IEEE Robotics and Automation Letters (RAL) paper presented at the
2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

IDDA: a large-scale multi-domain dataset for autonomous driving

Emanuele Alberti*,[1], Antonio Tavera*,[1], Carlo Masone[2] and Barbara Caputo[1]

***Abstract*— Semantic segmentation is key in autonomous driving. Using deep visual learning architectures is not trivial in this context, because of the challenges in creating suitable large scale annotated datasets. This issue has been traditionally circumvented through the use of synthetic datasets, that have become a popular resource in this field. They have been released with the need to develop semantic segmentation algorithms able to close the visual domain shift between the training and test data. Although exacerbated by the use of artificial data, the problem is extremely relevant in this field even when training on real data. Indeed, weather conditions, viewpoint changes and variations in the city appearances can vary considerably from car to car, and even at test time for a single, specific vehicle. How to deal with domain adaptation in semantic segmentation, and how to leverage effectively several different data distributions (source domains) are important research questions in this field. To support work in this direction, this paper contributes a new large scale, synthetic dataset for semantic segmentation with more than 100 different source visual domains. The dataset has been created to explicitly address the challenges of domain shift between training and test data in various weather and view point conditions, in seven different city types. Extensive benchmark experiments assess the dataset, showcasing open challenges for the current state of the art. The dataset will be available at: `https://idda-dataset.github.io/home/`.**

I. Introduction

With the latest advancements in Deep Learning, we are starting to see a glimpse of what the future of the automotive industry might look like: self-driving cars that increase travel safety, reducing, if not nullifying, accidents. To achieve this ambitious goal, cars need to be aware of the environment that surrounds them in order to take the most appropriate action in each different situation.

Even though object detection/recognition based approaches [1] are very precise and reliable in some cases, they are not enough to accomplish such objective. A more profound comprehension of the scene is necessary if we want fine-graned decisions capabilities, e.g. deciding to go against a fence instead of a wall after a maneuver done to avoid a vehicle or a person.

Semantic Segmentation (SemSeg) [2] is a technology that, by classifying each individual pixel of the scene instead of just recognizing the main actors (such as vehicles and pedestrians), can enable driving systems to reach a better understanding of the whole view. Given the broad variety of driving conditions encountered in the real world, it is of paramount importance for these SemSeg algorithms to be able to generalize well and also cope with the inevitable domain shifts. On one side, this implies developing more effective domain adaptation (DA) techniques [3] that are able to cope with such a diversity of unpredictable scenarios. On the other, this requires datasets with a large amount of labeled data from a diverse set of conditions to support the training and evaluation of such techniques.

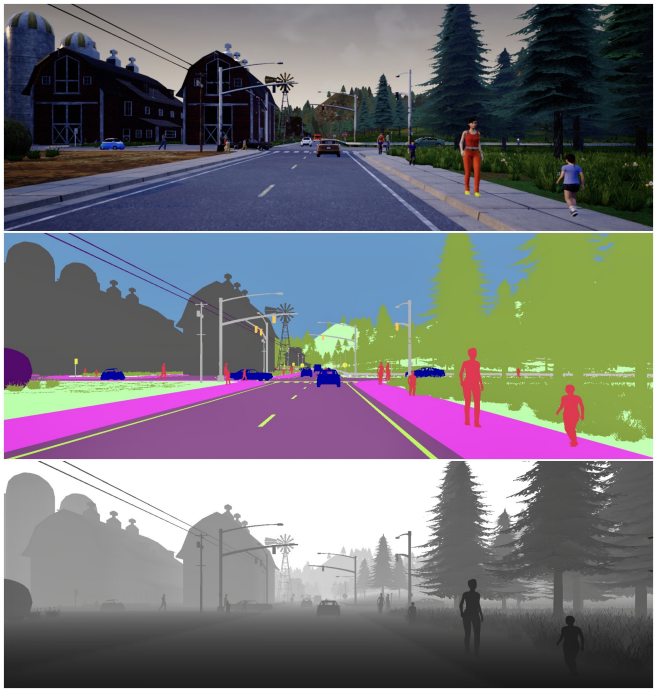

Fig. 1. The IDDA dataset. An example with an RGB image and its corresponding semantic and depth maps.

However, obtaining real labeled data in large quantities is far from trivial. Firstly, it is both arduous and costly to deploy multiple vehicles to collect images from a multitude of weather, lighting and environmental conditions.

Secondly, the task of manually classifying each image is excessively time-consuming, with a duration that can range from 60 to 90 minutes per image, as it was for the CamVid [4], [5] and Cityscapes [6] datasets respectively. Lastly, the accuracy of the manually produced labels might be inconsistent throughout the dataset.

