# BEVal: A Cross-dataset Evaluation Study of BEV Segmentation Models for Autononomous Driving

Manuel Alejandro Diaz-Zapata, Wenqian Liu, Robin Baruffa, Christian Laugier

# BEVal: A Cross-dataset Evaluation Study of BEV Segmentation Models for Autononomous Driving

Manuel Diaz-Zapata[1,2], Wenqian Liu[2], Robin Baruffa[2], Christian Laugier[2]

***Abstract*— Current research in semantic bird's-eye view segmentation for autonomous driving focuses solely on optimizing neural network models using a single dataset, typically nuScenes. This practice leads to the development of highly specialized models that may fail when faced with different environments or sensor setups, a problem known as domain shift. In this paper, we conduct a comprehensive cross-dataset evaluation of state-of-the-art BEV segmentation models to assess their performance across different training and testing datasets and setups, as well as different semantic categories. We investigate the influence of different sensors, such as cameras and LiDAR, on the models' ability to generalize to diverse conditions and scenarios. Additionally, we conduct multi-dataset training experiments that improve models' BEV segmentation performance compared to single-dataset training. Our work addresses the gap in evaluating BEV segmentation models under cross-dataset validation. And our findings underscore the importance of enhancing model generalizability and adaptability to ensure more robust and reliable BEV segmentation approaches for autonomous driving applications. The code for this paper available at https://github.com/manueldiaz96/beval/.**

## I. Introduction

Recently, the Bird's Eye View (BEV) representation has gained significant attention in the autonomous driving community as a crucial tool for scene understanding. Unlike traditional image or point cloud segmentation, the BEV encodes rich scene representations as a unified space for integrating information from multiple sensor modalities, offering advantages such as object size invariance and reduced occlusions [1]. Inspired by the Binary Occupancy Grids [2], recent methodologies aim to develop a BEV representation enriched with semantic information encoded in each cell.

Creating semantic BEV grid representations presents a unique challenge: generating a top-down view of the scene that differs from the perspectives offered by the vehicle's sensors. Some cutting-edge approaches [4], [5], [6] utilize camera features and geometry to construct the BEV representation, while others [7] leverage 3D point cloud data to extract relevant semantic information. More recently, there has been an increase in sensor fusion techniques [8], [3], [9] that combine features from various sensor types to enhance the quality of BEV representations. This top-down perspective is particularly useful for downstream tasks such as tracking [10] and planning [4]. Semantic grids enable systems to distinguish between different object types (e.g., vehicles, pedestrians, static obstacles) and scene areas (e.g., roads, walkways, pedestrian crossings), facilitating better decision-making processes.

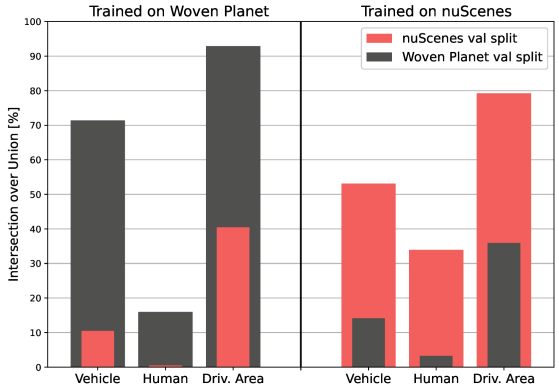



Fig. 1: Cross-dataset validation using the BEV semantic segmentation model LAPT-PP [3]. The left three columns show the Intersection Over Union (IoU) scores for three semantic categories when the model was trained on the Woven Planet dataset and tested on the nuScenes validation set (pink) and the Woven Planet validation set (gray). The right three columns show similar results when the model was trained on the nuScenes dataset initially. A significant performance drop is observed when the model is trained and tested on different datasets, highlighting its inherent limitations in generalization ability.

These advanced models require large volumes of diverse and accurately annotated data to learn the intricate details in the semantic BEV grids. Current research on BEV segmentation [11], [3], [12], [6] predominantly employs the nuScenes dataset for both training and evaluation. This raises critical questions about the robustness and generalizability of these models, as they are typically tested on a single dataset. While domain adaptation techniques are powerful and often used to improve model generalization, they often introduce additional complexity and computational overhead. Cross-dataset evaluation, on the other hand, provides a more direct and empirical verification of model robustness across different real-world conditions and scenarios without the need for additional training or fine-tuning.

