VS-Sim: A Synthetic Dataset for Viewpoint Shift Robustness

Riccardo Turra[1], Matteo Simoncini[1], Henrique Piñeiro Monteagudo[1,2]*, Aurel Pjetri[1,3], Samuele Salti[2], and Leonardo Taccari[1]

[1] Verizon Connect, Florence, Italy {riccardo.turra, matteo.simoncini, henrique.pineiro.monteagudo, aurel.pjetri, leonardo.taccari}@verizonconnect.com
[2] University of Bologna, Bologna, Italy samuele.salti@unibo.it
[3] University of Florence, Florence, Italy

henriquepm.github.io/VS-SIM

Abstract. We release VS-Sim, a synthetic dataset of road scene images that can be used to study the robustness to viewpoint shift of Computer Vision models for several tasks. Our dataset includes images from a frontal camera in different positions, and annotations for tasks both in frontal view (depth, semantic segmentation) and Bird's Eye View (semantic segmentation). The paper also includes an analysis of the robustness of a Bird's Eye View segmentation model to viewpoint shifts. Experiments indicate that viewpoint shift significantly degrades the performance of a model trained on data from a single viewpoint, and suggest that using multiple viewpoints during training helps to mitigate the performance drop across different scenarios.

Keywords: Synthetic Dataset · Domain Shift · Computer Vision

1 Introduction

Computer Vision models for tasks related to Autonomous Driving (AD) and driving scene understanding, such as object detection, depth estimation, Bird's Eye View reconstruction, are expected to work reliably in different domains and on a long tail of challenging conditions.

In many practical scenarios, due to either the lack of a sufficient amount of data or to the difficulties and costs associated to data annotation, it is common to use *off-the-shelf* models trained on public datasets, possibly fine-tuned on a small dataset for the target domain. However, the target domain can be significantly different from the one on which models have been developed and evaluated, and the performance gap can be wide.

All the largest and most popular datasets, such as KITTI [8], nuScenes [2], Waymo [20], Cityscapes [3], are acquired either from a single vehicle with a very specific camera setup, or from multiple vehicles with a camera setup which is as

* Corresponding author.



Fig. 1: Video frames from real-world frontal dashcams with different viewpoints.

close as possible to a single pre-defined one. This single-setup acquisition introduces an intrinsic Domain Shift problem when switching to applications where the installation setup of a camera, and, thus, the *viewpoint*, may be significantly different from the one in existing public datasets on which models are typically trained.

Indeed, in a real-world application, a frontal camera could be installed in several different positions: if it is an integrated camera, it is often behind the rear-view mirror, while dashcams can be installed virtually anywhere on the windshield or on the dashboard. The installation height is also directly related to the type of vehicle: comparing passenger cars with particularly tall vehicles, such as heavy duty trucks, the difference between the camera height could be in the order of 2 meters. Another source of variability is the camera angle, that can vary depending on the vehicle (cameras on tall vehicles need to point down to capture footage in front of the vehicle) and the quality of installation, that might be suboptimal *e.g.*, for aftermarket devices like dashcams. As a result, different camera setups can have different yaw, pitch and roll angles, height from the ground, and horizontal displacement, as shown in Fig. 1.

Although other domain shifts on driving scenes, such as adverse weather condition or image corruption, have been studied in the scientific literature [6], the robustness of neural network models to camera viewpoint shift is, to the best of our knowledge, an understudied problem in the literature, with few notable exceptions [12]. Building annotated datasets with data from multiple viewpoints covering most of the scenarios that might come up in practice is difficult, costly and time consuming. Such difficulties, however, could be easily overcome with synthetic data, that easily allows us to control for confounders (other sources of uncertainty of domain shift), isolate the effect of the viewpoint shift, and streamline evaluation.