All these reasons, together with the level of fidelity reached by 3D graphical engines, have fostered the creation and adoption of synthetic datasets for SemSeg [7], [8]. Furthermore, automatically producing the labels directly from

*The authors equally contributed.

[1]Emanuele Alberti, Antonio Tavera and Barbara Caputo are with Politecnico di Torino, Department of Control and Computer Engineering (DAUIN), Turin, Italy {emanuele.alberti, antonio.tavera, barbara.caputo}@polito.it

[2]Carlo Masone is with Italdesign Giugiaro S.p.A., Turin, Italy carlo.masone@italdesign.it

TABLE I

SUMMARY OF THE MOST POPULAR DATASET FOR SEMANTIC SEGMENTATION

				Semantic Segmentation						Data Variety			
Dataset	**Year**	**Size**	**Depth**	**Resolution (pixels)**	**FoV**	**#Classes**	**Annotation Time (min)**	**#Annotated Pixels (10^9)**	**#Weather Conditions**	**#Envs**	**#Viewpoints**	**#Selectable Domains**	**#images (avg*scene)**
Real-World Dataset													
CamVid	2008	701	No	920×720	-	32	60	0.62	1	1	1	1	-
KITTI	2012	400	**Yes**	1392×512	-	33	-	0.07	1	1	1	1	-
Cityscapes	2016	5k fine 20k coarse	No	2048×1024	90°	33	90 7	9.43 26.0	-	**50**	1	50	160
Mapillary Vistas	2017	25k	No	$\geq$ 1920×1080	-	**66**	94	-	-	-	-	1	-
BDD100K	2018	10k	No	1280×720	-	40	-	-	6	4	1	1	-
ApolloScape	2018	144k	**Yes**	**3384×2170**	-	25	-	-	-	1	1	3	**29k**
A2D2	2019	41k	No	1920×1280	**120°**	38	-	-	-	3	6 (different horizontal position)	23	1.7k
Synthetic Dataset													
Virtual KITTI	2016	21260	Yes	1242×375	29°	14	-	-	5	5	4 (different horizontal rotation)	50	426
Synthia-Rand Synthia-Seqs	2016	13,400 200k	**Yes**	960×720	100°	13	**Instant**	147.5	- **10**	4	**8** (different horizontal position)	1 51	- 8k
GTA V	2016	25k	No	1914×1052	-	19	7	50.15	-	1	-	1	-
IDDA	**2020**	**1M**	**Yes**	1920×1080	90°	24	**Instant**	**2087.70**	3	7	5 (different camera heights)	**105**	16k

the objects in the 3D engine allows to have perfect labeling and to easily add new classes. The downside of this approach is that models trained solely on virtual datasets have the tendency to perform very poorly in real case scenarios, suffering from the domain shift, even though ways to tackle these issues are being developed in the form of domain adaptation and generalization.

In our work we propose "IDDA" (ItalDesign DAtaset), a large synthetic dataset counting over one million labeled images as the sum of more than a hundred different scenarios over three axes of variability: 5 viewpoints, 7 towns and 3 weather conditions. The variety it offers allows for a deeper analysis and benchmarking of the performances of the current and future state-of-the-art SemSeg architectures, with a strong focus on DA tasks. For these reasons we believe that our dataset can bring a valuable contribution to the research community. The dataset, the experimental setups and all the algorithms used in this paper will be made publicly available through the dedicated webpage. The webpage will be periodically updated with new results and benchmark settings, with the explicit intention to make IDDA the reference resource for studying domain adaptive SemSeg in the automotive scenario.

To summarize, the main contributions of this paper are:

- the creation of the largest synthetic dataset for semantic segmentation currently available, featuring more than 1M images, more than 100 different combinations of scenarios, and fine pixel-wise semantic annotations and depth maps. The scenarios are well-divided using the three variability factors: weather condition, location and camera height.
- the evaluation of the performances of the current state-of-the-art SemSeg models with their DA variants, assessing how useful the dataset proves to be for benchmarking purposes, especially for a single-source DA task. We demonstrate how our dataset could potentially be employed to evaluate other tasks, such as multi-source DA or domain generalization.

II. RELATED WORK

The rapidly growing interest in the application of SemSeg to autonomous driving has led to the release of several datasets targeting this application (see Tab. I). Early datasets, such as CamVid [4], [5] and KITTI [9], while containing more than 30 classes of labeled objects, consisted of less than 1k semantically annotated images in low resolution and with little variability. The release of Cityscapes [6], with 5k finely annotated images and 20k coarsely annotated ones, led to the first benchmark to test SemSeg for autonomous driving.