Moreover, cross-dataset evaluation contributes to establishing standardized benchmarks, reveals inherent model lim-

This work was partially supported by Toyota Motor Europe.

[1]CHROMA team, Univ Lyon, Inria, INSA Lyon, CITI Lab, France. [2]CHROMA team, Univ. Grenoble Alpes, Inria, Grenoble, France. Correspondence: manuel.diaz-zapata@inria.fr

itations and strengths, offering clear insights into how it will perform when deployed in varied real-world settings. Despite the importance and advantages of cross-dataset validation, which ensures that models generalize well beyond their training data and mitigates overfitting, this area remains underexplored in the BEV semantic grid segmentation literature.

In this work, we aim to address this gap in evaluating BEV segmentation models across multiple datasets to verify their reliability and applicability in diverse real-world scenarios. We propose a novel cross-dataset framework for training and evaluating three BEV segmentation models on the nuScenes [13] and Woven Planet datasets [14]. We conduct experiments on three state-of-the-art BEV semantic segmentation models, evaluating their performance across three semantic categories using the Intersection Over Union (IoU) score. As shown in Fig. 1, there is a significant performance drop when models are tested on unseen data from a different dataset. Our proposed cross-dataset validation framework aims to identify specific weaknesses or failure modes that may not be apparent within a single dataset, helping to develop more robust and reliable models for autonomous driving. To our knowledge, this is the first work to address such topic.

Our contributions are:

- We introduce the first cross-dataset validation framework for BEV semantic segmentation task. This framework is flexible, which can be extended to additional models, datasets and semantic categories.
- We perform a comparative study using two real-world large-scale datasets, assessing three BEV segmentation models with a variety of input sensor modalities, across three semantic segmentation categories.
- Additionally, we investigate the models' generalization ability by training them simultaneously on both datasets.

## II. Related work

### A. Semantic BEV Segmentation

Recently, innovative approaches have significantly advanced BEV semantic segmentation for autonomous driving. For instance, PillarSegNet [7] utilized LiDAR point clouds to generate 2D BEV feature maps for predicting semantic classes in a BEV grid. LSS [4] and FIERY [15] addressed the challenge of image plane to BEV projection by learning probability distributions for discrete depth values. PON [12] proposed encoding images into 1D feature tensors, subsequently sampled to generate the BEV. More recent works [6], [16], [17] have employed attention operations to enhance BEV segmentation using RGB images.

Recognizing the advantages of both camera and LiDAR, researchers have developed sensor fusion approaches to address the weaknesses of one sensor with the strengths of another. For example, LAPTNet [3] resolved the challenge of projecting camera features to BEV by utilizing LiDAR depth information across multiple image scales. TransFuseGrid [9] fused camera and LiDAR feature maps using attention operations, while SimpleBEV [8] proposed enriching camera features with LiDAR or Radar information for semantic grid prediction.

Given the diverse approaches in state-of-the-art BEV segmentation, most of them are primarily trained and evaluated using the nuScenes dataset. However, this single-dataset approach raises concerns about the robustness and generalizability of these models. To address this, we are introducing cross-dataset validation for BEV segmentation models, with the goal of developing more robust and reliable models.

