SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation

Tao Sun[1*] Mattia Segu[1*] Janis Postels[1] Yuxuan Wang[1]
Luc Van Gool[1] Bernt Schiele[2] Federico Tombari[3,4] Fisher Yu[1]

[1]ETH Zürich [2]MPI Informatics [3]Google [4]Technical University of Munich
{taosun47, segum, jpostels, yuxuwang}@ethz.ch
vangool@vision.ee.ethz.ch, schiele@mpi-inf.mpg.de, tombari@in.tum.de, i@yf.io

Abstract

Adapting to a continuously evolving environment is a safety-critical challenge inevitably faced by all autonomous-driving systems. Existing image- and video-based driving datasets, however, fall short of capturing the mutable nature of the real world. In this paper, we introduce the largest multi-task synthetic dataset for autonomous driving, SHIFT. It presents discrete and continuous shifts in cloudiness, rain and fog intensity, time of day, and vehicle and pedestrian density. Featuring a comprehensive sensor suite and annotations for several mainstream perception tasks, SHIFT allows to investigate how a perception systems' performance degrades at increasing levels of domain shift, fostering the development of continuous adaptation strategies to mitigate this problem and assessing the robustness and generality of a model. Our dataset and benchmark toolkit are publicly available at www.vis.xyz/shift.

1. Introduction

Recent years have witnessed the remarkable progress of perception systems for autonomous driving. Betting on the role that autonomous driving will serve for society, industry [5, 7, 18, 29, 31, 52, 76] and academia [10, 17, 45, 50, 94] have joined forces to collect and release several large-scale driving datasets, raising hopes for a forthcoming successful deployment of self-driving cars.

Providing a playground for different techniques to compete and thrive on multiple tasks, large-scale driving datasets have played a pivotal role in the prosperity of perception algorithms. However, while their accuracy surges, progress in terms of generalization to unforeseen environmental conditions has been underwhelming [11, 47].

*Equal contribution.

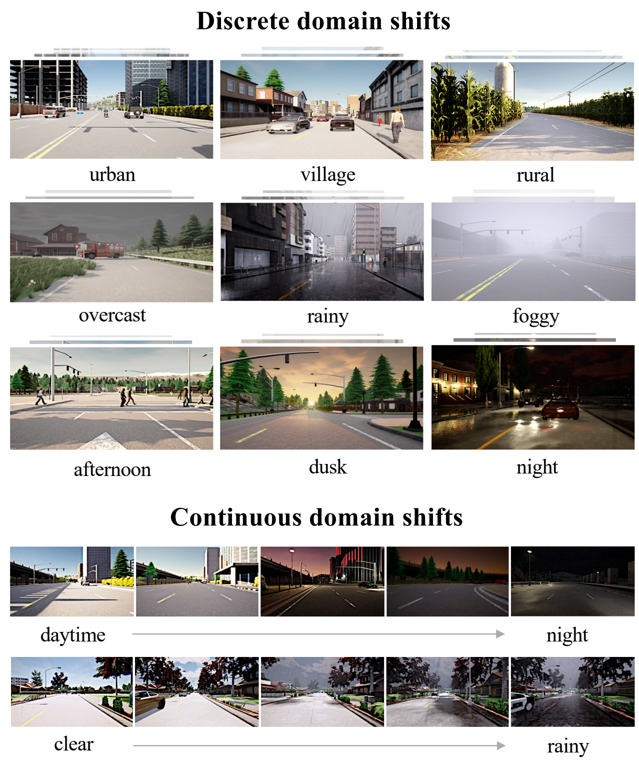


Figure 1. SHIFT provides **(a) discrete domain shifts**: sequences are collected using separated domain parameters and random initial states, used for *robustness evaluation* and *domain adaptation*; **(b) continuous domain shifts**: domain parameters change continuously during driving, used for *continuous domain adaptation*.

To achieve full autonomy, self-driving cars must adapt to new environments and identify life-threatening failure cases to promptly prevent crashes. Examples of domain shifts affecting driving are changes in weather and lighting conditions, scenery, and behavior, appearance, and quantity of agents on the road. Domain shift [2] is a well-known problem for learning algorithms, causing unforeseeable performance drops under conditions different from the training ones. Techniques to prevent, counteract or assess its impact have been developed in the form of, respectively, domain generalization [32, 48, 80, 88], domain adaptation [15, 42, 84, 91], uncertainty estimation [14, 36, 43, 58] and out-of-distribution (OOD) detection [26, 57, 68, 93]. However, such approaches are typically deployed and tested on toy datasets [39,69,86] or synthetically corrupted ones [24]. Although there are preliminary attempts at providing driving datasets with different domains [5, 11, 55, 67, 72, 73, 83, 94], each only covers a limited amount of perception tasks (*e.g.* only semantic segmentation [72,73]) and a narrow selection of domain shift directions (*e.g.* only rain [83] or snow [55]). Consequently, current solutions to domain shift cannot undergo scrutiny in controlled autonomous driving scenarios, making it difficult to verify their safety without risking real-world car crashes.

Given their short length, sequences from existing driving datasets are captured under approximately stationary conditions, and only *discrete shifts* are witnessed among sets of sequences presenting different homogeneous conditions from one set to another (*e.g.* clear weather and rainy). However, as the ancient Greek philosopher Heraclitus of Ephesus uttered, nothing in this world is constant except change and becoming. *Continuous shifts* - the intra-sequence shifts from one domain into another - are a certainty in the real world, where a sunny day can rapidly turn into a rainy one, or a quiet road can quickly become busy. Moreover, continuous distributional shift has recently been shown to represent a critical challenge for current learning systems [59].