The success of Cityscapes was later followed by the release of larger datasets from academic research (BDD100K [10]), image providers (Mapillary Vistas [11]) and automotive industry (Apolloscape [12], A2D2 [13]). Despite the availability of multiple datasets, none of these has yet provided a good benchmark to evaluate how well a SemSeg network performs when tested on a different domain. Some datasets, such as CamVid, KITTI or Apolloscape, simply lack variability since they contain images taken from a single city or point of view. Others, such as Mapillary Vistas and BDD100K, that offer scene diversity but lack a way to easily pick scenarios from different domains, or Virtual KITTI [14], that provides few images per scenario, make it hard to use them to evaluate DA approaches.

The problem of collecting and labeling large quantities of images with a rich diversity of conditions has led to the creation of datasets based on 3D games engines such as SYNTHIA [8] and GTA V [7]. Using data from game engines also allows to get finely annotated images without the cost of manual labeling. Unfortunately, even these two datasets have limitations for what concerns their use to evaluate DA. GTA V does not currently offer the possibility of picking scenes from different domains whereas SYNTHIA-Seqs only contains low resolution images and few labels.

In comparison to these prior datasets IDDA is designed to provide a benchmark to test not only the generalization capabilities of SemSeg architectures, but also to assess how

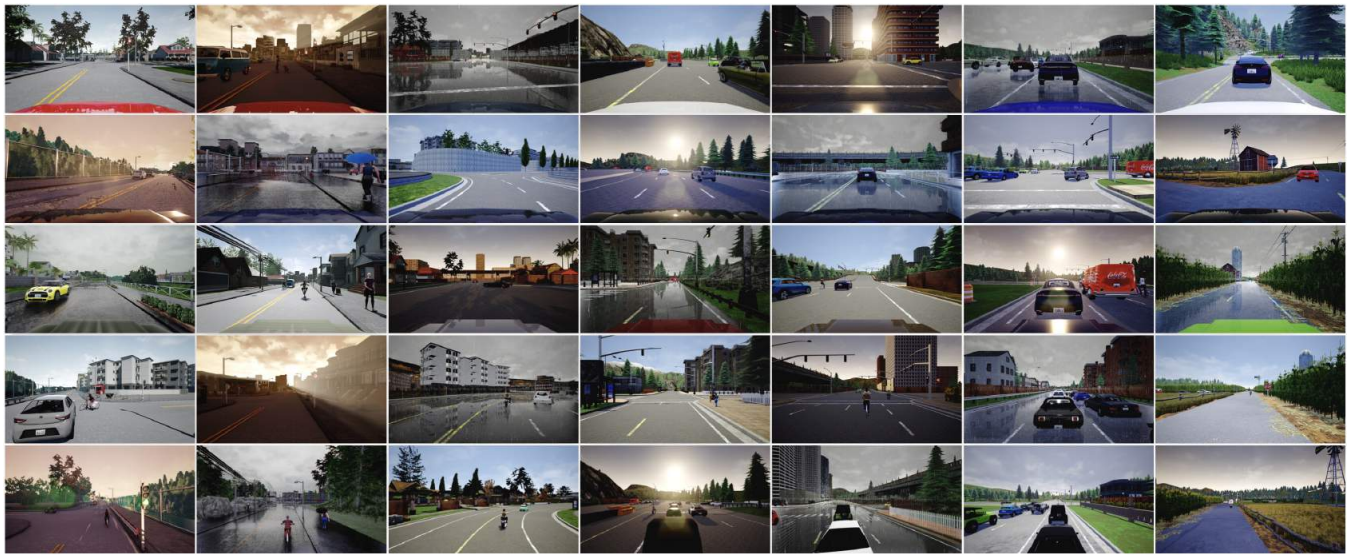

Fig. 2. Samples for any instance of variety provided by the IDDA dataset. On the row the 5 viewpoints (Audi, Mustang, Jeep, Volkswagen T2 and Bus), on the column the 7 environments (from Town1 to Town7). Images iterate over the 3 weather conditions (Clear Noon, Clear Sunset and Hard Rain Noon).

well they adapt to a domain shift. Our large-scale dataset consists of more than 1 FHD million images and it offers multiple domains easily and separately selectable. Together with each RGB image the dataset contains also its respective depth map and its high-quality semantic annotation for a total of 24 semantic classes, as shown in Fig. 1.