### B. Cross-dataset Validation

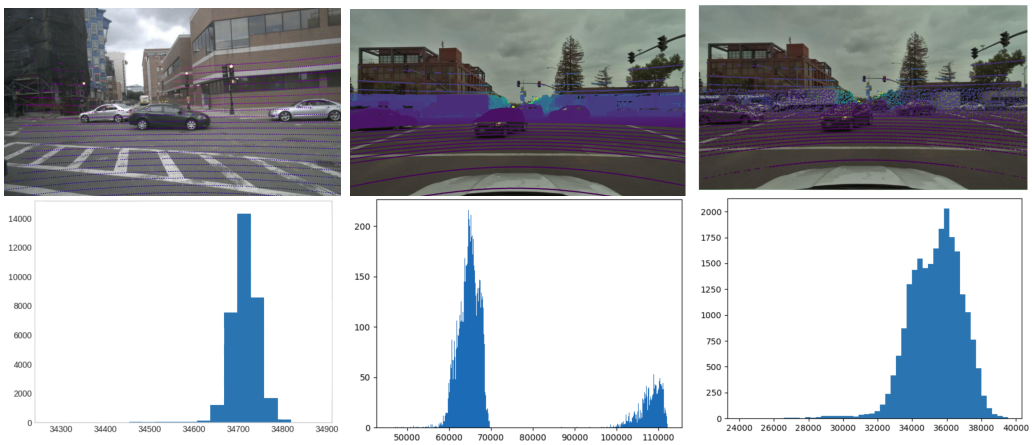
There has been a growing interest in cross-dataset validation research in autonomous driving. Gilles *et al*. [18] evaluated vehicle trajectory prediction methods across four datasets, highlighting that the size of the dataset is not the most contributing factor in increasing performance, but rather its ability to faithfully represent real conditions. Gesnouin *et al*. [19], studied pedestrian intention prediction across three datasets, finding that models often overfit to one dataset and underperform on others, highlighting the need for quantifying a model's uncertainty when evaluating on unseen data. Stäcker *et al*. [20] evaluated a 3D detection network across two datasets using a camera and radar fusion approach, demonstrating that visual variability in pretraining benefits camera features but not radar features, while the fusion of both modalities leads to the best performance overall. Furthermore, they only evaluated their models on the same dataset used for training. Despite the exploration of cross-dataset validation in autonomous driving tasks, there is a notable lack of research in BEV semantic grid segmentation, a gap that our work aims to fill.

## III. Methodology

In this section, we outline the methodology of our study. First, we discuss the datasets utilized, highlighting their characteristics and differences. Next, we detail the processing of sensor data common to both datasets and the generation of ground truth. Finally, we describe the models, as well as the training and evaluation strategies employed in our study.

### A. Datasets

To conduct cross-dataset evaluation study in this paper, we use nuScenes dataset [13] and Woven Planet Perception Dataset [14], given their relevance in the BEV segmentation literature [4], [8], [15] and their comparable sensor setups. The **nuScenes dataset** [13] focuses on driving-specific scenarios and was collected in Boston (USA) and Singapore. It provides sensor information from six cameras, five radars and one 32-layer LiDAR across 1000 driving scenes of 20 seconds each. Given the different sampling rates of each sensor, the dataset provides a set of synchronized keyframes across all sensors with a frequency of 2Hz. It also provides 3D bounding box annotations for the different agents in each scene and a set of high-definition maps of the traversed areas. The **Woven Planet Perception Dataset** [14], formerly known as the Lyft Level 5 Perception Dataset, is a large-scale dataset for research on self-driving vehicles. Captured across




(a) nuScenes (b) Woven Planet (c) Woven Planet (subsampled)

Fig. 2: Point cloud sample illustration (top) and histogram of the number of points per sample (bottom) for (a) nuScenes, (b) Woven Planet and (c) the subsampled Woven Planet point clouds. Best viewed with digital zoom.

the city of Palo Alto (USA), it provides sensor data from a set of six cameras and three 64-layer LiDARs, 3D bounding box annotations for pedestrians and vehicles in the scene, as well as a semantic map raster at a resolution of 10cm/px.

Both of the two datasets provide a comprehensive 360° field of view, including six surrounding cameras and a roof-mounted LiDAR. To maintain consistency in sensor configurations between both datasets, we omitted the use of radar sensors from the nuScenes dataset and the point clouds from the frontal LiDARs from the Woven Planet dataset.

### B. Point Cloud Processing

Given the difference in the LiDAR specifications between nuScenes (32 layers) and Woven Planet (64 layers), we conducted a preliminary study to compare the distribution of point clouds across both datasets. We generated histograms depicting the number of points per sample, as shown in Fig. 2a for nuScenes and Fig. 2b for Woven Planet. A considerable difference in point cloud density was observed between the two datasets, which is expected due to their different LiDAR systems. Notably, nuScenes' point clouds exhibit greater uniformity across samples, with both the median and average number of points being 34,720, whereas Woven Planet's point clouds have an average of 72,431 points and a median of 65,568 points.