In this paper, with the aim of raising awareness on the problem and speeding up the scientific research toward *viewpoint robustness*, we introduce VS-Sim, a dataset of synthetic driving scenes acquired from different viewpoints. Moreover, we estimate the impact of the viewpoint shift in driving scenes on the problem of Bird's Eye View (BEV) semantic segmentation (SS), a popular task in road scene understanding in which the position of the camera with respect to the scene is crucial. Due to the inherent geometric nature of the task, even small viewpoint changes are likely to impact the performance of a model more than for other tasks, such as semantic segmentation or object detection in perspective view. We show that a BEV SS model trained on a single viewpoint transfers poorly to different ones, especially if the angle with the ground changes, and that simply

training on multiple viewpoints yields a model that generalizes better also on viewpoints that were not used during training, indicating that adding viewpoint augmentations during training can be useful also for tasks that typically require an accurate camera calibration.

2 Related Work

2.1 Domain Generalization in Road Scenes

Domain adaptation and domain generalization are well-known problems in autonomous driving and road scene understanding. Indeed, driving scenes in the wild are known to have a wide range of conditions and contain a long tail of non-nominal events that may introduce a significant domain shift with respect to popular datasets: examples are adverse weather conditions [15], rare safety-critical scenarios [21], as well as different scene geometry [12, 23]. A simple yet effective approach is to accurately curate datasets so that they better align to the data distribution of the target domain, as well as using data augmentation techniques to cope with the scarcity of such data [13, 15]. This is not always possible or sufficient, and there is a wide body of research on techniques to either adapt to a target domain [14, 19] or to train models with better generalization capabilities [24].

2.2 Datasets with Multiple Viewpoints

Compared to other sources of domain shift, viewpoint shift comes with a substantial change in the geometry of the scene, that is challenging to obtain from real-world sensors. Most autonomous driving datasets, such as nuScenes [2], Waymo [20], Cityscapes [3], contain data acquired from rigs with several viewpoints from 360 surrounding cameras. However, this data cannot be directly used to study robustness to small viewpoint shifts of cameras that have the same function (e.g., frontal view).

There are few real-world datasets with driving scenes from multiple frontal viewpoints. A notable example is Mapillary Vistas [16], that contains road images from an extremely diverse set of cameras across the world, with very diverse geographies, environmental conditions and viewpoints. The only available annotations are for semantic segmentation. BDD100K [26] includes 100k annotated images captured by mobile phones used as dashcams. The dataset contains annotations for object detection and some semantic segmentation tasks. Note that in both BDD100K and Mapillary Vistas, camera extrinsics are not known.

Several synthetic datasets in the literature include data acquired from different viewpoints, although robustness to viewpoint shifts is usually not the main focus of the dataset. Virtual KITTI [7] is a synthetic replica of KITTI [8], that includes some variations of the orientation of frontal camera ($\pm 15^\circ$ in the yaw angle). IDDA [1] has data from 5 different vehicles. The main difference is the camera height and the shape of the vehicle hood that is visible in the image. SYNTHIA [18] is another large-scale sythetic dataset focused on semantic

segmentation of urban scenes. It includes 2 frontal cameras with a horizontal displacement, and images from other 6 side/rear cameras. A notable exception is the dataset described in [12], with synthetic data for a total of 36 frontal viewpoints, with perturbations in pitch, yaw, height, and depth.

2.3 Viewpoint Robustness

Several articles have shown that tasks such as Object Detection or Semantic Segmentation are generally quite robust to variability in camera pose, although there might be some degradation of performance to the different appearance that some objects might have from orientations not seen in the training set. [7] reports experiments on Virtual-KITTI around Multiple Object Tracking with 2 non-standard camera orientations ($\pm 15^\circ$ in the yaw angle), showing a degradation of $\sim 5\%$ in MOTA. The work in [1] reports experiments on IDDA about the impact of viewpoint shifts on Semantic Segmentation of the frontal view image. Their results show that SS predictions for an unseen viewpoint are mainly affected by the presence of a different vehicle hood, whose pixels are misclassified. Overall, models do not appear to be heavily affected by the viewpoint shift (much less than other perturbations such as weather), nor do Domain Adaptation methods improve much (up to 4%). The work in [22] shows that Semantic Segmentation does not degrade much, on average around 3% of mIOU, also in very different viewpoints from what was seen during training. The work in [23] proposes a domain adaptation method to reduce the gap due to viewpoint shift for the detection of pedestrians in different real-world datasets (Cityscapes [3] and MOT [4]) in which they appear under significantly different perspectives.

