two front-view images taken at the same location with two different weather conditions: Clear noon and cloudy sunset.

optimal number of clusters is identified using the silhouette score method [23]. Then, the final partitioning is achieved. This approach ensures balanced spawn-point distribution, minimizing the computational strain and optimizing utilization for GPU rendering. Each simulated vehicle is equipped with six RGB cameras (front, front-right, front-left, back, back-right, and

back-left), a top-mounted BEV semantic segmentation camera, and a GPS sensor to capture vehicle-related data. The overall dataset generation is described in Algorithm 1.

Each generated BEV map covers an area of $30 \times 30 \text{ m}^2$ with a GSD of 10 cm. These maps include classes, such as *roads*, *walkways*, *road markings*, *traffic poles*, *vegetation*, and

Algorithm 1: Vehicle-Related Dataset Generation.

```

1: for town in towns do
2:   for weather in weathers do
3:     for region in regions do
4:       vehicles  $\leftarrow$  []
5:       for spawn_point in region do
6:         vehicle  $\leftarrow$  SPAWNVEHICLE(spawn_point)
7:         ATTACHSENSORS(vehicle)
8:         vehicles.APPEND(vehicle)
9:       for tick in range(10) do
10:        WORLD TICK()  $\triangleright$  Vehicles advance by 5 meters
11:        CAPTUREDATA(vehicles)
12:        DESTROY(vehicles)

```

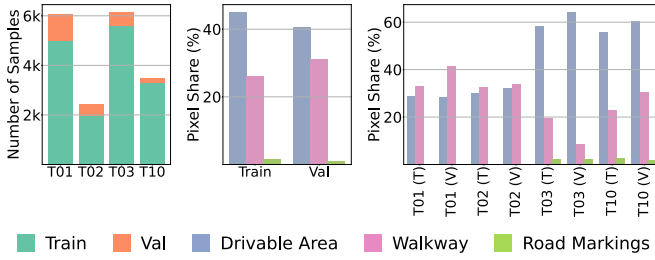


Fig. 4. Statistics from our Synthetic Dataset; distribution of samples across different towns as well as the class distribution of pixels across sets and across towns (“T” and “V” referring to Training and Validation sets, respectively).

buildings. Fig. 1 shows an example of the synthetic data. To make it comparable to the nuScenes dataset, we only include the *road*, *walkway*, and *road markings* classes.

There are a total of 18 000 samples, 12% of which constitute the validation set. Samples were split into training and validation sets based on the vehicle’s position. Those located in the south-eastern area of a town, constituting 16% of the total town area, were designated for the validation set. The remaining samples were assigned to the training set.

We take this measure to prevent data leakage, whereby scenes or maps from the training set may appear in the validation set. This was not taken into consideration in the official nuScenes [3] split, as mentioned by Roddick and Cipolla [6]. Fig. 4 shows statistics on our Synthetic dataset.

To assess the compatibility of our synthetic dataset with nuScenes, we verify alignment across three dimensions. First, regarding sensor configuration: the six RGB cameras are placed at equivalent positions around the ego-vehicle, using the same front, front-right, front-left, back, back-right, and back-left arrangement as nuScenes, ensuring that the model receives structurally comparable multiview input. Second, regarding annotation structure: BEV maps are generated at the same spatial extent and resolution used in our experiments ($30 \times 30 \text{ m}^2$ at 10 cm/pixel), and the *Drivable* and *Walkway* class definitions are kept consistent with those in nuScenes. Unlike nuScenes, annotations are derived directly from the simulator’s semantic segmentation output, making them noise-free and geometrically accurate. Third, regarding data distribution: the class statistics shown in Fig. 4 confirm that both classes are consistently

represented across all towns and in both splits, with no severe imbalance. A known limitation is the visual domain gap between CARLA-rendered imagery and real-world data, which we analyze quantitatively in Section V-D.

In the experiments related to the synthetic data we omit the road markings class due to the difference in definition between the two datasets; in the synthetic data all kinds of road markings are grouped into one class that is hard to separate, meanwhile in nuScenes they are divided into different classes with inconsistent definition which we discuss in further details in Section V-C.

IV. EXPERIMENTAL SETUP

We design experiments to evaluate how key dataset characteristics affect models that generate precise BEV maps from six surrounding RGB images.

To ensure a fair comparison across different experiments, we use a consistent training setup. Each model is trained for 6000 iterations with five gradient accumulation steps and a batch size of 2 on a NVIDIA A100 accelerator.

We evaluate performance using both mask intersection over union (IoU) and Boundary IoU [24]. Mask IoU measures the overlap between predicted and ground-truth class regions, where higher values indicate better agreement between the predicted segmentation and the ground-truth. Boundary IoU instead focuses on the overlap of boundary regions, making it more sensitive to edge accuracy. This is particularly important in our setting, since differences in GSD and annotation quality directly affect the precision of class boundaries, especially for fine structures, such as road dividers, lane markings, and stop lines.

For all experiments, we consider camera-only BEV mapping models and focus on the map segmentation task, which predicts traffic-related infrastructure, such as roads and walkways, rather than dynamic objects, such as vehicles. This allows us to isolate the impact of dataset characteristics without introducing additional complexity from multimodal fusion.

We evaluate two models: BEVCar [16] and SimpleBEV [12]. For BEVCar, we follow the original setup of the authors, including the same hyperparameters, a final input image size of 448×896 pixels, and augmentation techniques, such as camera shuffling. For SimpleBEV, we use the architecture proposed by the authors, while keeping the same training protocol and hyperparameters as for BEVCar to ensure a fair comparison between the two models.

The use of two architecturally different models allows us to assess whether the observed trends are model-specific or consistent across different BEV mapping approaches.

Unless otherwise specified, all experimental factors other than the variable under study (e.g., GSD, coverage, or class configuration) are kept fixed across experiments.

Following common practice in the literature, models are trained on the training split of nuScenes [3] and evaluated on the validation split. We use the same protocol for experiments involving synthetic data, ensuring consistent comparisons between real-only and mixed-domain settings.

TABLE I

COMPARISON OF THE BEV MAP PREDICTIONS ACROSS DIFFERENT CLASSES AND THE MEAN FOR THE BEVCAR MODEL TRAINED TO PREDICT TWO DIFFERENT COVERAGE AREAS ON THE NUSCENES DATASET

Cov. [m]	Mean	Driv.	Carp.	Ped. Cros.	Walkway	Stop Line	R. Div.	L. Div.
IoU								
30 × 30	56.1	90.6	55.4	58.9	63.3	30.4	51.4	43.1
100 × 100	48.9	89.0	44.7	48.3	58.6	22.6	43.1	36.1
Boundary IoU								
30 × 30	51.8	67.5	56.7	57.9	58.7	27.6	51.4	42.8
100 × 100	43.6	61.6	44.7	46.6	53.4	19.9	43.1	36.1

Best values are in bold.