An adequate dataset design is thus needed to quantify and address domain shift both at discrete and continuous levels. Consequently, we set the goal of overcoming the outdated paradigm of previous driving datasets and introduce SHIFT, a new synthetic dataset capturing the continuously evolving nature of the real world through realistic discrete and continuous shifts along safety-critical environmental directions: time of day, cloudiness, rain, fog strength, and vehicle and pedestrian density. Collected by means of the CARLA Simulator [13], SHIFT includes a comprehensive sensor suite and covers the most important perception tasks. Counting 4,800+ sequences captured from a multi-view sensor suite in 8 different locations, it supports 13 perception tasks for multi-task driving systems: semantic/instance segmentation, monocular/stereo depth regression, 2D/3D object detection, 2D/3D multiple object tracking (MOT), optical flow estimation, point cloud registration, visual odometry, trajectory forecasting and human pose estimation.

With our dataset, we aim to foster research in several under-explored fields related to the generality and reliability of perception systems for autonomous driving, *e.g.* domain generalization, domain adaptation, and uncertainty estimation. Moreover, by collecting incremental discrete shifts from one domain to another, we hope to foster research in the field of continual learning [20, 87, 90] for autonomous driving, so far only studied on discrete levels of synthetic corruptions [24] of traditional image classification datasets [12, 35]. Finally, by collecting sequences with realistic intra-sequence continuous domain shifts, we provide the first driving dataset allowing research on continuous test-time learning and adaptation [56, 77, 81, 82, 90].

We summarize the main contributions of this work:

- We introduce SHIFT, a multi-task driving dataset featuring the most important perception tasks under a variety of conditions and with a comprehensive sensor setup. To the best of our knowledge, it is the largest synthetic dataset for autonomous driving and provides the most inclusive set of annotations and conditions.
- Using SHIFT, we analyze the importance of modeling discrete and continuous domain shifts, and demonstrate new findings on different adaptation and uncertainty estimation methods under continuous shifts.

2. Related Work

During the past decade, a large variety of realistic and synthetic driving datasets emerged, providing a playground for researchers to develop novel algorithms. Contextually, domain shift has been identified as a common threat to the performance and safety of learning-based methods.

We here introduce the most-notable driving datasets and the techniques to mitigate the domain shift effect. For an overview of the current driving datasets, refer to Tab. 1.

Real-world driving datasets typically focus on a specific subset of perception tasks due to the high cost of data collection and annotation. After almost a decade of development, the pioneering real-world dataset KITTI [17] supports almost all the perception tasks for autonomous driving, including semantic / instance segmentation, depth estimation, 2D and 3D object detection and tracking, optical flow, scene flow, and visual odometry. However, its small scale represents an obvious problem and its diversity is severely limited compared to modern large-scale datasets. CamVid [4], Cityscapes [10], and Mapillary [50] are image-based driving datasets for segmentation, A*3D [54] for 3D object detection, and HD1K [34] for optical flow estimation. Recently, many large-scale datasets, *e.g.*, BDD100K [94], Waymo Open [76], H3D [52], and nuScenes [5], have been released with multi-task annotations, although mainly focusing on object detection and tracking. Our dataset offers a complete set of annotations for all the frames, comprehensive of all the most important perception tasks supported by other datasets, and enabling multi-task learning on a broader range of tasks and conditions.

	Dataset	Cities	Tracking sequences	Max length for sequence	Labels for domain shifts	Annotated frames for					
						Seg.	2D Det.	3D Det.	MOT	Depth	Flow
Real-world	KITTI [17]	1	22	106 sec	no	200	15k	15k	15k	93k	389
	CamVid [4]	4	-	-	no	700	-	-	-	-	-
	Cityscapes [10]	27	-	-	no	25k	-	-	-	-	-
	Cityscapes-C† [47]	27	-	-	discrete	25k	-	-	-	-	-
	H3D [52]	4	160	20 sec	discrete	-	-	27k	27k	-	-
	HD1K [34]	1	-	-	discrete	-	-	-	-	-	1k
	A*3D [53]	1	-	-	discrete	-	-	39k	-	-	-
	nuScenes [5]	2	1,000	20 sec	discrete	-	-	40k	40k	-	-
	Waymo Open [76]	3	1,150	20 sec	discrete	-	200k	230k	230k	-	-
	BDD100K [94]	multiple	2,000	40 sec	discrete	10k	100k	-	318k	-	-
Synthetic	SYNTHIA [67]	3	-	-	discrete	9,000	200k	200k	-	-	-
	GTA-V [65]	1	-	-	no	25k	-	-	-	-	-
	VIPER [64]	1	184	10 min	discrete	320k	320k	-	320k	-	320k
	AIODrive [92]	8	100	100 sec	discrete	100k	100k	100k	100k	100k	-
	SHIFT (*ours*)	8	4,850	33 min	discrete + continuous	2.5M	2.5M	2.5M	2.5M	2.5M	2.5M

Table 1. Comparison of size and supported tasks of existing driving datasets. SHIFT is the largest synthetic dataset and, most notably, the only dataset providing realistic continuous domain shifts, diverse annotations, and longer annotated sequences. † artificially corrupted.

Synthetic driving datasets are collected using graphic engines and physical simulators. SYNTHIA [67] contains images and segmentation annotations generated by its simulator. AIODrive [92] is produced using CARLA Simulator with multiple sensor support, focusing on high-density long-range LiDAR sets. Compared to ours, these datasets present sequences of limited length and are restricted to discrete domain labels (Tab. 1). Further, video games have also been used for data generation. GTA-V [28, 65] provides images and segmentation masks captured from a popular game. VIPER [64] extends GTA-V by providing optical flow masks and discrete environmental labels. However, low-level control of video game engines is hardly accessible, impeding fine-grained environmental control and the collection of continuous shifts.