To achieve a uniform number of points per sample across both datasets, we subsampled the point clouds in the Woven Planet dataset to match, as closely as possible, the density of those in the nuScenes dataset. We first transformed each point cloud available in the Woven Planet from the original Cartesian coordinates $(x, y, z)$ to spherical coordinates $(\rho, \theta, \phi)$. Then we divided the range of $\theta$ values into 32 sectors, corresponding to the 32 LiDAR layers in nuScenes. For the $\phi$ values, we divided them into 1500 sectors, as this produced a distribution most similar to nuScenes. We sampled one point from each sector and saved the resulting point cloud for later use. The difference between the original and subsampled point clouds is shown in the top of Fig. 2b and Fig. 2c respectively. The histogram at the bottom of Fig. 2c illustrates that the subsampled point cloud distribution in Woven Planet is finally closer to the nuScenes distribution, with a median of 35,498 points and a mean of 35,360 points.

### C. Image Processing

The nuScenes and the Woven Planet datasets present different image sizes for their camera input. nuScenes provides a set of six camera images of $(1600 \times 900)$ pixels. In contrast, Woven Planet includes two different image sizes across scenes: some are $(1920 \times 1080)$ pixels with a $16:9$ aspect ratio (same as nuScenes), and other scenes are $(1124 \times 1024)$ pixels with a $1:1.1$ aspect ratio.

To ensure consistency across both datasets, we followed previous works [4], [3], resizing and center cropping each image to have dimensions of $(128 \times 352)$ pixels. We also adjusted the intrinsic camera matrices accordingly. Additionally, we applied standard ImageNet [21] normalization before passing the images to the models for evaluation.

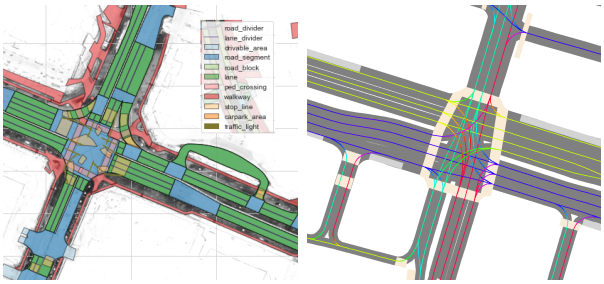
### D. Ground Truth Generation

We evaluated BEV segmentation within a 100m by 100m area surrounding the ego vehicle. Consistent with previous studies [4], [15], [8], [3], we discretized this space at a resolution of 0.5m per pixel, resulting in a 200 by 200 pixel grid. We used three semantic categories in our experiments, including the Human, Vehicle and the Drivable Area, given that these are the only semantic categories in common across both datasets.

To obtain the required ground truth for Human and Vehicle classes, we discretized the provided 3D bounding box coordinates and sizes and projected them onto the BEV to generate the corresponding semantic ground truth. We did not filter these annotations based on visibility levels, as suggested in [16] and [22], since these visibility levels are only available in the nuScenes dataset.

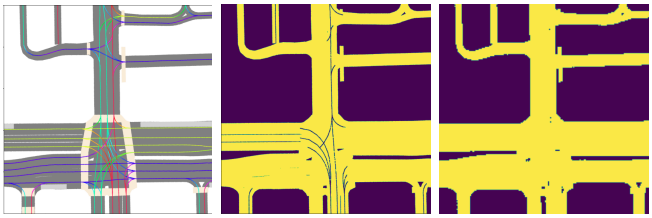
To discuss the ground truth generation for the Drivable Area class, we first show an example of the original map annotations provided by each dataset in Fig. 3. It is straightforward for the nuScenes dataset using its provided map API.

By providing the ego position, area of interest, resolution, and required class, we can generate any ground truth map representation for any sample.




(a) nuScenes (b) Woven Planet

Fig. 3: Example of map annotations provided by (a)nuScenes dataset and (b) Woven Planet dataset.



(a) (b) (c)

Fig. 4: Drivable Area ground truth generation for the Woven Planet Dataset. (a) Region of interest cropping. (b) Color filtering. (c) Gap filling and image resizing. Best viewed with digital zoom.

For the Woven Planet dataset, which provides its map as one RGB image, the procedure for ground truth generation differs. First, we crop the area of interest from the original map image. Next, we apply a color filter to isolate pixels representing drivable areas and crossings, followed by a morphological closing operation using a $(5 \times 5)$ kernel to fill gaps left by the centerlines. Finally, we resize the image to match the required BEV resolution. This process is illustrated in Fig. 4.