TABLE II

COMPARISON OF THE BEV MAP PREDICTIONS ACROSS DIFFERENT CLASSES AND THE MEAN FOR THE SIMPLEBEV MODEL TRAINED TO PREDICT TWO DIFFERENT COVERAGE AREAS ON THE NUSCENES DATASET

Cov. [m]	Mean	Driv.	Carp.	Ped. Cros.	Walkway	Stop Line	R. Div.	L. Div.
IoU								
30 × 30	50.7	88.4	45.3	55.4	56.2	26.2	46.0	37.6
100 × 100	44.6	86.3	38.1	46.5	49.8	21.1	37.7	31.8
Boundary IoU								
30 × 30	18.2	48.9	8.9	11.0	24.7	5.8	11.9	16.0
100 × 100	18.9	43.0	19.2	9.2	24.5	5.9	17.4	13.0

Best values are in bold.

We consider four experimental settings in the following: 1) varying the BEV Map Coverage, 2) Varying the GSD, 3) modifying the class configuration, and 4) combining real and synthetic data to study domain shift generalization.

V. RESULTS

We study four complementary aspects of dataset design for BEV mapping: the impact of spatial coverage, the role of GSD, the effect of class configuration, and the influence of domain shift when combining real and synthetic data. Each section introduces the corresponding experimental setting and presents the quantitative and qualitative results.

A. Effect of Spatial Coverage

We investigate whether reducing the BEV map coverage improves model performance. We compare two spatial extents: the standard [10], [11], [13] 100 × 100 m² footprint and a reduced 30 × 30 m² area. Given the 50 cm GSD of the dataset, these correspond to 200 × 200 and 60 × 60 pixels, respectively. All other factors are kept fixed.

To ensure a fair comparison between the two settings, the evaluation for the 100 × 100 m models is performed on a central cropped region of size 60 × 60 pixels, corresponding to the 30 × 30 m² area. This guarantees that both models are evaluated on the same spatial extent.

Tables I and II present the quantitative results for BEVCAR and SimpleBEV, respectively. For BEVCAR, the model trained on the smaller coverage achieves higher performance across all classes for both Mask IoU and Boundary IoU. The improvement is particularly pronounced for fine-grained structures, such as *Lane Dividers* and *Stop Lines*.

A similar trend is observed for SimpleBEV, where the 30 × 30 m² configuration also outperforms the larger coverage setting in terms of mean IoU and most classwise metrics. However, unlike BEVCAR, SimpleBEV does not benefit consistently from reduced coverage in terms of boundary IoU, where the 100 × 100 m setting outperforms the 30 × 30 m² setting for several fine-detail classes, such as *Carpark* (19.2 versus 8.9) and *Road Divider* (17.4 versus 11.9). This discrepancy can be attributed to the combination of two factors. First, these classes appear sparsely and are more prone to partial visibility within the smaller coverage area, which particularly affects boundary-level predictions. Second, BEVCAR's richer DINOv2-based feature extraction allows it to overcome this limitation and maintain precise boundary delineation even for sparse structures, whereas SimpleBEV's simpler pipeline remains more sensitive to the reduce spatial diversity introduced by the smaller coverage. While the magnitude of the improvement differs across classes and metrics, the overall behavior remains consistent between the two models.

The qualitative results in Figs. 5 and 6, corresponding to BEVCAR and SimpleBEV, respectively, further illustrate these differences. In particular, the smaller coverage setting produces more accurate predictions of road and walkway widths in regions farther from the ego-vehicle.

B. Effect of Ground Sampling Distance

We investigate the effect of the GSD on BEV map prediction. High-resolution maps are generally desirable for downstream tasks, such as localization and navigation [25], as they allow finer spatial detail to be represented. At the same time, the choice of GSD also affects the generated ground-truth maps. As illustrated in Fig. 7, maps of the same nuScenes scene generated with the official toolbox [3] at 50 cm/pixel and 10 cm/pixel exhibit visible differences, particularly for thin structures and class boundaries.

To evaluate the impact of different GSDs, we train each model twice on the same 30 × 30 m² area: once using a GSD of 50 cm/pixel, resulting in BEV maps of 60 × 60 pixels, and once using a GSD of 10 cm/pixel, resulting in BEV maps of 300 × 300 pixels. We assess performance using two evaluation protocols. First, the output of the 50 cm/pixel is upsampled and evaluated against 10 cm/pixel ground truth. Second, the output of the 10 cm/pixel is downsampled and evaluated against 50 cm/pixel ground truth. This allows both training resolutions to be compared under both target resolutions.

The quantitative results for BEVCAR and SimpleBEV are reported in Tables III and IV, respectively. For both models, the 10 cm/pixel training setting achieves the strongest results when evaluated against 10 cm/pixel ground truth, particularly in Boundary IoU. For BEVCAR, the mean Boundary IoU increases from 16.9 to 19.0, while for SimpleBEV it increases from 4.8 to 5.4. Similar improvements are observed for several thin or boundary-sensitive classes.

When evaluated against 50 cm/pixel ground truth, the opposite trend is observed. For both models, the 50 cm/pixel training setting achieves higher mean IoU and Boundary IoU than the