Adverse conditions datasets support the evaluation of robustness under different OOD conditions. A recent work [44] collects meteorological and air temperature measurements under discrete real-world shifts. Image-based datasets, *e.g.* CIFAR10/100-C [47], ImageNet-R [23] and Cityscapes-C [24], have been generated by applying artificial corruptions such as blurring, additive Gaussian noise and addition of specific patterns on the original dataset. Though carefully designed, such ad-hoc corruptions cannot fully represent the challenges presented by visual shifts in the real world. To this end, recent driving datasets [5, 45, 53, 76, 94] provide manually labeled tags for various weather conditions, scene categories, and day periods. However, each only covers a limited amount of perception tasks (see Tab. 1) and a narrow selection of domain shift directions. Moreover, ad-hoc datasets have been collected for specific underrepresented domains, *e.g.* rain [30, 83], fog [71, 72, 78], night [11], snow [55]. However, domain tags remain coarse-grained and only certain tasks and domain shift directions are supported. Recently, the ACDC dataset [73] has been proposed, featuring images evenly distributed between fog, nighttime, rain, and snow. However, it supports only semantic segmentation. Interestingly, the India Driving Dataset [85] is the only dataset to provide extremely busy roads as adverse conditions. Overall, BDD100K [94] is the large-scale real-world dataset presenting the largest diversity of perception driving tasks and discrete domain labels for the time of day and weather conditions. For this reason, we use it as a reference to validate empirical observations drawn from our dataset. Nevertheless, compared to our dataset, BDD100K only provides annotated images from single cameras, does not provide 3D bounding boxes and optical flow annotations, distribution of domains is highly imbalanced and the domain is stationary within each sequence. In contrast, our dataset provides a full sensor suite, annotations for multiple tasks, balanced domain distribution and sets of sequences with continuously changing time of day, weather conditions (cloudiness, rain and fog strength), and vehicle and pedestrians density.

Unsupervised domain adaptation (UDA) means simultaneously learning on a labeled source and an unlabeled target domain to find transferable features across domains. UDA is mainly achieved via feature-space alignment [60, 75], domain-consistent regularization [15, 16, 27] and minimization of surrogate functions of domain gaps [70, 89]. The discrete shifts provided in our dataset can be directly used for training and evaluating UDA approaches.

Continual domain adaptation aims at performing consecutive discrete adaptation steps from one domain to multiple others. Incremental domain adaptation (IncDA) is a subset of continual DA that requires the source data and assumes availability of intermediate domains where domain shifts

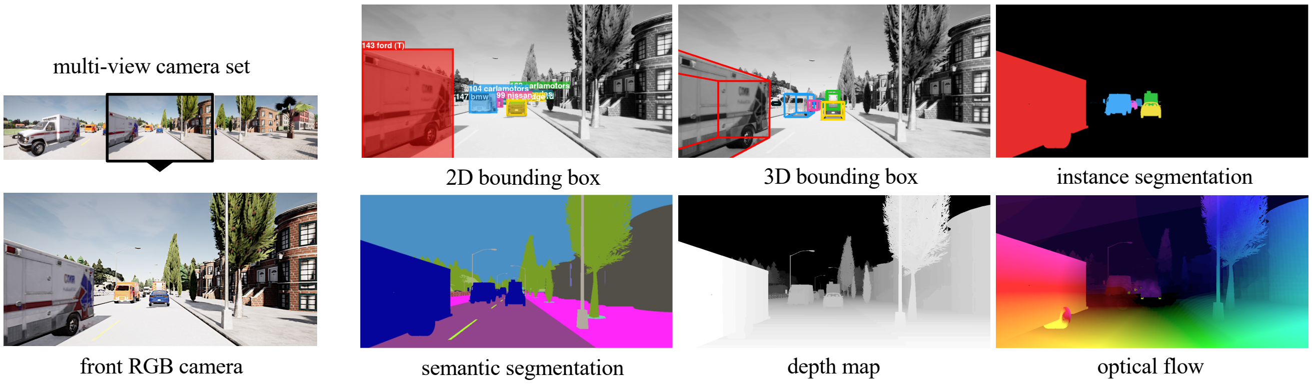


Figure 2. The annotation set of the RGB camera in our dataset. Each frame is associated with annotations of 2D/3D bounding boxes with tracking identities (visualized by different colors), semantic/instance segmentation, depth map and optical flow label.

occur gradually [37, 87, 90], allowing to minimize the gap between adaptation steps and performing adaptation from the source to the final target domain more effectively than with direct UDA. Providing different strengths of variations along natural axes, our dataset is suitable for IncDA.

Continuous test-time adaptation (ContinuousTTA) assumes that gradual domain shifts occur within the same test sequence, and adaptation is performed at test time on the incoming data stream. ContinuousTTA is a suitable choice for any scenario where a model is required to adapt on the go to a shifting domain and no large labeled or unlabeled collection of data from the target domain is available in advance. Recent works [49, 77, 90] show the efficiency of TTA when applied to artificial corruptions in the image-based datasets ImageNet-C/-R [23, 24]. The continuously shifting video sequences in our dataset provide instead realistic domain shift along natural directions, facilitating the development of ContinuousTTA methods transferable to the real world.

Uncertainty Estimation is a fundamental task for safety-critical vision applications. Quantifying the confidence about a model's prediction allows avoiding dangerous failures in autonomous driving. However, current uncertainty estimation techniques [14, 36, 40, 57] mainly focus on classification on toy datasets [35, 38], while recent work [59] has observed poor calibration, *i.e.* uncertainty uncorrelated with prediction's error, when such techniques are extended to more difficult datasets [25] and tasks under distributional shift. We hope that the domain shifts and multiple tasks supported in SHIFT will enable the study of uncertainty estimation methods on a wide variety of tasks for autonomous driving and their calibration under distributional shift.

3. The SHIFT Dataset

We provide a driving dataset with a comprehensive sensor suite (Sec. 3.1) and a rich set of annotations (Sec. 3.2), supporting multiple image- and video-based perception and forecasting tasks against environmental changes. We detail our design choices regarding domain shifts in Sec. 3.3.

3.1. Sensor Suite

We collect the data through a comprehensive sensor suite. Our sensor suite features 11 different sensors, including a multi-view RGB camera set with 5 cameras, a stereo RGB camera set, an optical flow sensor, a depth camera, a GNSS sensor, and an IMU. All the cameras have a field-of-view of $90°$ and resolution of 1280×800 pixel. Moreover, we provide point clouds captured by a 128-channel LiDAR sensor. All sensors are synchronized and captured at a 10Hz rate. We follow the Scalabel [1] format and right-hand coordinate systems for storing all the annotations. More details are in the Appendix.