### *E. Models*

We perform cross-dataset validation experiments using state-of-the-art BEV semantic segmentation models with various input sensor modalities:

**Camera-only**: Lift-Splat-Shoot (LSS) [4] is a prominent semantic BEV segmentation model that exclusively uses camera images as input. It predicts an implicit depth distribution to project image features into 3D space and assigns these features to BEV cells via sum-pooling. This model serves as a benchmark to assess performance variations across datasets when using only image inputs.

**Early Camera-LiDAR Sensor Fusion**: LAPT [3] adopts an early fusion approach by combining camera and LiDAR data at the initial stage. This model utilizes LiDAR depth information to link image features with the BEV, projecting features from multiple image scales to enhance BEV coverage. It illustrates the impact of limited sensor fusion, focusing on depth values from point clouds rather than their complete 3D structure.

**Late Camera-LiDAR Sensor Fusion**: LAPT-PP [3] is a variant of LAPT that employs late fusion techniques. This model integrates a LiDAR-specific encoder to generate BEV features exclusively from point-cloud data. The resulting feature map is then fused with camera-derived BEV features to predict final semantic segmentation. This model evaluates performance changes in networks relying on the 3D structure of point clouds, showcasing the effects of late-stage sensor fusion.

### *F. Experimental Details*

We evaluated each of the models described in Section III-E for single-class BEV segmentation on Vehicle, Human, and Drivable Area respectively. Additionally, we also evaluated each model on multi-class BEV segmentation by jointly predicting Vehicle and Drivable Area, to assess their performance in a broader context.

For both single-class and multi-class segmentation tasks, each model underwent two distinct training setups. Firstly, models were individually trained on the nuScenes dataset and the Woven Planet dataset separately, then evaluated on both datasets to measure cross-dataset generalization. Secondly, models were trained simultaneously on both datasets to investigate the impact of augmented training data on performance across individual datasets.

For training, we utilized binary cross-entropy loss for single-class prediction, and cross-entropy loss for multi-class prediction. Training employed a batch size of 10 and the Adam optimizer with a learning rate of $0.001$, continuing for 50 epochs or until convergence of the evaluation metric, whichever happens first.

During evaluation, we adhered to standard practices in the field, using the Intersection over Union (IoU) metric to gauge the overlap between model predictions and ground truth annotations.

## IV. Results and Discussion

In this section, we first present the experimental results for cross-dataset evaluation for both single-class and multi-class BEV segmentation. Then we discuss the experiments on multi-dataset training setups. We will show results quantitatively and qualitatively.

### *A. Cross-dataset Evaluation*

We present quantitative results for cross-dataset evaluation in Table I. For single-class prediction results shown in Table I(a), I(b) and I(c), models trained on Woven Planet mostly exhibit worse generalization to nuScenes across different semantic classes compared to the reverse case, as indicated by the significant drop in IoU scores.

Among the models, we observe that LAPT-PP suffers the most than the other two models in cross-dataset evaluation, showing the largest performance drops across different semantic classes and respective train-test setups. This significant degradation can be attributed to LAPT-PP-FPN's heavy

| | NS | NS* | WP | WP* |
|---|---|---|---|---|
| **LSS** | 32.95 | 10.5 (68.13% ↓) | 27.07 | 22.63 (16.4% ↓) |
| **LAPT** | 47.03 | 22.3 (52.58% ↓) | 57.98 | 37.18 (35.87% ↓) |
| **LAPT-PP** | 53.1 | 10.46 (80.3% ↓) | 71.37 | 14.1 (80.24% ↓) |

(a) Vehicle

| | NS | NS* | WP | WP* |
|---|---|---|---|---|
| **LSS** | 12.21 | 0.4 (96.72% ↓) | 5.55 | 4.1 (26.13% ↓) |
| **LAPT** | 22.69 | 2.8 (87.66% ↓) | 10.35 | 7.73 (25.31% ↓) |
| **LAPT-PP** | 33.95 | 0.5 (98.53% ↓) | 15.95 | 3.25 (79.62% ↓) |

(b) Human

| | NS | NS* | WP | WP* |
|---|---|---|---|---|
| **LSS** | 75.41 | 33.5 (55.58% ↓) | 84.88 | 48.55 (42.86% ↓) |
| **LAPT** | 77.15 | 47.6 (38.3% ↓) | 90.71 | 58.57 (35.43% ↓) |
| **LAPT-PP** | 79.23 | 40.42 (48.98% ↓) | 92.88 | 35.9 (61.34% ↓) |