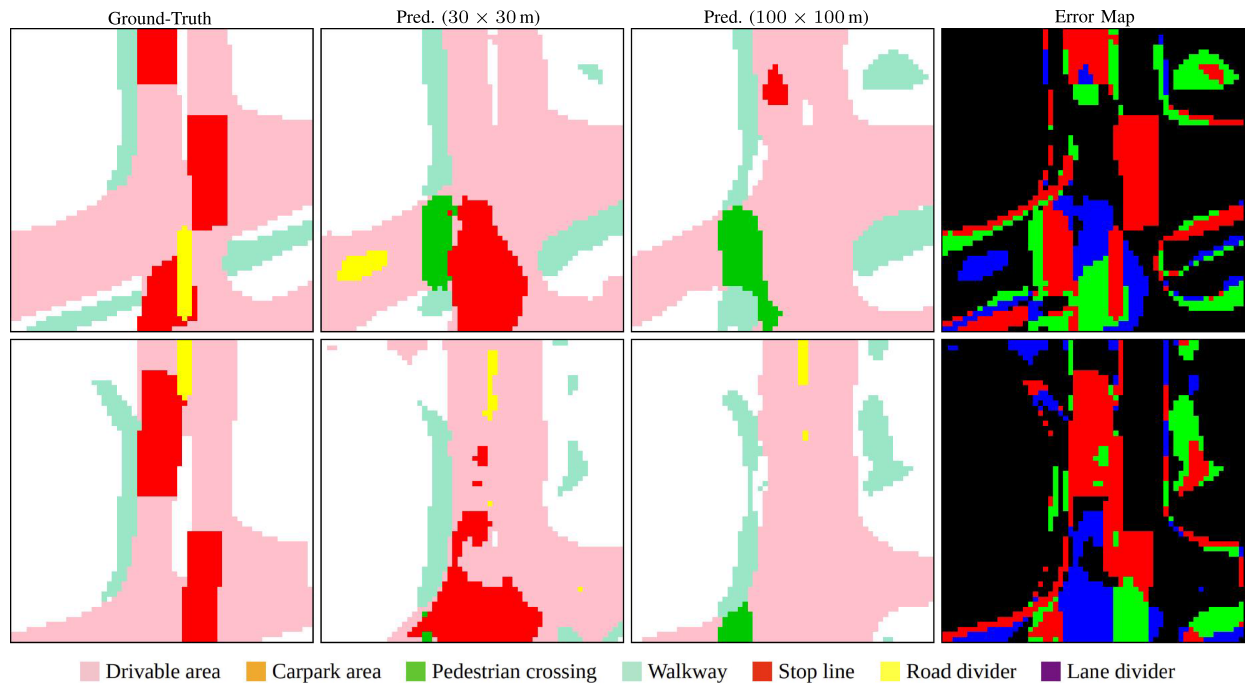


Fig. 5. BEVCar: Two examples comparing model predictions for different coverages ($30 \times 30 \text{ m}^2$ and $100 \times 100 \text{ m}^2$), showing the predicted BEV maps and corresponding error maps. In the error maps, ■ pixels misclassified by the 100 m model but correctly predicted by the 30 m model, ■ those misclassified by the 30 m model but correctly predicted by the 100 m model, ■ pixels misclassified by both models.

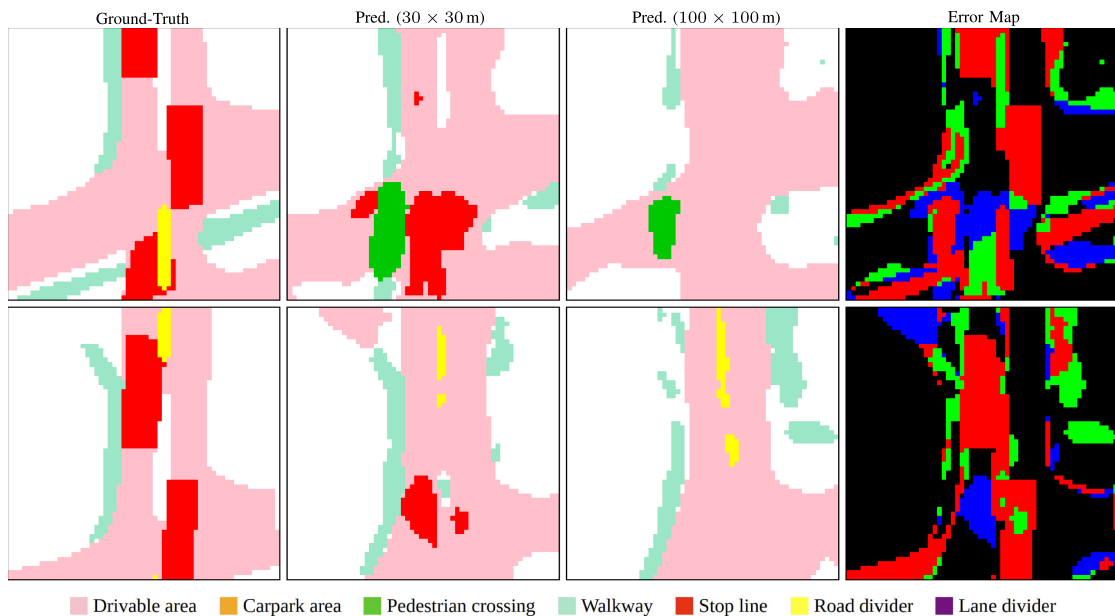


Fig. 6. SimpleBEV: Two examples comparing model predictions for different coverages ($30 \times 30 \text{ m}^2$ and $100 \times 100 \text{ m}^2$), showing the predicted BEV maps and corresponding error maps. In the error maps, ■ pixels misclassified by the 100 m model but correctly predicted by the 30 m model, ■ those misclassified by the 30 m model but correctly predicted by the 100 m model, ■ pixels misclassified by both models.

10 cm/pixel training setting. For BEVCar, the mean IoU increases from 51.5 to 56.1 and the mean Boundary IoU from 46.3 to 51.8. For SimpleBEV, the mean IoU increases from 47.0 to 50.7 and the mean Boundary IoU from 16.7 to 18.2.

A more detailed look at the class-wise results shows that coarse classes, such as *Drivable* and *Walkway* remain relatively stable across the different settings, while thinner structures, such as *Road Dividers*, *Lane Dividers*, and *Stop Lines* show larger

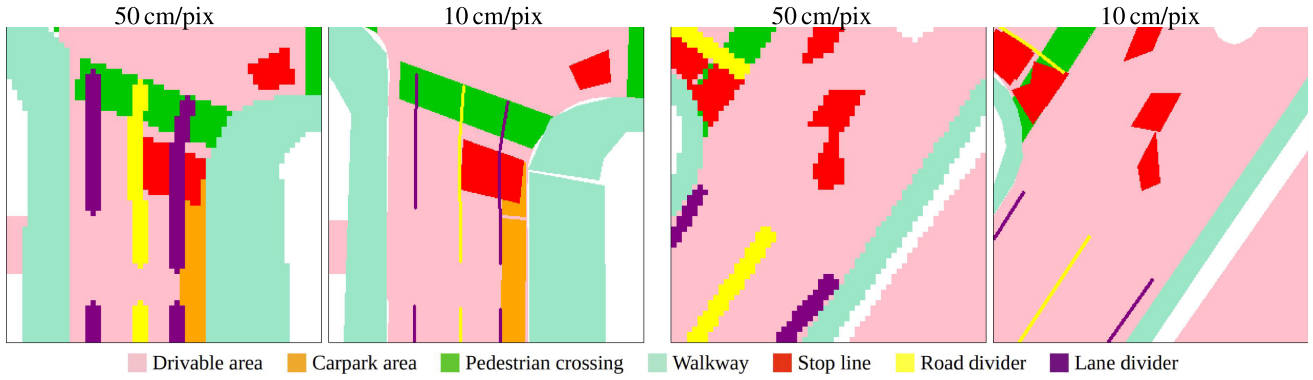


Fig. 7. Two sample BEV maps from the nuScenes dataset generated with different GSD: 50 cm/pixel and 10 cm/pixel.