3.2. Annotations

We provide annotations for multiple mainstream perception tasks in autonomous driving, including 2D/3D bounding box trajectories, instance/semantic segmentation, optical flow and dense depth. Unlike real-world datasets, whose annotations are often limited to a group of keyframes due to prohibitive labeling cost, we offer full annotations for each frame in the sequences. More details are in the Appendix.

3.3. Dataset Design

Given their short sequence length, existing driving datasets are captured under approximately stationary conditions, and only discrete shifts are witnessed among sets of sequences presenting different homogeneous conditions (*e.g.* clear weather and rainy). We set the goal of overcoming the outdated paradigm of previous driving datasets and introduce SHIFT, a new synthetic dataset capturing the continuously evolving nature of the real world through realistic discrete and continuous shifts along safety-critical environmental directions: time of day, cloudiness, rain, fog strength, and vehicle and pedestrian density. We collect 4,850 sequences, of which 4,250 contain stationary environ-

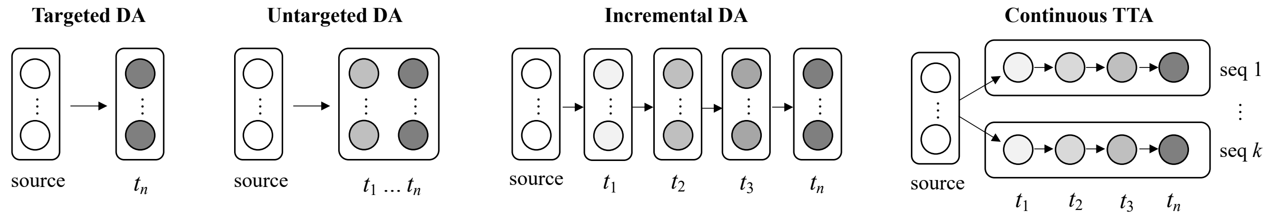


Figure 3. We evaluate four adaptation strategies: targeted domain adaptation (Targeted DA), untargeted domain adaptation (Untargeted DA), incremental domain adaptation (Incremental DA) and continuous test-time adaptation (Continuous TTA). The dots in the same row represent frames from the same sequence; their grayscale marks the degree of domain shift (white dots = source, dark gray dots = target.)

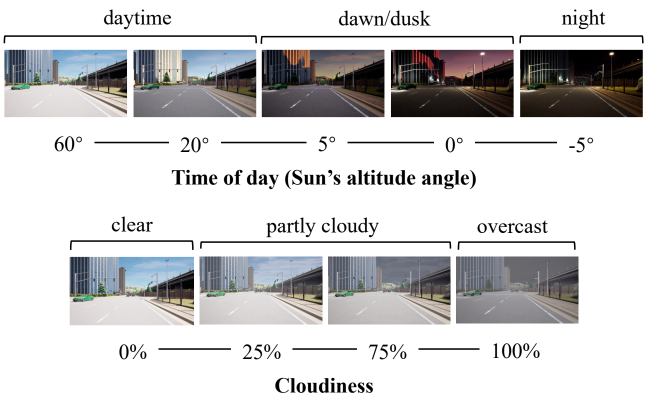


Figure 4. Examples of the two-level structure for domain labels. Each discrete label (tag above images) corresponds to an interval of continuous labels (*i.e.*, severity, axis below images).

mental conditions, *i.e.* inter-sequence domain shift. Each sequence is composed of 500 frames collected at 10 Hz, equivalent to 50 seconds of driving time. The remaining 600 sequences have continuously shifting conditions, *i.e.* inter-sequence domain shift. Totalling 70+ hours of driving and 2,500,000 annotated frames, SHIFT is the largest synthetic driving dataset available.

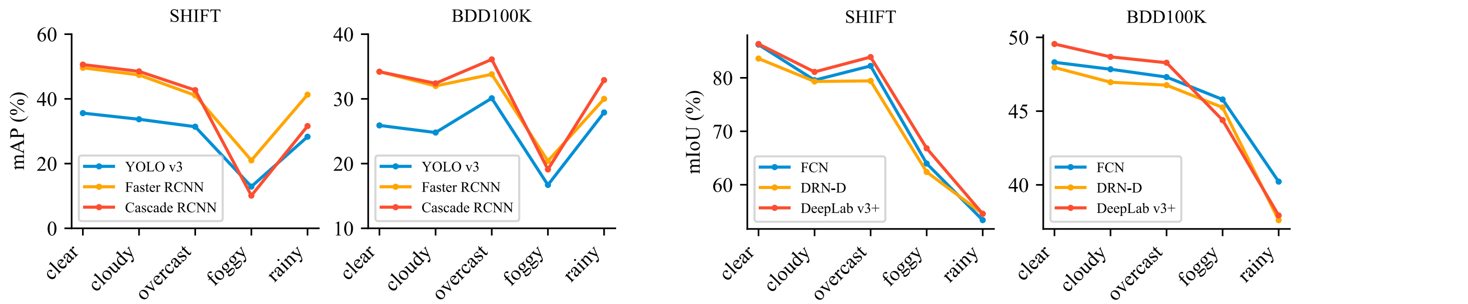
Domain shift types. We consider the most-frequent real-world environmental changes. SHIFT provides domain shifts in (a) weather conditions, including cloudiness, rain, and fog intensity, (b) time of day, (c) the density of vehicles and pedestrians, and (d) camera orientation.