Arguably, viewpoint shift is even more relevant for tasks that require some kind of 3D understanding of the world, such as depth estimation or Bird's Eye View reconstruction. In the depth estimation community, much effort in the last few years has gone towards foundation models that work zero-shot to estimate depth across multiple domains and geometries [9, 11, 25], though we know of no work that attempts to directly measure the effect of different viewpoints. In tasks where the aim is to reconstruct a Bird's Eye View representation, to the best of our knowledge the only relevant existing work is [12], in which the authors measure the impact of viewpoint shift in BEV reconstruction of surrounding vehicles. On their synthetic dataset, they train a model on a standard "source" viewpoint and test it on data from several different viewpoints, unseen during training. They show how a small change in the viewpoint significantly affects BEV reconstruction. Moreover, they propose a viewpoint augmentation technique involving novel view synthesis to transform images to a specific target viewpoint.

In this paper we propose to further expand the work of [12] on *viewpoint robustness* by measuring the performance drop due to viewpoint shift in several realistic positions for camera installations. In addition to [12] we include the case of a non-negligible roll angle, and consider dense BEV semantic segmentation rather then just the detection of surrounding vehicles. Moreover, we look at the problem through the lens of domain generalization, rather than only of domain

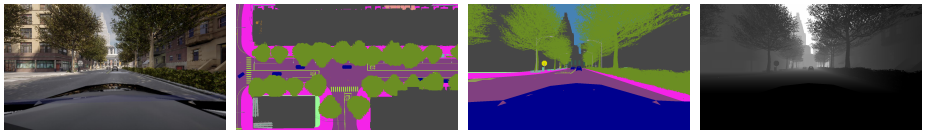


Fig. 2: Frontal RGB image and pixel-level ground truth for BEV semantic segmentation, semantic segmentation, depth.



Fig. 3: Positions of the cameras on the windshield and reference system.

adaptation as in [12], thus under the assumption that the target domain is not known and, potentially, challenging. We show that a naive approach, training a BEV model on a mix of different viewpoints, already helps to obtain a model with better generalization capabilities.

3 The VS-Sim Dataset

The VS-Sim dataset is generated using CARLA [5], a popular open-source simulator for autonomous driving research based on the Unreal engine. To acquire the raw RGB data and the annotations, we use three types of virtual sensors: the RGB camera, acting as a road-facing dashcam to collect frontal view images; the depth camera, that generates depth masks corresponding to the RGB images; the semantic segmentation (SS) camera, used to generate SS ground truth both for frontal views and for the Bird's Eye View. An example of RGB frame and corresponding annotations from the other virtual sensors is presented in Fig. 2.

We place frontal view cameras inside the cabin of the ego-vehicle, in positions that commonly occur in real dashcam installations: one in a default position (top-center, above the rear-view mirror), one on the top-left corner of the windshield, and one at the bottom touching the dashboard. Moreover, we include some common angle perturbations: a roll by 5° to the right and four different pitch rotations (-10°, -5°, 5°, 10°) with respect to the default position. For each position, we capture RGB, depth and semantic segmentation. All the virtual cameras simultaneously record the same scene. The available viewpoints are summarized in Fig. 3 and Table 1.

We configure the intrinsic parameters of the frontal cameras to resemble a commercial dashcam, with a resolution of 1280x720 and a horizontal field of view of 120°. In Fig. 4, we report a sample set of images with the various

Table 1: Summary of the available viewpoints.

Name	Position	Roll	Pitch	Yaw
Default	Top-center of windshield (A)	0	0	0
Bottom	Bottom center of windshield (C)	0	0	0
Top-Left	Top-left of windshield (B)	0	0	0
Roll	Top-center of windshield (A)	+5°	0	0
Pitch^{--}	Top-center of windshield (A)	0	−10°	0
Pitch^{-}	Top-center of windshield (A)	0	−5°	0
Pitch^{+}	Top-center of windshield (A)	0	+5°	0
Pitch^{++}	Top-center of windshield (A)	0	+10°	0