TABLE III
BEVCAR: COMPARISON OF MODELS TRAINED ON BEV MAPS WITH GSDs OF 10 AND 50 CM/PIXEL UNDER TWO EVALUATION PROTOCOLS

Train GSD	Evaluated at 10 cm/pixel								Evaluated at 50 cm/pixel							
	Mean	Driv.	Carp.	Ped. Cros.	Walkway	Stop Line	R. Div.	L. Div.	Mean	Driv.	Carp.	Ped. Cros.	Walkway	Stop Line	R. Div.	L. Div.
IoU																
10 cm/pix	46.7	90.2	55.6	57.8	60.4	27.2	21.1	14.5	51.5	90.2	57.2	59.0	62.4	27.6	32.7	31.2
50 cm/pix	45.7	89.5	51.3	55.8	59.1	28.5	20.8	14.9	56.1	90.6	55.4	58.9	63.3	30.4	51.4	43.1
Boundary IoU																
10 cm/pix	19.0	27.9	20.4	18.8	22.7	7.8	21.0	14.4	46.3	66.8	52.9	57.6	57.4	25.8	32.7	31.2
50 cm/pix	16.9	23.0	18.7	15.7	19.6	7.7	20.1	13.5	51.8	67.5	56.7	57.9	58.7	27.6	51.4	42.8

In the first protocol, predictions are evaluated against 10 cm/pixel ground truth, with the output of the 50 cm/pixel model upsampled accordingly. In the second protocol, predictions are evaluated against 50 cm/pixel ground truth, with the output of the 10 cm/pixel model downsampled accordingly. Best results in each evaluation protocol are highlighted in bold.

TABLE IV
SIMPLEBEV: COMPARISON OF MODELS TRAINED ON BEV MAPS WITH GSDs OF 10 AND 50 CM/PIXEL UNDER TWO EVALUATION PROTOCOLS

Train GSD	Evaluated at 10 cm/pixel								Evaluated at 50 cm/pixel							
	Mean	Driv.	Carp.	Ped. Cros.	Walkway	Stop Line	R. Div.	L. Div.	Mean	Driv.	Carp.	Ped. Cros.	Walkway	Stop Line	R. Div.	L. Div.
IoU																
10 cm/pix	41.7	87.5	42.8	54.5	53.1	24.3	18.2	11.7	47.0	87.8	44.6	55.3	54.7	25.6	31.8	28.9
50 cm/pix	41.6	87.3	42.8	52.3	53.4	24.9	17.3	12.9	50.7	88.4	48.3	55.4	56.2	26.2	46.0	37.6
Boundary IoU																
10 cm/pix	5.4	15.7	2.2	1.9	8.8	1.6	4.6	3.0	16.7	47.5	6.0	7.9	24.1	6.1	12.5	13.0
50 cm/pix	4.8	15.4	2.9	2.7	7.9	1.6	1.8	1.8	18.2	48.9	8.9	11.0	24.7	5.8	11.9	16.0

In the first protocol, predictions are evaluated against 10 cm/pixel ground truth, with the output of the 50 cm/pixel model upsampled accordingly. In the second protocol, predictions are evaluated against 50 cm/pixel ground truth, with the output of the 10 cm/pixel model downsampled accordingly. Best results in each evaluation protocol are highlighted in bold.

variations. The qualitative results in Figs. 8 and 9 illustrate the same behavior. In particular, the 10 cm/pixel models produce finer and sharper boundaries, whereas the 50 cm/pixel models align more closely with the coarser 50 cm/pixel annotations. This is most evident for the *Road Divider* class, which at 10 cm/pixel appears as a very thin structure that both models frequently miss, while its artificially exaggerated width at 50 cm/pixel makes it considerably easier to detect.

C. Effect of Class Configuration

The nuScenes dataset includes seven map classes, most of which are commonly used in existing studies. However, closer inspection reveals that several class annotations are noisy or inconsistent. For example, in Fig. 1, the *Stop Line* class occupies a large portion of the crossroad in the BEV map, although the ground-view image only shows thin lines at the end of the turns behind the ego-vehicle. Similarly, *Road Divider* and *Lane Divider* annotations do not always reflect their visual appearance

and may appear interchangeable in some cases. In addition, inconsistencies are also visible for the *Carpark* class, where visually similar parking regions may be annotated differently within the same sample.

Since such label noise appears more prominent in fine-detail classes than in broader classes, such as *Drivable Area* and *Walkway*, we evaluate whether restricting training to the two main classes affects performance on those classes. To this end, we train models using either all classes or only the *Drivable* and *Walkway* classes. The comparison is performed on BEV maps covering a $30 \times 30 \text{ m}^2$ area at both 10 cm/pixel and 50 cm/pixel GSD.

The quantitative results for BEVCAR and SimpleBEV are shown in Tables V and VI, respectively. For BEVCAR, the differences between the all-class and the two-class settings are small at both resolutions. At 10 cm/pixel, the two-class model achieves a slightly higher mean IoU (75.4 versus 75.3) and *Drivable* IoU (90.4 versus 90.2), while *Walkway* remains unchanged at 60.4. In Boundary IoU, the all-class setting performs slightly better