Domain shifts level. To facilitate research on domain adaptation in different scenarios, SHIFT provides two levels of domain shifts, namely discrete domain shifts and continuous domain shifts. The *discrete* set contains 4,250 sequences generated with fixed environmental parameters and random initial states. We group these sequences into different domains, according to their severity. Fig. 4 shows grouping examples. All possible domain combinations are uniformly distributed across all sequences. The *continuous* set contains additional 600 sequences with continuous domain variations. In particular, each sequence presents a gradual shift from one domain to another, where the shift happens through the intermediate domains that would naturally occur in the real world. In total, we collect 500 sequences of a basic 20 seconds length (1x), 80 sequences 10x longer than the basic length, and 20 100x longer. Each set is uniformly divided among the following shifts, each of which also loops back to the source domain: day $\rightarrow$ night, clear $\rightarrow$ rain, clear $\rightarrow$ foggy, clear $\rightarrow$ overcast. Given a domain shift direction, *e.g.* day to night, all other domain parameters are uniformly distributed across all sequences. Different sequence lengths allow analyzing the impact of domain shift speed on continuous TTA strategies (Sec. 4.2).

4. Experiments

SHIFT allows studying the robustness of perception systems for driving under both discrete and continuous distributional shifts. We first (Sec. 4.1) assess the impact of discrete domain shifts on model performance for multiple perception tasks available in our dataset and empirically demonstrate that observations from our simulation dataset transfer to real-world datasets. Moreover, we compare different discrete adaptation strategies and assess the calibration of uncertainty estimation methods under domain shifts. In Sec. 4.2 we extend the analysis to continuous domain shifts and investigate properties of continuous domain adaptation methods [90] against incremental adaptation and unsupervised domain adaptation [89]. Further experiments, implementation details, and ablations on the data collection choices are reported in the Appendix, together with additional experiments on multitask learning.

Domain adaptation strategies. To analyze the impact of our dataset design choice, we examine the four domain adaptation strategies allowed by our dataset (Fig. 3). As *Baseline*, we consider the model trained on the source domain only and directly tested on the other domains. *Targeted DA* [91] is a traditional computer vision problem consisting of adapting from a labeled source domain to a specific unlabeled target domain. We define *Untargeted DA* [39, 74] as adapting from a labeled source domain to a set of various unlabeled shifted domains. *Incremental DA* [87] consists in performing incremental steps of targeted



(a) Object detection (b) Semantic segmentation

Figure 5. Performance degradation for different object detection (left) and semantic segmentation (right) methods under different weather conditions. Every model is trained under clear weather conditions and tested on other domains. SHIFT shows a similar trend as BDD100K.

Task	**Method**	**Metric**	<u>clear-daytime</u>	partly cloudy	overcast	foggy	rainy	dawn/dusk	night
Semantic segmentation	DRN-D [95]	mIoU (%) ↑	83.6	79.3	79.4	62.4	54.6	60.8	42.8
Instance segmentation	Mask R-CNN [21]	mAP (%) ↑	39.3	39.4	34.0	18.7	35.0	30.7	13.1
Object detection	Faster R-CNN [6]	mAP (%) ↑	46.9	47.4	41.1	21.0	41.3	37.3	15.4
MOT	QDTrack [51]	MOTA (%) ↑	56.2	53.4	46.2	25.0	41.9	44.7	16.5
Mono. depth estimation	AdaBins-UNet [3]	SILog ↓	9.6	10.0	8.9	12.0	10.3	19.7	27.9
Optical flow estimation	RAFT [79]	EPE (px) ↓	2.26	2.01	2.35	2.60	2.43	4.17	8.85

Table 2. Performance degradation on SHIFT of different methods for different perception tasks under discrete domain shifts. Training domain is underlined. The test domains are weather variations in daytime (partly cloudy, overcast, foggy, rainy) and time of day variations in clear weather (dawn/dusk, night). ↑ (↓): the higher (lower) the better.

Scenario	Baseline	Targeted DA	Incremental DA
<u>daytime</u> → night	42.8	45.3	**47.3**
<u>clear</u> → foggy	**62.4**	59.1	57.3
<u>clear</u> → rainy	54.6	61.0	**64.9**

Table 3. Comparison of different adaptation strategies for semantic segmentation under three directions of domain shift. The source domain is underlined. Incremental DA improves over Targeted DA, except for the case when Targeted DA underperforms the baseline. (Baseline = without DA)

DA from the source domain to the target domain passing through intermediate discretely-shifted domains. *Continuous TTA* [90] aims at adapting frame by frame to a sequence presenting a continuously shifted domain from source to target domain.

Implementation details. For the adaptation tasks, we focus on semantic segmentation and use ADVENT [89] for the Targeted and Untargeted DA. The segmentation backbone is DRN-D-54 [96]. Incremental DA is performed as a series of Targeted DA steps, while for Continuous TTA we extend TENT [90] to semantic segmentation and iteratively apply it on every incoming frame. Every model is trained in the clear-daytime domain and tested under different weather domains. While our dataset provides finer domain labels depending on the severity of the perturbation, we group different degrees of severity to match the environmental labels in BDD100K [94] in order to assess the compatibility of conclusions drawn from our dataset with real-world trends.

4.1. Discrete Shifts

As outlined in Sec. 3.3, our dataset provides incremental discrete shifts along natural environmental directions. We investigate properties of discrete shifts on the multitude of supported tasks and report findings on domain adaptation and uncertainty estimation performance.

Impact of domain shift. We find that many mainstream algorithms for different perception tasks suffer performance drops under domain shift (Tab. 2), where the severity increases with the distance from the source domain. In particular, we train all models in the clear-daytime domain and test under different weather conditions, showing the overall negative impact of domain shift on all the vision tasks supported by our dataset. Nevertheless, in some specific cases a model may even perform better on a shifted domain, *e.g.* instance segmentation on overcast. We leverage the incremental domain shifts provided in our dataset to investigate in Tab. 3 different discrete adaptation strategies for semantic segmentation, *i.e.* Incremental DA and Targeted DA. We find that incrementally adapting from source to target domain improves the generalization to the target domain compared to direct Targeted DA. However, clear → foggy represents a challenging scenario for which both the adaptation strategies worsen the baseline performance.