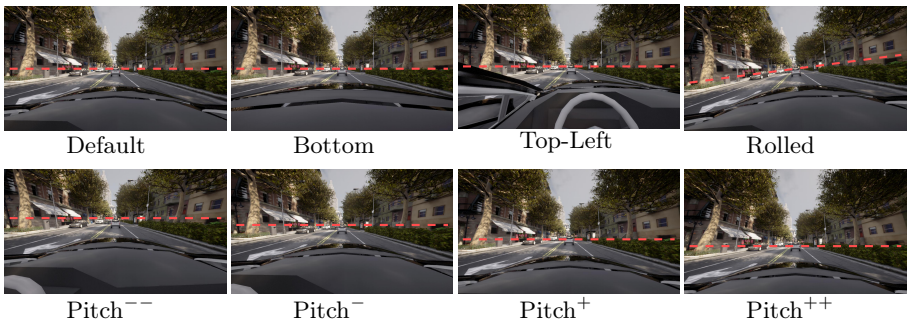


Fig. 4: Viewpoints included in the dataset for each location. Horizon highlighted by a dashed line.

viewpoints. To capture BEV ground truth, we position a SS camera with the same resolution and a FOV of 10° at a height of 650 meters above the scene. With this configuration, we aim to minimize perspective distortion and approximate an orthographic camera, which is not available in CARLA. Semantic segmentation masks contain pixel-level information for a total of 11 classes: road, sidewalk, building, wall, vegetation, terrain, occlusion, person, two-wheeler, car, and truck, analogous to categories in similar real-world datasets [2, 8].

To provide sufficient diversity, data are captured in seven different CARLA maps, namely Town 1, 2, 3, 4, 5, 7, and 10. For each map, we generate 100 different 20-second trajectories, varying the position of the vehicle in the map and the surrounding objects. In total, using a sensor sampling rate of 1Hz, the dataset consists of a total of 14,000 samples for each of the 8 viewpoints, each one including RGB images and frontal view annotations (depth and semantic segmentation), and a mask with BEV semantic segmentation ground truth.

4 Experimental Evaluation

For a fair representation of model performance, all experiments are carried out on spatially separated train/test sets, splitting each town in an area used for training and one used for testing, so that images extracted in the same part of

the map belong to the same split. The train set is composed of a total of 10466 images per position while the test set is composed of 1491 images per position. In the following sections training is always performed on the aforementioned train split of the specified position and evaluation is performed on the test split.

4.1 BEV Semantic Segmentation

The Bird's Eye View (BEV) is a popular representation of the environment surrounding the ego-vehicle for assisted and autonomous driving. It has several desirable properties: the scale is constant regardless of the viewpoint (contrary to perspective view images), distances are metric if the ground is flat, and it is compact, containing most of the relevant information for road scene understanding in a single two-dimensional raster. The task of BEV semantic segmentation from a camera consists in obtaining a class value for each pixel in this top-down representation taking as input a RGB monocular image.

We use SimpleBEV [10] as the BEV semantic segmentation network for our experiments. SimpleBEV is a minimal yet high performance architecture composed of three stages. First, an image encoder processes the RGB image. Then, a parameter-free 3D lifting mechanism fills a 3D feature volume by bilinearly sampling the encoder features in positions obtained by projecting the voxel centers using the camera instrinsics. Finally, those features are collapsed along the vertical direction and fed to a UNet-like network with a semantic segmentation head that generates the final BEV semantic segmentation. We slightly modify the network as in [17] for our setting by switching to a multiclass segmentation task from the original monoclass paradigm and getting rid of the auxiliary task heads.

4.2 Measuring Performance Drop from Viewpoint Shift

To estimate the effect of the viewpoint shift on a BEV model for dense semantic segmentation, we train what we call a "baseline model" on the Default position. Then, we measure the quality of the predicted BEV on acquisitions from other viewpoints. The results are presented in Tab. 2.

Compared to the performance on the default viewpoint that the model was trained on, the results show a drop in performance for all viewpoints, although with very different levels of degradation. On the Bottom and Top-Left positions, despite the ego-vehicle occluding a bigger part of the frames (see Fig. 4), the network demonstrates quite good generalization, resulting in a modest decrease in mIoU (around 10%). The largest drop occurs for classes with rather small objects (car, truck), where even small localization errors have a huge impact on the IoU.