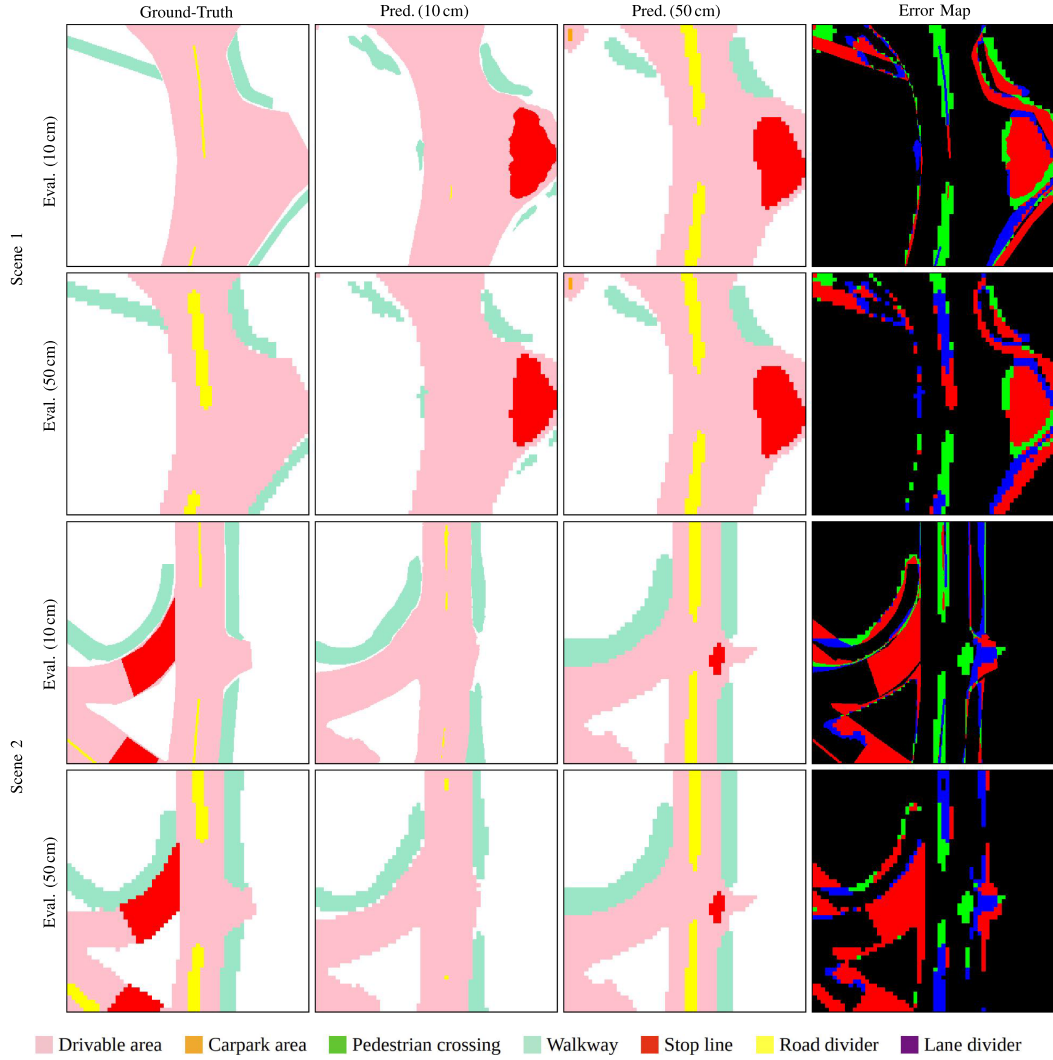


Fig. 8. BEVCAR: Two examples comparing training and evaluation at different GSDs (10 cm/pixel and 50 cm/pixel). For each sample, the ground truth at both GSDs is shown alongside: predictions from the model trained on 10 cm GSD (original and downsampled to 50 cm), predictions from the model trained on 50 cm GSD (original and upsampled to 10 cm), and the corresponding error maps. In the error maps, ■ pixels misclassified by model trained on 50 cm GSD and correctly predicted by model trained on 10 cm GSD, ■ pixels misclassified by model trained on 10 cm GSD and correctly predicted by model trained on 50 cm GSD, ■ pixels misclassified by both.

TABLE V

BEVCAR: COMPARISON OF THE PREDICTED BEV MAPS FOR *DRIVABLE* AND *WALKWAY* CLASSES WHEN TRAINING ON ALL CLASSES VERSUS ONLY THESE TWO CLASSES ON 10 CM/PIXEL AND 50 CM/PIXEL GSD

Trained on	10 cm/pix GSD			50 cm/pix GSD		
	Mean	Drivable	Walkway	Mean	Drivable	Walkway
IoU						
All Classes	75.3	90.2	60.4	76.9	90.6	63.3
Two Classes	75.4	90.4	60.4	76.9	90.7	63.0
Boundary IoU						
All Classes	25.3	27.9	22.7	63.1	67.5	58.7
Two Classes	25.1	27.6	22.5	63.0	67.8	58.3

Best results are highlighted in bold.

TABLE VI

SIMPLEBEV: COMPARISON OF THE PREDICTED BEV MAPS FOR *DRIVABLE* AND *WALKWAY* CLASSES WHEN TRAINING ON ALL CLASSES VERSUS ONLY THESE TWO CLASSES ON 10 CM/PIXEL AND 50 CM/PIXEL GSD

Trained on	10 cm/pix GSD			50 cm/pix GSD		
	Mean	Drivable	Walkway	Mean	Drivable	Walkway
IoU						
All Classes	70.3	87.5	53.1	72.3	88.4	56.2
Two Classes	71.1	88.3	53.8	73.3	88.8	57.6
Boundary IoU						
All Classes	12.3	15.7	8.8	36.8	48.9	24.7
Two Classes	12.2	16.5	7.8	36.9	49.6	24.2

Best results are highlighted in bold.

at 10 cm/pixel. At 50 cm/pixel, both settings again yield very similar results, with a small advantage for the two-class setting on the *Drivable* class and for the all-class setting on *Walkway*.

For SimpleBEV, the two-class settings yield slightly higher Mask IoU at both resolutions. At 10 cm/pixel, the IoU increases from 70.3 to 71.1, and at 50 cm/pixel from 72.3 to 73.3.