Real-world compatibility. To establish a reliable benchmark we must first confirm that trends witnessed in our simulation dataset are compatible with real-world observations. We use BDD100K [94] for comparison because it

	Method	clear-daytime	cloudy	overcast	foggy	rainy	dawn/dusk	night	**OOD avg.**
SHIFT	Softmax	3.3	32.6	14.2	48.8	64.3	43.7	64.7	45.2
	MCDO	1.2	13.1	7.6	20.8	10.0	27.2	39.6	19.7
	Ensemble	1.4	12.3	7.5	23.4	8.9	18.7	36.9	18.0
BDD	Softmax	9.6	23.2	9.9	9.7	7.7	10.6	48.6	18.4
	MCDO	12.3	22.0	7.8	13.0	11.4	13.1	41.4	18.1
	Ensemble	12.6	18.8	9.2	11.7	11.8	13.9	39.8	17.5

Table 4. Calibration (ECE, %) of uncertainty estimation methods under distributional shift for semantic segmentation. The lower, the better. Source domain is clear-daytime. We find that calibration worsens far from the source, both for SHIFT and BDD100K.

features the largest subset of our tasks available in a real-world dataset with discrete domain labels. We study the domain shift effect on two fundamental perception tasks, *i.e.* 2D object detection and semantic segmentation, and show compatible trends for different methods trained on SHIFT and BDD100K (Fig. 5). We evaluate the one-stage method YOLO v3 [62], as well as the two-stage methods Faster R-CNN [63] and Cascade R-CNN [6] for object detection. For semantic segmentation, we consider three different methods, FCN [41], DRN-D [95], and DeepLab v3+ [8]. Our experiments suggest that the performance of different methods for semantic segmentation and object detection degrades under different domain shifts. Moreover, we find that the ranking of methods and the relative degradation trend is compatible between SHIFT and the real-world dataset BDD100K, confirming the usefulness of SHIFT and its consistency with the real world.

Uncertainty estimation. Autonomous driving systems must deal with life-threatening failure cases. To this end, uncertainty estimation represents a powerful tool to assess the reliability of a model's predictions. Following [19], we evaluate the Expected Calibration Error (ECE) to assess the calibration, *i.e.* correlation with model error, of uncertainty estimation methods under domain shift. In particular, we evaluate the Softmax Entropy baseline and traditional Bayesian techniques such as Monte-Carlo Dropout (MCDO) [14] and Deep Ensembles [36]. We observe that such uncertainty estimation methods are not well calibrated under domain shift, and that calibration worsens under incremental shifts on both SHIFT and BDD100K (Tab. 4). While some domains are more challenging in SHIFT than in BDD100K, the overall degradation of calibration observed on SHIFT is confirmed on BDD100K and the ranking of methods is preserved, further highlighting that conclusions drawn from our dataset transfer to the real world.

We hope that our dataset will help researchers providing solutions to the potentially life-threatening shortcomings of current DA and uncertainty estimation techniques.

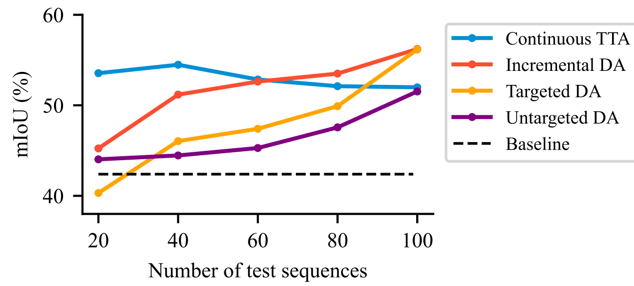


Figure 6. Comparison of different adaptation strategies for semantic segmentation on daytime → night shifts at varying amounts of available sequences. TTA is the most effective under limited amounts of data. When enough data becomes available, Incremental DA outperforms all other alternatives.

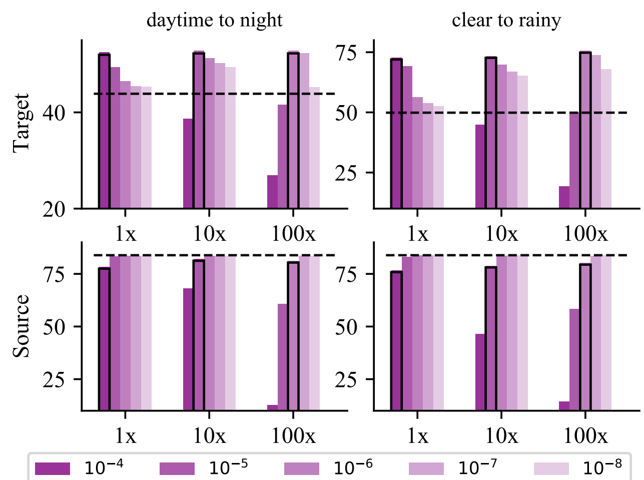


Figure 7. Performance on the target domain of TTA for different sequence lengths. Best learning rate on target domain is highlighted by black boxes. Both source and target performance are highly sensitive to the learning rates. Dashed lines = before TTA.

4.2. Continuous Shifts

A key feature of SHIFT is that of providing a set with continuous intra-sequence domain shifts, allowing to compare different adaptation strategies under continuous shifts and provide an in-depth analysis on TTA and its properties.

Continual domain adaptation. Fig. 6 compares four different adaptation strategies for semantic segmentation on an increasing number of sequences. Given a model pretrained on the source domain, *i.e.* clear-daytime, and the set of continuously shifting sequences from one domain to another, *i.e.* clear-daytime → night, we train the TTA algorithm on each frame of the incoming data stream. TTA is thus performed independently on each sequence. Final performance is averaged over all the sequences. For the other adaptation strategies, we divide the length of the sequence in 20 bins, consider each bin as a separate domain, and group corresponding bins from all the provided sequences. For