(c) Drivable Area

| | NS | NS* | WP | WP* |
|---|---|---|---|---|
| **LSS** | 19.4 / 62.03 | 12.77 (34.18% ↓) / 32.75 (47.2% ↓) | 27.82 / 73.33 | 16.73 (39.86% ↓) / 44.52 (39.29% ↓) |
| **LAPT** | 15.5 / 60.0 | 19.13 (31.93% ↑) / 36.11 (39.82% ↓) | 23.62 / 68.07 | 16.05 (32.05% ↓) / 49.28 (27.60% ↓) |
| **LAPT-PP** | 20.88 / 58.76 | 15.51 (24.72% ↓) / 37.04 (36.96% ↓) | 43.3 / 68.33 | 9.08 (79.03% ↓) / 34.04 (50.18% ↓) |

(d) Vehicle / Drivable Area

TABLE I: IoU [%] scores for cross-dataset evaluation of single-class BEV segmentation: (a) Vehicle, (b) Human, (c) Drivable Area, and multi-class BEV segmentation with joint prediction: (d) Vehicle / Drivable Area. The columns with a asterisk sign ($*$) indicate models trained on one dataset and tested on the other separately. The values in gray beside the IoU scores denote the performance difference in percentage when models were trained on a different dataset compared to their baselines (trained and tested on the same dataset). And the arrows indicate whether the performance drops (↓) or increases (↑). (**NS** - nuScenes / **WP** - Woven Planet)

reliance on LiDAR point cloud features. Models based on LiDAR often struggle with cross-dataset generalization due to sensor-specific variations, particularly the differences in LiDAR configurations between Woven Planet and nuScenes datasets. In contrast, image-based models, such as LSS and LAPT, benefit from more consistent visual data, standardized preprocessing, and annotation practices, resulting in better cross-dataset performance.

We also present multi-class segmentation results in Table I(d) by training and testing models to jointly predict Vehicle and Drivable Area in the BEV grid. We notice that multi-class prediction demonstrates less performance drop compared to single-class across datasets and models. This finding suggests that jointly predicting multiple classes (vehicle and drivable area) helps mitigate the impact of dataset variations on model performance. By jointly learning multi-class prediction, the model captures robust and redundant features from the scene that can help stabilize predictions against dataset-specific biases.

Last but not least, we realize that the human segmentation performance yields the lowest absolute IoU scores across datasets and models. On one hand, human BEV segmentation is challenging given the limited number of samples available within datasets. On the other hand, the chosen BEV grid resolution of 0.5m/px results in each human annotation being represented by only one to two pixels. This fine granularity can lead to the development of highly specialized models when trained on individual datasets.

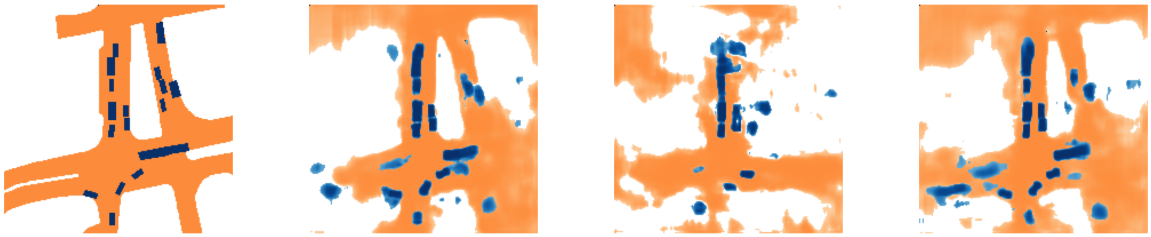
### *B. Multi-dataset Training*

Next, we conducted experiments using the combined training sets from both the Woven Planet and nuScenes datasets. We then tested each model on each dataset separately and present the IoU scores in Table II. The performances for both single-class and multi-class segmentation are shown. After training on both datasets, the models demonstrated consistent accuracy across both datasets. Specifically, the IoU scores for each dataset were similar to their baselines shown in Table I. Additionally, the IoU scores across the two datasets exhibited balanced performance, without bias towards one dataset or the other. Notably, as nuScenes data is included in the training process, all models for Human segmentation improve by 10% to 15% on the Woven Planet. This improvement can be attributed to the more varied data in nuScenes, as noted by Gilles et al. [18], which helps the model better understand the task.