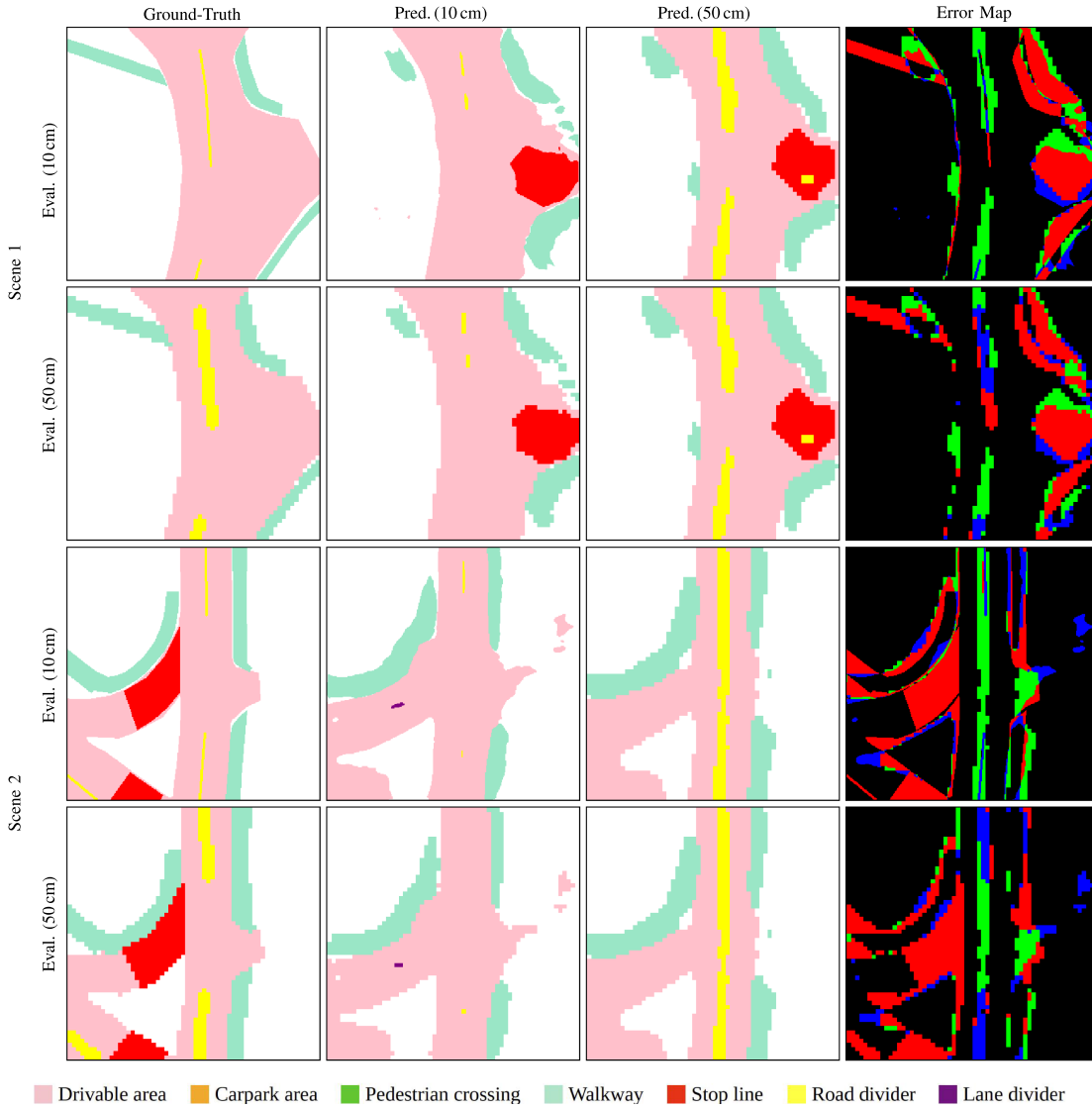


Fig. 9. SimpleBEV: Two examples comparing training and evaluation at different GSDs (10 cm/pixel and 50 cm/pixel). For each sample, the ground truth at both GSDs is shown alongside: predictions from the model trained on 10 cm GSD (original and downsampled to 50 cm), predictions from the model trained on 50 cm GSD (original and upsampled to 10 cm), and the corresponding error maps. In the error maps, ■ pixels misclassified by model trained on 50 cm GSD and correctly predicted by model trained on 10 cm GSD, ■ pixels misclassified by model trained on 10 cm GSD and correctly predicted by model trained on 50 cm GSD, ■ pixels misclassified by both.

Boundary IoU remains very similar between the two settings, with only small class-dependent differences.

Overall, both models show that restricting training to *Drivable* and *Walkway* does not lead to large changes in performance on these two classes. The observed differences are small compared to those seen in the coverage and GSD experiments.

D. Effect of Training Data Size and Domain Composition

This experiment addresses two complementary questions about training data composition. First, how does model performance scale with the amount of available real-world data, and how much of the nuScenes training set is actually needed to achieve strong results? Second, can synthetic data effectively complement or partially substitute for real data, improving

generalization without requiring additional real-world collection?

Models trained on nuScenes typically perform well on its own validation and test splits, but their behavior across domains is less clear, especially because few publicly available datasets provide a comparable camera setup and map annotation structure. To study this question, we use the synthetic dataset introduced in Section III-B, which was designed to be compatible with the nuScenes setting while taking into account the dataset-quality considerations discussed in the previous experiments.

Following the observations from Section V-C, we focus on the *Drivable* and *Walkway* classes. To address the first question, we train models on progressively smaller subsets of nuScenes only (100%, 75%, 50%, and 25%), which allows us to characterize

TABLE VII

BEVCAR: COMPARISON OF THE BEV MAP PREDICTIONS FOR *DRIVABLE* AND *WALKWAY* CLASSES WHEN TRAINED ON 10 CM/PIXEL BEV MAPS USING DIFFERENT COMBINATIONS OF NUSCENES (N) AND SYNTHETIC (S) DATASETS

Train [%]		Test on N			Test on S			Test on N+S		
N	S	Mean	Driv.	Walkway	Mean	Driv.	Walkway	Mean	Driv.	Walkway
IoU										
100	0	75.4	90.4	60.4	26.6	36.8	16.2	62.1	75.8	48.3
75	0	72.4	87.6	57.1	30.9	35.7	26.2	61.1	73.5	48.7
50	0	69.1	58.9	52.3	31.0	35.6	26.4	58.7	52.5	45.2
25	0	65.5	83.8	47.2	32.3	33.9	30.7	56.5	70.2	42.7
75	25	72.6	88.7	56.5	81.5	91.4	71.6	75.0	89.5	60.6
50	50	69.6	86.5	52.7	83.9	93.6	74.3	73.5	88.4	58.6
25	75	65.0	83.1	46.9	84.3	93.8	74.8	70.3	86.0	54.5
0	100	36.7	55.7	18.0	85.1	93.2	76.9	49.8	65.9	34.0
Boundary IoU										
100	0	25.1	27.6	22.5	5.5	5.9	5.2	19.8	21.7	17.8
75	0	22.1	24.3	20.0	6.2	5.3	7.2	17.8	19.1	16.5
50	0	20.3	21.9	18.7	6.2	5.4	7.1	16.5	17.4	15.5
25	0	18.1	19.3	16.9	5.7	4.7	6.8	14.7	15.3	14.2
75	25	22.1	24.2	19.9	27.8	33.6	22.0	23.7	26.8	20.5
50	50	20.3	21.9	18.8	32.7	40.1	25.3	23.7	26.9	20.6
25	75	17.2	18.8	15.6	34.9	42.6	27.3	22.0	25.3	18.8
0	100	5.3	5.0	5.6	35.7	42.3	29.0	13.6	15.2	12.0

Models are evaluated on the validation sets of nuScenes, the synthetic dataset, and their combination. Best results are highlighted in bold.

TABLE VIII

SIMPLEBEV: COMPARISON OF THE BEV MAP PREDICTIONS FOR *DRIVABLE* AND *WALKWAY* CLASSES WHEN TRAINED ON 10 CM/PIXEL BEV MAPS USING DIFFERENT COMBINATIONS OF NUSCENES (N) AND SYNTHETIC (S) DATASETS

Train [%]		Test on N			Test on S			Test on N+S		
N	S	Mean	Driv.	Walkway	Mean	Driv.	Walkway	Mean	Driv.	Walkway
IoU										
100	0	71.1	88.3	53.8	35.1	40.1	30.2	61.3	75.2	47.4
75	0	64.9	80.6	49.3	34.9	40.4	29.4	59.6	69.7	49.4
50	0	62.6	79.6	45.6	34.8	39.0	30.7	58.3	68.5	48.1
25	0	56.6	76.3	36.8	32.6	38.6	26.7	54.5	66.0	43.0
75	25	65.8	82.4	49.1	76.2	87.2	65.1	68.3	83.7	52.6
50	50	62.1	79.4	44.8	79.5	89.9	69.2	66.5	82.3	51.4
25	75	55.1	74.5	35.7	80.3	90.5	70.1	62.0	78.9	45.1
0	100	32.8	50.4	15.3	83.2	91.9	74.4	46.5	61.7	31.3
Boundary IoU										
100	0	12.2	16.5	7.8	5.4	4.9	5.9	8.6	13.3	4.0
75	0	9.9	13.0	6.8	5.5	4.9	6.1	7.6	10.8	4.5
50	0	8.9	11.3	6.5	5.2	4.3	6.0	6.9	9.5	4.8
25	0	7.6	8.1	7.1	5.4	4.3	6.4	6.0	7.1	4.9
75	25	9.0	12.2	5.7	22.2	24.7	19.7	11.9	15.6	8.1
50	50	9.2	12.9	5.5	26.8	32.2	21.3	12.5	18.1	6.8
25	75	7.1	9.0	5.2	28.3	34.4	22.2	12.0	16.0	7.8
0	100	5.1	5.2	5.1	34.2	40.4	28.0	10.4	14.8	6.1