The Rolled viewpoint has a significant impact on the performance, impacting the reconstruction especially on pixels that are outside the central portion of the BEV (i.e., far from the vanishing point). This is quite evident looking at Fig. 5, that compares the prediction of the baseline model on the RGB input from the default viewpoint and the prediction on the rolled RGB input.

Table 2: Performance on BEV SS of a baseline model trained on the Default position, and tested with different camera viewpoints. Best is in bold, second best underlined.

Test viewpoint	Road	Sidew.	Build.	Wall	Veget.	Terrain	Car	Truck	mIoU	Δ
Default	**81.20**	**61.61**	**63.45**	**3.67**	**66.31**	**42.45**	**42.45**	**21.81**	**48.23**	–
Bottom	<u>79.27</u>	55.96	61.60	2.45	65.03	38.99	30.84	<u>12.62</u>	<u>43.22</u>	-10.4%
Top-Left	78.81	<u>56.25</u>	<u>62.25</u>	<u>3.18</u>	<u>65.71</u>	<u>39.79</u>	27.19	8.57	42.72	-11.4%
Roll	58.56	36.86	39.99	1.31	52.68	26.79	<u>31.63</u>	10.90	32.34	-32.9%
Pitch^{--}	34.31	25.10	32.97	0.49	44.19	22.3	1.45	0.18	20.12	-58.3%
Pitch^{-}	55.48	22.00	21.21	0.37	37.94	15.99	1.38	0.71	19.38	-59.8%
Pitch^{+}	30.85	23.43	26.50	0.35	33.56	21.91	0.09	0.00	17.09	-64.6%
Pitch^{++}	41.73	14.38	10.97	0.09	16.16	10.05	0.13	0.00	11.69	-75.8%

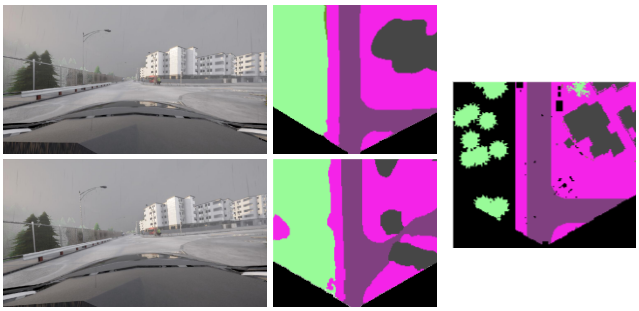


Fig. 5: Example of BEV reconstructions for a non-default viewpoint. On the left, the RGB image from the default position and the rolled viewpoint, respectively. In the middle, prediction of the baseline model on the 2 images. On the right, the BEV ground truth.

Finally, the largest decrease occurs when the viewpoint is pitched. BEV models, such as SimpleBEV, typically construct a 3D feature volume aligned with the camera reference system. If the camera optical axis is not parallel to the ground plane – which is the case when there is pitch – reducing the vertical dimension of this volume results in a plane which is tilted with the respect to the ground plane. This causes the output BEV to be deformed and explains the degradation in the results. In Fig. 6 it is easy to appreciate that even slight rotations along the pitch axis cause a drastic degradation of the quality of the prediction.

4.3 Increasing Viewpoint Robustness on Multiple In-domain Viewpoints

We run a second set of experiments by training the same SimpleBEV architecture on data acquired from a mix of positions. We refer to it as *mixed model*. We hypothesize that augmenting the training dataset with different viewpoints could enhance the model generalization and mitigate the degradation due to viewpoint shifts. Our aim with these experiments is to validate this hypothesis.

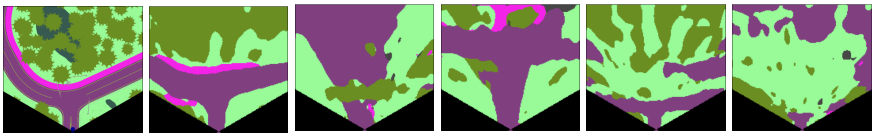


Fig. 6: Example of BEV maps predicted by the baseline model on viewpoints with different pitch.

Table 3: Performance on BEV SS of a model trained on a mixture of viewpoints (Default, Bottom, Top-Left, Roll) and tested on those (in-domain) viewpoints.