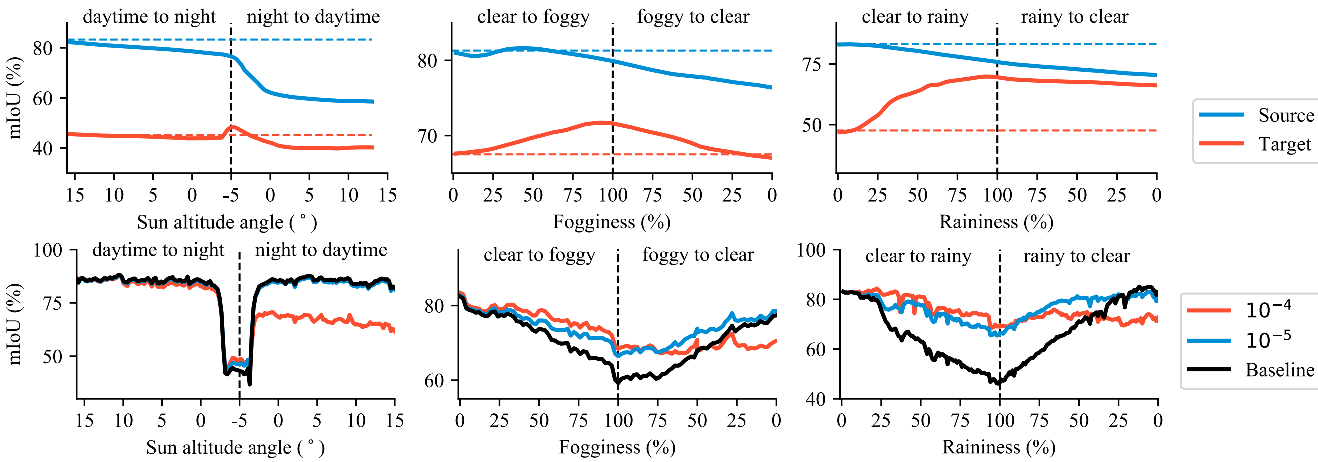


Figure 8. Performance of TTA for semantic segmentation under three types of domain shift: daytime → night, clear → foggy, clear → rainy. Each point corresponds to the performance of the model on the source (top-blue) / target (top-red) / current (bottom) domain finetuned up to that level of domain shift in the sequence. Horizontal lines in the bottom figure represent the original performance on source (blue) and target domain (red). After reaching the target domain, every sequence loops back to the original source domain. Catastrophic forgetting can be observed by the drop in source performance during TTA.

Targeted DA, we thus adapt directly to the last bin, corresponding to the night domain. Untargeted DA is instead applied on all the bins but the source one. Incremental DA is performed by incrementally adapting from one bin to the consecutive one until the end of the sequence is reached. In particular, we plot the average mIoU against the number of training sequences (Fig. 6). We find that TTA is extremely efficient under small target data availability compared to all other alternatives, and that Incremental DA is consistently more effective than both Targeted and Untargeted DA.

Test-time adaptation. As intra-sequence continuous shifts represent one of the main contributions of SHIFT, we further focus on TTA by using TENT [90] and evaluate the effect of the speed at which domain shift happens within a sequence on TTA performance (Fig. 7). This is made possible by the sets of sequences of different lengths (1x, 10x, 100x the basic sequence length).

Given a source and a target domain, *e.g.* daytime and night, each sequence starts from the source domain and reaches the target domain at mid-sequence length; then, it loops back to the original domain. We first observe that, depending on the domain shift speed, the learning rate can highly affect the outcome of the TTA (Fig. 7). Slower (faster) shifts will require lower (higher) learning rates. Moreover, after reaching the target domain at mid-sequence, the performance on the target domain has improved compared to its original value, while that on the source domain has dropped. According to Fig. 7 (1x), we find that the optimal learning rate in terms of adaptation to the target domain leads to the largest performance drop on the original source (Fig. 8, top). This problem, known as catastrophic forgetting [33] in the continual learning literature, has already been observed for class- and task-incremental learning.

To further investigate this issue, we loop back to the original domain after adapting to the target and find that, while the performance on the current target domains largely improves over the baseline (Fig. 8, bottom), the original source domain accuracy cannot be recovered (Fig. 8, top). While TTA has shown to be extremely effective to adapt on the go, a model adapted with TTA cannot be safely deployed on the original source domain. Showing that catastrophic forgetting also affects test-time adaptation further demonstrates the importance of providing continuously shifted sequences in driving datasets, and we hope that future research will attempt to mitigate this problem.

5. Conclusion

We introduce SHIFT, a multi-task driving dataset featuring the most important perception tasks under discrete and continuous domain shifts. Thanks to our dataset design, we demonstrate several new findings on different adaptation strategies and uncertainty estimation methods. Although simulation environments are still far from being a perfect representation of the real world, they allow inexpensive data collection and annotation. Moreover, we empirically demonstrate that conclusions drawn from our dataset hold in real-world datasets. To the best of our knowledge, SHIFT is the largest synthetic dataset for autonomous driving, providing the most inclusive set of annotations and conditions. We hope that providing the first dataset with realistic continuous domain shifts will contribute to shaping the data collection paradigm for real-world driving datasets and promote advances in test-time learning and adaptation.

References

[1] Scalabel: A scalable open-source web annotation tool, howpublished = https://scalabel.ai/, note = Accessed: 2021-11-16,. 4

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. 1

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 6, 19

[4] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 2, 3

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2, 3, 13

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 6, 7, 19

[7] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3D tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019. 1

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 7

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 19

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2, 3, 13

[11] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE, 2018. 1, 2, 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[13] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 2

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 2, 4, 7

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 3

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 3

[17] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 2, 3, 13, 14, 19

[18] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2D2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 1

[19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 7

[20] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International Conference on Machine Learning*, pages 2535–2544. PMLR, 2019. 2

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6, 16, 19

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 19

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 3, 4

[24] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 2, 3, 4

[25] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 4

[26] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 2

[27] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 3

[28] Hou-Ning Hu, Yung-Hsu Yang, Tobias Fischer, Trevor Darrell, Fisher Yu, and Min Sun. Monocular quasi-dense 3d object tracking. *arXiv preprint arXiv:2103.07351*, 2021. 3

[29] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The ApolloScape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2702–2719, 2019. 1