Models are evaluated on the validation sets of nuScenes, the synthetic dataset, and their combination. Best results are highlighted in bold.

how performance degrades as real training data becomes scarcer. To address the second question, we then replace the withheld real data with varying proportions of synthetic data, training models on mixed real (N)–synthetic (S) combinations (N75%–S25%, N50%–S50%, N25%–S75%) as well as on synthetic data alone (S100%). This design allows us to directly assess the effect of adding synthetic data at each level of real-data availability, by comparing each mixed setting to its nuScenes-only counterpart trained on the same proportion of real data. For the mixed settings, all models are trained for the full 6000 iterations regardless of data composition, ensuring that each model is exposed to the same number of training steps. For the nuScenes-only subsets, the number of iterations is scaled proportionally to the data size. The full set of configurations is reported in Tables VII and VIII. All experiments are conducted using $30 \times 30 \text{ m}^2$ BEV maps at 10 cm/pixel GSD. The trained models are evaluated on nuScenes, the synthetic dataset, and their combination.

For both BEVCAR and SimpleBEV, the strongest performance on nuScenes is obtained when training only on nuScenes data. For BEVCAR, the N100% model achieves the highest mean IoU on nuScenes (75.4), while for SimpleBEV the corresponding

score is 71.1. Likewise, the strongest performance on the synthetic dataset is obtained when training only on synthetic data, where BEVCAR reaches a mean IoU of 85.1 and SimpleBEV 83.2. These results indicate a clear domain dependence, with each model performing best on the domain it was trained on.

The nuScenes-only rows also reveal how model performance scales with the amount of available real training data. For both models, reducing the nuScenes subset leads to a gradual but consistent drop in performance. For BEVCAR, the mean IoU on nuScenes decreases from 75.4 at N100% to 72.4, 69.1, and 65.5 at N75%, N50%, and N25%, respectively, with SimpleBEV following a similar trend (71.1, 64.9, 62.6, and 56.6). This confirms that real data remains valuable. Substituting the withheld real data with synthetic data has a largely neutral effect on nuScenes performance, with differences of less than one point in either direction across all settings and both models. However, it yields a clear benefit in a different dimension: substantially improved generalization across domains, as shown next.

Specifically, mixed training improves performance when evaluation is carried out on the combined test set. For BEVCAR, the best mean IoU on the combined set is obtained with the N75%–S25% setting (75.0), while the highest mean Boundary IoU is obtained with the N50%–S50% setting (23.7, rounded). For SimpleBEV, the best mean IoU on the combined set is also obtained with N75%–S25% training (68.3), while the best mean Boundary IoU is achieved by the N50%–S50% setting (12.5).

The tables also show that, for both models, increasing the proportion of synthetic data generally improves performance on the synthetic test set, while reducing performance on nuScenes. For example, in BEVCAR, the mean IoU on the synthetic test set increases from 26.6 for N100% to 85.1 for S100%, whereas the corresponding mean IoU on nuScenes decreases from 75.4 to 36.7. A similar trend is observed for SimpleBEV, where the mean IoU on the synthetic set rises from 35.1 to 83.2, while the nuScenes score drops from 71.1 to 32.8.

The qualitative examples in Fig. 10 are consistent with these quantitative results. Across the selected real and synthetic samples, models trained on mixed real and synthetic data generally produce more balanced predictions across both domains than models trained on a single domain only. At the same time, domain-specific failure cases remain visible, particularly in regions with occlusions or ambiguous class appearance.

VI. DISCUSSION

In this section, we interpret the experimental results presented in Section V and discuss their implications for dataset design in BEV mapping. We analyze how coverage, resolution, class configuration, and domain composition jointly affect model performance, boundary quality, and generalization.

A. Spatial Coverage and Hallucination Risk

The coverage experiment highlights a fundamental limitation of BEV mapping from camera-only inputs: models are required to predict regions that are not directly observable from the available views. When large spatial extents, such as 100×100



Fig. 10. BEVCar: Qualitative comparison across three samples from the nuScenes dataset and two samples from the synthetic dataset with 30×30 m coverage. On the left is the ground-view, followed by the ground-truth and predictions from models trained on different percentages of real and synthetic data. Drivable area in ■ and Walkway in ■. We also include quantitative measures for the different predictions including mean and class-wise (boundary) IoU.

m^2 are used, a significant portion of the ground truth lies outside the camera field of view. As a result, the model must infer these regions based on learned dataset priors, rather than direct visual evidence.

This behavior is reflected in the results, where reducing the coverage to $30 \times 30 \text{ m}^2$ consistently improves performance for both BEVCar and SimpleBEV. The improvement is particularly evident for fine-grained structures, which are more sensitive to incorrect spatial inference. These findings suggest that limiting the prediction area to regions that are more reliably observable can reduce hallucinations and lead to more accurate and interpretable predictions.

B. GSD and Annotation Consistency

The results of the GSD experiments reveal a strong dependency between model performance and the resolution at which both training and evaluation are performed. Models consistently perform best when the training and evaluation GSD match, indicating that the learned representations are closely tied to the scale of the annotation.

A key factor behind this behavior is the way ground truth annotations are generated. In the nuScenes dataset, thin

structures, such as lane markings, are represented with a fixed pixel width. As a result, their physical size varies significantly depending on the GSD. For example, a three-pixel-wide road divider corresponds to approximately 1.5 m at 50 cm/pixel, but only 30 cm at 10 cm/pixel. This inconsistency introduces a mismatch between the geometric reality and the annotation, which affects both training and evaluation.

This also explains the asymmetric behavior observed in the results. Upsampling coarse predictions can artificially improve overlap with fine-resolution ground truth, while downsampling fine-resolution predictions may cause thin structures to disappear. These findings indicate that improving spatial resolution alone is not sufficient; the consistency and accuracy of the annotation process across scales are equally important.

C. Class Configuration and Label Noise

The class configuration experiment shows that restricting training to the main classes (*Drivable* and *Walkway*) has only a limited impact on performance compared to training on all classes. This suggests that the prediction of these dominant classes is relatively robust to the inclusion or exclusion of finer, noisier classes.

However, the analysis of the dataset reveals that several fine-detail classes suffer from annotation inconsistencies, including ambiguous definitions and irregular labeling patterns. These inconsistencies introduce noise into the training process, which can affect the learning of precise boundaries and class relationships.

Interestingly, for BEVCar, training on all classes yields slightly better Boundary IoU at higher resolution, suggesting that fine-detail classes may still provide useful contextual information when accurate spatial detail is available. In contrast, SimpleBEV benefits slightly from the reduced class setting in terms of Mask IoU, indicating that simpler label spaces can also improve learning stability. Overall, these results suggest that the usefulness of additional classes depends on both annotation quality and model capacity.