Test viewpoint	Road	Sidewalk	Building	Wall	Veget.	Terrain	Car	Truck	mIoU
Default	82.83	60.67	66.01	1.85	69.29	45.23	33.00	1.51	45.05
Bottom	80.96	57.01	62.21	1.20	68.68	43.38	29.46	0.29	42.90
Top-Left	82.61	60.64	65.16	2.01	69.03	44.41	32.55	1.40	44.73
Roll	82.85	60.78	66.16	1.90	69.38	45.10	31.07	1.60	44.86

We train the model on a dataset composed of a mix of viewpoints (Default, Top-Left, Bottom, Rolled) and evaluate on a test set that includes only positions seen at training time. From the results in Tab. 3, we observe that the mixed model can perform almost equally well across all those viewpoints. The only major difference is the class "truck", which drops to almost 0 for all viewpoints. However, note that the number of trucks in the dataset is very small.

In Fig. 7 we compare the results obtained on all viewpoints of the baseline model (trained only on the Default viewpoint) and of the mixed model (trained on Default, Bottom, Top-Left, Roll). The mixed model does significantly better on all positions compared to the baseline model, except for the default viewpoint, indicating that targeting a single viewpoint is preferable, if possible. Still, this experiment shows that a BEV reconstruction model can learn to perform its task on multiple viewpoints, if they do not differ too much and they are provided at training time.

4.4 Increasing Viewpoint Robustness on Unseen Viewpoints

Our last experiment intends to measure if training on multiple viewpoints can help generalize to viewpoints that were *not seen* at training time. We consider again the mixed model trained on a mix of viewpoints (Default, Top-Left, Bottom, Rolled) and evaluate on a test set of unseen viewpoints (Pitch^{--}, Pitch^{-}, Pitch^{+}, Pitch^{++}). The results are reported in Tab. 4.

The performance is not competitive with the results obtained by the same model on viewpoints that were seen in training (Tab. 3). The very low scores on small objects, especially car, show that localization remains very imprecise. However, in Fig. 7 we can observe that the training on multiple viewpoints was able to increase the viewpoint robustness, even for unseen perturbations. More specifically, the results for negative pitch perturbations are rather good, and

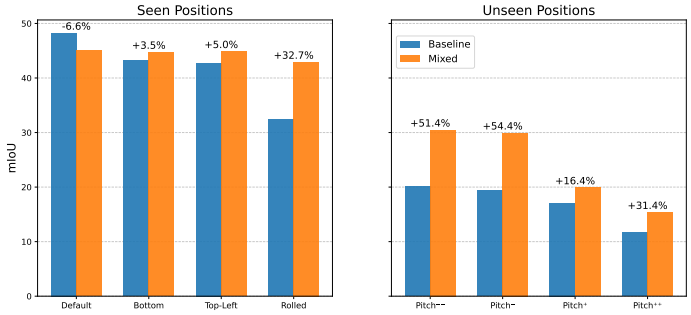


Fig. 7: Comparison of the results of the baseline model (in blue, trained on Default only) and of the mixed model (in orange, trained on Default, Bottom, Top-Left, Roll). Results on viewpoints seen (left) and unseen (right) during training.

Table 4: Performance on BEV SS of a model trained on a mixture of viewpoints (Default, Bottom, Top-Left, Roll) and tested on unseen viewpoints (Pitch^{--}, Pitch^{-}, Pitch^{+}, Pitch^{++}).

Test viewpoint	Road	Sidewalk	Building	Wall	Veget.	Terrain	Car	Truck	mIoU
Pitch^{--}	63.80	40.25	49.26	0.48	56.62	32.79	0.57	0.00	30.47
Pitch^{-}	66.02	34.90	49.40	0.66	56.62	29.40	2.18	0.33	29.94
Pitch^{+}	38.11	27.95	35.64	0.19	36.64	20.58	0.02	0.00	19.89
Pitch^{++}	48.62	16.79	18.92	0.06	26.97	11.41	0.14	0.00	15.36

show that the mixed model generalized significantly better than the Baseline. We hypothesize that this result is due to the similarity between the appearance of objects in the bottom camera pose (seen in training) and this viewpoint, reducing the domain shift between the train and test sets.