[30] Jiongchao Jin, Arezou Fatemi, Wallace Lira, Fenggen Yu, Biao Leng, Rui Ma, Ali Mahdavi-Amiri, and Hao Zhang. RaidaR: A rich annotated image dataset of rainy street scenes. *arXiv preprint arXiv:2104.04606*, 2021. 3

[31] 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. Level 5 perception dataset 2020. https://level-5.global/level5/data/, 2019. 1

[32] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012. 2

[33] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 8

[34] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–28, 2016. 2, 3

[35] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55:5, 2014. 2, 4

[36] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017. 2, 4, 7

[37] Qicheng Lao, Xiang Jiang, Mohammad Havaei, and Yoshua Bengio. Continuous domain adaptation with variational domain-agnostic feature replay. *arXiv preprint arXiv:2003.04382*, 2020. 4

[38] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 4

[39] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 2, 5

[40] Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Conference on Neural Information Processing Systems*, 2020. 4

[41] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 7, 19

[42] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2

[43] Antonio Loquercio, Mattia Segu, and Davide Scaramuzza. A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2):3153–3160, 2020. 2

[44] Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021. 3

[45] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Chunjing Xu, et al. One million scenes for autonomous driving: Once dataset. 2021. 1, 3

[46] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 19

[47] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 1, 3

[48] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013. 2

[49] Chaithanya Kumar Mummadi, Robin Hutmacher, Kilian Rambach, Evgeny Levinkov, Thomas Brox, and Jan Hendrik Metzen. Test-time adaptation to distribution shift by confidence maximization and input transformation. *arXiv preprint arXiv:2106.14999*, 2021. 4

[50] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 1, 2

[51] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 164–173, 2021. 6

[52] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9552–9557. IEEE, 2019. 1, 2, 3

[53] Quang-Hieu Pham, Pierre Sevestre, Ramanpreet Singh Pahwa, Huijing Zhan, Chun Ho Pang, Yuda Chen, Armin Mustafa, Vijay Chandrasekhar, and Jie Lin. A* 3d dataset: Towards autonomous driving in challenging environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2267–2273. IEEE, 2020. 3

[54] Quang-Hieu Pham, Pierre Sevestre, Ramanpreet Singh Pahwa, Huijing Zhan, Chun Ho Pang, Yuda Chen, Armin Mustafa, Vijay Chandrasekhar, and Jie Lin. A* 3d dataset: Towards autonomous driving in challenging environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2267–2273. IEEE, 2020. 2

[55] Matthew Pitropov, Danson Evan Garcia, Jason Rebello, Michael Smart, Carlos Wang, Krzysztof Czarnecki, and Steven Waslander. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research*, 40(4-5):681–690, 2021. 2, 3

[56] Matteo Poggi, Alessio Tonioni, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Continual adaptation for deep stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[57] Janis Postels, Hermann Blum, Yannick Strümpler, Cesar Cadena, Roland Siegwart, Luc Van Gool, and Federico Tombari. The hidden uncertainty in a neural networks activations. *arXiv preprint arXiv:2012.03082*, 2020. 2, 4

[58] Janis Postels, Francesco Ferroni, Huseyin Coskun, Nassir Navab, and Federico Tombari. Sampling-free epistemic uncertainty estimation using approximated variance propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2931–2940, 2019. 2

[59] Janis Postels, Mattia Segu, Tao Sun, Luc Van Gool, Fisher Yu, and Federico Tombari. On the practicality of deterministic epistemic uncertainty. *arXiv preprint arXiv:2107.00649*, 2021. 2, 4

[60] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and N Lawrence. Covariate shift and local learning by distribution matching, 2008. 3

[61] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 19

[62] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 7, 19

[63] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 7, 18, 19

[64] Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2232–2241, 2017. 3, 19

[65] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 3

[66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 16, 19

[67] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. 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*, pages 3234–3243, 2016. 2, 3

[68] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021. 2

[69] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 2

[70] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019. 3

[71] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 687–704, 2018. 3

[72] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 2, 3

[73] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. *arXiv preprint arXiv:2104.13395*, 2021. 2, 3

[74] Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *arXiv preprint arXiv:2011.12672*, 2020. 5

[75] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. In *Domain Adaptation in Computer Vision Applications*, pages 153–171. Springer, 2017. 3

[76] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 1, 2, 3, 13

[77] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pages 9229–9248. PMLR, 2020. 2, 4

[78] Jean-Philippe Tarel, Nicholas Hautiere, Laurent Caraffa, Aurélien Cord, Houssam Halmaoui, and Dominique Gruyer. Vision enhancement in homogeneous and heterogeneous fog. *IEEE Intelligent Transportation Systems Magazine*, 4(2):6–20, 2012. 3

[79] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 6, 19

[80] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. 2

[81] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip HS Torr. Learning to adapt for stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9661–9670, 2019. 2

[82] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 195–204, 2019. 2

[83] Frederick Tung, Jianhui Chen, Lili Meng, and James J Little. The raincouver scene parsing benchmark for self-driving in adverse weather and at night. *IEEE Robotics and Automation Letters*, 2(4):2188–2193, 2017. 2, 3

[84] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 2

[85] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1743–1751. IEEE, 2019. 3

[86] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 2

[87] Riccardo Volpi, Diane Larlus, and Grégory Rogez. Continual adaptation of visual representations via domain randomization and meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4443–4453, 2021. 2, 4, 5

[88] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018. 2

[89] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 3, 5, 6

[90] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *Int. Conf. Learn. Represent. (ICLR)*, 2021. 2, 4, 5, 6, 8

[91] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 2, 5

[92] Xinshuo Weng, Yunze Man, Dazhi Cheng, Jinhyung Park, Matthew O'Toole, and Kris Kitani. All-In-One Drive: A Large-Scale Comprehensive Perception Dataset with High-Density Long-Range Point Clouds. *arXiv*, 2020. 3

[93] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 2

[94] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3, 6

[95] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 6, 7, 16, 18, 19

[96] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 6