D. Training Data Composition and Generalization

The generalization experiments demonstrate a clear domain gap between real and synthetic data. Models trained exclusively on one domain perform best on that domain but show a significant drop in performance when evaluated on the other. This indicates that current BEV mapping models do not generalize well across domains without additional adaptation.

At the same time, combining real and synthetic data during training leads to more balanced performance across domains. Mixed training improves results on the combined test set and consistently enhances boundary quality, as reflected in higher Boundary IoU scores. This suggests that synthetic data introduces useful variability that complements real-world data, particularly for learning sharper and more consistent boundaries.

However, the qualitative results also show that mixed training introduces new challenges, including occasional semantic confusion and errors in ambiguous regions. Furthermore, models trained solely on nuScenes appear to learn dataset-specific annotation biases, which can both help and hinder performance depending on the test domain.

Overall, these findings highlight the importance of dataset diversity and annotation quality for improving generalization in BEV mapping. While synthetic data can mitigate some limitations of real-world datasets, careful alignment between domains remains necessary to fully bridge the gap.

VII. LIMITATIONS AND FUTURE WORK

This study has limitations that also point to directions for future research. First, while the analysis is conducted using BEVCar and SimpleBEV, extending the evaluation to a broader range of architectures, including multimodal models, would help assess the generality of the observed dataset effects. Second, the results reveal limitations in current annotation practices, particularly for fine-detail classes and across different GSDs. Future work should therefore focus on developing datasets with more consistent and physically accurate annotations across scales. Third, although the synthetic dataset enables controlled analysis of domain effects, further validation on additional real-world datasets would strengthen conclusions regarding generalization. In particular, improving the alignment between real

and synthetic data distributions remains an important challenge. Finally, the strong impact of coverage and domain composition suggests that future work should explore more structured dataset design strategies. For example, restricting training to observable regions, prioritizing task-relevant classes, and incorporating targeted synthetic data could further improve the robustness of BEV mapping models. More broadly, to reliably advance BEV-based urban perception and remote sensing, future research should prioritize developing diverse, high-quality datasets with consistent and physically accurate annotations, enabling more resilient and generalizable data-driven remote-sensing models.

VIII. CONCLUSION

In this work, we conducted a dataset-centric analysis of the BEV map prediction task using the nuScenes dataset as a case study. We investigated how key dataset characteristics, including spatial coverage, ground sampling distance, class configuration, and training data composition, affect model performance and reliability. Importantly, the consistency of the observed trends across both BEVCar and SimpleBEV suggests that the impact of dataset characteristics is not limited to a specific architecture, but reflects more general properties of the BEV mapping task.

Our results show that reducing BEV map coverage leads to consistently improved performance by restricting predictions to regions that are both observable and closer to the ego-vehicle, where visual cues are more reliable, thereby reducing hallucinations and improving prediction accuracy. We therefore demonstrate that model performance is strongly influenced by the interaction between resolution and annotation design. In particular, inconsistencies in the representation of fine structures across different GSDs significantly affect boundary quality and evaluation outcomes.

In addition, we find that restricting training to dominant classes has only a limited impact on performance for those classes, suggesting that they are relatively robust to label noise in finer categories. Finally, our experiments highlight a clear domain gap between real and synthetic data, while also showing that mixed training improves performance across domains and leads to more balanced predictions.

Overall, our findings emphasize that dataset design, particularly coverage, annotation consistency, and domain diversity, play a central role in the performance and generalization of BEV mapping models. While some of these outcomes may seem intuitive, this work provides systematic, quantitative evidence that validates these assumptions and clarifies their practical impact, offering concrete guidance for the design and curation of future datasets in BEV mapping and beyond.

REFERENCES

- [1] M. Bojarski et al., "End to end learning for self-driving cars," 2016, *arXiv:1604.07316*.
- [2] Z. Liu et al., "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2023, pp. 2774–2781.
- [3] H. Caesar et al., "Nuscenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11618–11628.

- [4] A. E. B. Zekri, A. Latrach, R. Bahmanyar, and H. C.-Chouayakh, "Towards using synthetic data in aerial image segmentation," in *Proc. Joint Urban Remote Sens. Event*, 2025, pp. 1–4.
- [5] L. u. C. M. J. G. Van De Molengraft, and G. Dubbelman, "Monocular Semantic Occupancy Grid Mapping With Convolutional Variational Encoder–Decoder Networks," *IEEE Robot. Automat. Lett.*, vol. 4, no. 2, pp. 445–452, Apr. 2019, doi: [10.1109/LRA.2019.2891028](https://doi.org/10.1109/LRA.2019.2891028).
- [6] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern. Recognit.*, 2020, pp. 11138–11147.
- [7] N. Gosala, and A. Valada, "Bird's-Eye-View Panoptic Segmentation Using Monocular Frontal View Images," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 1968–1975, Apr. 2022, doi: [10.1109/LRA.2022.3142418](https://doi.org/10.1109/LRA.2022.3142418).
- [8] A. Saha, O. M. Maldonado, C. Russell, and R. Bowden, "Translating images into maps," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 9200–9206.
- [9] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robot. Automat. Lett.*, vol. 5, no. 3, pp. 4867–4873, Jul. 2020, doi: [10.1109/LRA.2020.3004325](https://doi.org/10.1109/LRA.2020.3004325).
- [10] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. Comput. Vis. - Europ. Conf. Comp. Vis.*, 2020, pp. 194–210.
- [11] X. Zhu, V. Zyrianov, Z. Liu, and S. Wang, "Mapprior: Bird's-eye view perception with generative models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8228–8239.
- [12] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, "Simple-BEV: What really matters for multi-sensor BEV perception?," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 2759–2765.
- [13] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13750–13759.
- [14] Z. Li et al., "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Computer Vision—ECCV*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Berlin, Germany: Springer, 2022, pp. 1–18.
- [15] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "BEVSegFormer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 5924–5932.
- [16] J. Schramm et al., "BEVCar: Camera-radar fusion for BEV map and object segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024, pp. 1435–1442.
- [17] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, 2024, pp. 1–32. [Online]. Available: <https://openreview.net/forum?id=a68SUt6zFt>
- [18] Z. Chen et al., "Vision Transformer adapter for dense predictions," in *Proc. Int. Conf. Learn. Representations*, 2023.
- [19] M. Cordts et al., "The Cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.
- [20] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, pp. 1231–1237, Sep. 2013.
- [21] M.-F. Chang et al., "Argoverse: 3D tracking and forecasting with rich maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8740–8749.
- [22] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Ann. Conf. Robot Learn.*, 2017, pp. 1–16.
- [23] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0377042787901257>
- [24] B. Cheng, R. Girshick, P. Dollar, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15329–15337.
- [25] J. Jeong, J. Yoon, H. Lee, H. Darweesh, and W. Sung, "Tutorial on high-definition map generation for automated driving in urban environments," *Sensors*, vol. 22, Sep. 2022, Art. no. 7056.