5 Conclusion

With VS-Sim, we release a synthetic dataset of road scene images that can be used to study the robustness to viewpoint shifts of modern Computer Vision methods. We also provide the first analysis of viewpoint robustness for Bird's Eye View dense semantic segmentation from monocular camera.

The experimental results show that a BEV model trained on data in a single position is significantly affected even by small variations in viewpoint. Among the camera positions we studied, translations turned out to affect performance the least, while changes of orientation have a much larger impact. Even small positive rotations along the pitch axis (*i.e.*, lower horizon) greatly reduce mIoU scores across all classes, and completely destroy the ability to localize smaller objects.

We plan on extending the dataset to include additional viewpoints and more driving scenarios. Future research can exploit the dataset for analysis on other

tasks such as depth estimation, frontal-view semantic segmentation, and object detection. The dataset could also be used for research on data augmentations that introduce different points of view (*e.g.*, using novel view synthesis) to improve viewpoint robustness.

Acknowledgments

Henrique Piñeiro Monteagudo is supported by the SMARTHEP project, funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086.

References

1. Alberti, E., Tavera, A., Masone, C., Caputo, B.: Idda: A large-scale multi-domain dataset for autonomous driving. IEEE Robotics and Automation Letters **5**(4), 5526–5533 (2020)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
4. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020)
5. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
6. Drenkow, N., Sani, N., Shpitser, I., Unberath, M.: A systematic review of robustness in deep learning for computer vision: Mind the gap? (2022)
7. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: CVPR (2016)
8. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
9. Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9233–9243 (2023)
10. Harley, A.W., Fang, Z., Li, J., Ambrus, R., Fragkiadaki, K.: Simple-bev: What really matters for multi-sensor bev perception? In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2759–2765 (2023)
11. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
12. Klinghoffer, T., Philion, J., Chen, W., Litany, O., Gojcic, Z., Joo, J., Raskar, R., Fidler, S., Alvarez, J.M.: Towards viewpoint robustness in bird's eye view segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8515–8524 (2023)

13. Kong, L., Xie, S., Hu, H., Ng, L.X., Cottereau, B.R., Ooi, W.T.: Robodepth: Robust out-of-distribution depth estimation under corruptions. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
14. Man, Y., Gui, L.Y., Wang, Y.X.: DualCross: Cross-Modality Cross-Domain Adaptation for Monocular BEV Perception. In: IROS (2023)
15. Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M., Brendel, W.: Benchmarking robustness in object detection: Autonomous driving when winter is coming (2020)
16. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE international conference on computer vision. pp. 4990–4999 (2017)
17. Piñeiro Monteagudo, Henrique and Taccari, Leonardo and Pjetri, Aurel and Sambo, Francesco and Salti, Samuele: Rendbev: Semantic novel view synthesis for self-supervised bird's eye view segmentation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025)
18. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234–3243 (2016)
19. Shenaj, D., Fanì, E., Toldo, M., Caldarola, D., Tavera, A., Michieli, U., Ciccone, M., Zanuttigh, P., Caputo, B.: Learning across domains and devices: Style-driven source-free domain adaptation in clustered federated learning. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 444–454 (2023)
20. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
21. Taccari, L., Sambo, F., Bravi, L., Salti, S., Sarti, L., Simoncini, M., Lori, A.: Classification of crash and near-crash events from dashcam videos and telematics. In: 2018 21st International Conference on intelligent transportation systems (ITSC). pp. 2460–2465. IEEE (2018)
22. Testolina, P., Barbato, F., Michieli, U., Giordani, M., Zanuttigh, P., Zorzi, M.: Selma: Semantic large-scale multimodal acquisitions in variable weather, daytime and viewpoints. IEEE Transactions on Intelligent Transportation Systems **24**(7), 7012–7024 (2023)
23. Vidit, V., Engilberge, M., Salzmann, M.: Learning transformations to reduce the geometric shift in object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17441–17450 (2023)
24. Wang, S., Zhao, X., Xu, H., Chen, Z., Yu, D., Chang, J., Yang, Z., Zhao, F.: Towards domain generalization for multi-view 3d object detection in bird-eye-view. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13333–13342 (2023)
25. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
26. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)