While both LSS and LAPT achieve improved performance on both datasets, LAPT-PP suffers a slight performance degradation, which is within our expectation. This is likely due to the domain shift between the two datasets, which have different data distributions (e.g., different sensor configurations, environmental conditions, annotation styles).

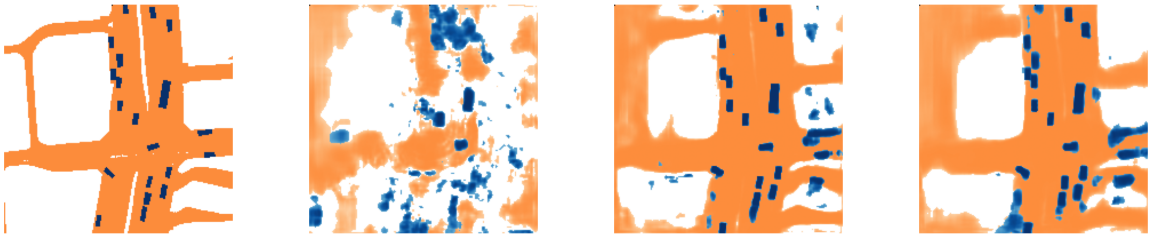
### *C. Qualitative Results*

In Fig. 5 and Fig. 6, we show qualitative results yielded by the LAPT-PP model on nuScenes and Woven Planet datasets respectively. We jointly predict Vehicle and Drivable Area in the scene onto the same BEV grid. Following the discussion above, the model performs the best on the dataset it was trained on. Fig. 5(b) and 6(c) show two examples when LAPT-PP was trained and tested using the same dataset, and the resultant BEV grids closely match the ground truth. However, when evaluating on a different dataset (Fig. 5(c) and Fig. 5(b)), the model's predictions fail to accurately



(a) Ground Truth (b) Trained on Nuscenes (c) Trained on Woven Planet (d) Trained on both datasets

Fig. 5: Qualitative BEV semantic segmentation results for LAPT-PP on nuScenes dataset.



(a) Ground Truth (b) Trained on Nuscenes (c) Trained on Woven Planet (d) Trained on both datasets

Fig. 6: Qualitative BEV semantic segmentation results for LAPT-PP on Woven Planet dataset.

| | Woven Planet | | | | nuScenes | | | |
|---|---|---|---|---|---|---|---|---|
| | **Vehicle** | **Human** | **Driv. A.** | **Vehicle / Driv. A.** | **Vehicle** | **Human** | **Driv. A.** | **Vehicle / Driv. A.** |
| **LSS** | 28.58 (↑) | 6.11 (↑) | 83.19 (↓) | 25.95 (↓) / 70.22 (↓) | 33.47 (↑) | 12.35 (↑) | 76.63 (↑) | 24.74 (↑) / 65.52 (↑) |
| **LAPT** | 58.32 (↑) | 11.96 (↑) | 87.65 (↓) | 20.84 (↓) / 63.66 (↓) | 48.55 (↑) | 21.76 (↓) | 78.16 (↑) | 13.76 (↓) / 51.57 (↓) |
| **LAPT-PP** | 70.57 (↓) | 17.31 (↑) | 90.57 (↓) | 29.75 (↓) / 66.03 (↓) | 52.48 (↓) | 33.76 (↓) | 79.77 (↑) | 16.72 (↓) / 52.68 (↓) |

TABLE II: Multi-dataset training for BEV semantic segmentation tasks. The models are trained on the combined training sets from both the Woven Planet and nuScenes datasets. They are then tested on each dataset separately. The IoU [%] scores are calculated for single-class semantic segmentation: Vehicle, Human, Drivable Area, and multi-class prediction: Vehicle/Drivable Area. Additionally, the arrows in gray color indicate whether the performance drop (↓) or increase (↑) compared to the baseline scores shown in Table I.

represent the scene semantics. When the model is trained on both datasets, as shown in Fig. 5(d) and Fig. 6(d), it is able to represent most of the scene accurately but does not achieve the same level of performance as single-dataset training. For more qualitative comparisons of the models across all datasets and classes, please refer to the following video: https://youtu.be/z9-wJ-FTc8Y

## V. Conclusions

In this paper, we addressed the critical gap in cross-dataset evaluation research for BEV semantic grid segmentation tasks. We evaluated three BEV segmentation models across three semantic categories using two autonomous driving datasets. Additionally, we proposed a preprocessing procedure to standardize the setups of the two datasets, ensuring similar data distributions and groundtruth annotations. Our results indicate that models utilizing images as the primary feature source demonstrate superior generalization across datasets, whereas those relying on LiDAR point clouds are more sensitive to dataset-specific characteristics. Furthermore, our study suggests that multi-dataset training achieves performance comparable to single-dataset training, albeit with potential slight performance drops due to domain shift. These findings underscore the importance of diverse data exposure in developing robust and reliable autonomous driving systems.

In future work, we plan to explore additional data augmentation methods, such as experimenting with different ratios of dataset splits, and investigate techniques in domain adaptation. These steps aim to identify specific factors contributing to performance drops and develop strategies to mitigate them, ultimately leading to models capable of generalizing to broader conditions and scenarios.

## Acknowledgments

The experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations.

## REFERENCES


[1] B. Siciliano, O. Khatib, and T. Kröger, *Springer handbook of robotics*. Springer, 2008, vol. 200.

[2] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, 1989.

[3] M. Diaz-Zapata, D. Sierra-Gonzalez, Ö. Erkent, C. Laugier, and J. Dibangoye, "Laptnet-fpn: Multi-scale lidar-aided projective transform network for real time semantic grid prediction," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 712–718.

[4] 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*. Springer, 2020, pp. 194–210.

[5] N. Gosala and A. Valada, "Bird's-eye-view panoptic segmentation using monocular frontal view images," *IEEE Robotics and Automation Letters*, 2022.

[6] A. Saha, O. Mendez Maldonado, C. Russell, and R. Bowden, "Translating images into maps," *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[7] J. Fei, K. Peng, P. Heidenreich, F. Bieder, and C. Stiller, "Pillarsegnet: Pillar-based semantic grid map estimation using sparse lidar data," in *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, pp. 838–844.

[8] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, "A simple baseline for bev perception without lidar," *arXiv preprint arXiv:2206.07959*, 2022.

[9] G. Salazar-Gomez, D. S. González, M. A. Diaz-Zapata, A. Paigwar, W. Liu, Ö. Erkent, and C. Laugier, "Transfusegrid: Transformer-based lidar-rgb fusion for semantic grid prediction," in *ICARCV 2022-17th International Conference on Control, Automation, Robotics and Vision*, 2022.

[10] J. Dequaire, P. Ondrúška, D. Rao, D. Wang, and I. Posner, "Deep tracking in the wild: End-to-end tracking using recurrent neural networks," *The International Journal of Robotics Research*, p. 0278364917710543, 2017.

[11] S. Srivastava, F. Jurie, and G. Sharma, "Learning 2d to 3d lifting for object detection in 3d for autonomous vehicles," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4504–4511.

[12] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 138–11 147.

[13] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[14] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet, "Woven planet perception dataset 2020," https://woven.toyota/en/perception-dataset, 2019.

[15] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "Fiery: Future instance prediction in bird's-eye view from surround monocular cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 273–15 282.

[16] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 760–13 769.

[17] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," *arXiv preprint arXiv:2203.17270*, 2022.

[18] T. Gilles, S. Sabatini, D. Tsishkou, B. Stanciulescu, and F. Moutarde, "Uncertainty estimation for cross-dataset performance in trajectory prediction," *CoRR*, vol. abs/2205.07310, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2205.07310

[19] J. Gesnouin, S. Pechberti, B. Stanciulescu, and F. Moutarde, "Assessing cross-dataset generalization of pedestrian crossing predictors," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 419–426.

[20] L. Stäcker, P. Heidenreich, J. Rambach, and D. Stricker, "Cross-dataset experimental study of radar-camera fusion in bird's-eye view," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 810–814.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[22] F. Bartoccioni, E. Zablocki, A. Bursuc, P. Perez, M. Cord, and K. Alahari, "Lara: Latents and rays for multi-camera bird's-eye-view semantic segmentation," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=abd_D-iVjk0