III. Data Creation

A. The virtual simulator

The simulator used for the generation of the dataset is CARLA [15] (version 0.8.4 and 0.9.6), an open-source project developed to support prototyping, training, and validation of autonomous driving systems. The motivation behind the choice of this simulator is the high degree of customization that it offers: the developer can set the number of pedestrians and vehicles, the environment conditions, the map and the speed of the simulation. Moreover, CARLA uses Unreal Engine 4 which is the current state-of-the-art in computer graphics. From a practical perspective, CARLA is based on a client-server architecture, where the client controls a chosen individual agent (*player*) while the server simulates the world and the remaining agents. This split allowed us to focus on implementing a custom made data-collection client without rewriting the server.

B. Data-Collection Client

Our client can start new simulations (*episodes*), defining each time the parameters and the meta-parameters. The number of frames captured by the player in each episode is limited by the client depending on the size of the town: the smaller the town the fewer the images (i.e. the shorter the episodes). Furthermore, to create different traffic scenarios, each episode is initialized with a random number of vehicles and pedestrians in the range of [20, 150] and [0, 100], respectively. Lastly, players are spawned in new locations and in each episode the distributions of the vehicle models and colors keep changing. These choices were made to limit the occurrence of deadlocks and, thus, the times in which the ego vehicle is stationary for any reason. Overall, these factors ensure that the collected data is rich and diverse.

The client also specifies the sensors equipped on the player vehicle. Out of all the sensors available in CARLA, for the creation of the dataset we used an RGB camera, a semantic segmentation sensor and a depth sensor, all with a field-of-view (FoV) of 90 degrees. The semantic segmentation sensor produces instantly pixel-wise labeled images directly from the blueprints of the objects in the Unreal Engine. The depth sensor provides images that codify depth in the 3 channels of the RGB color space, from the least to the most significant bytes: $R > G > B$.

The sensors are mounted coincidentally on the player's windshield, roughly at the height of the rear-view mirror. Since we used 5 different player vehicle models to collect the data (two sport cars, a jeep, a minivan and a bus), the camera height ranges between 1.2 and 2.5 meters. Additionally, the portion of the image occupied by the player's hood varies depending on the model of the vehicle, ranging from 11.08% to 13.99% when the hood is visible (sedans and jeep) and equaling 0% in the other cases.

All sensors are synchronized to capture a frame every 3 seconds, leading to episodes lasting from 3 to 4 minutes each (simulation time).

At the moment of capture, six frames are simultaneously stored: one RGB, three depth (raw, grayscale, and log-grayscale), two semantic (raw and colored using the Cityscapes color palette). For the RGB camera, post-processing effects such as bloom, lens flare and motion blur are applied in order to increase the realism of the images.

IV. The Dataset

IDDA (ItalDesign DAtaset) consists of 1,006,800 frames taken from the virtual world simulator CARLA. The creation

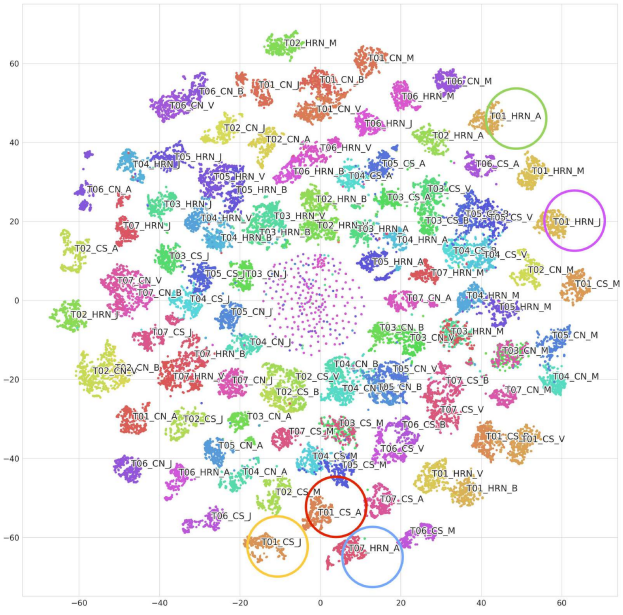

Fig. 3. The tSNE representation of the 105 different IDDA's scenarios. Inside the circles are the domains tested in Sec. V-A.

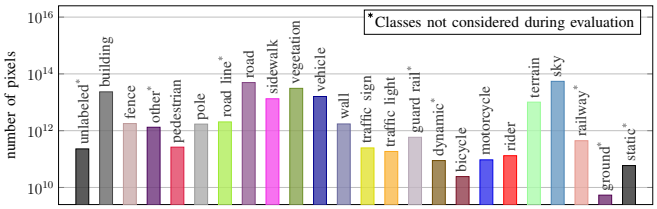

Fig. 4. Number of high-quality annotated pixels (y-axis) per class (x-axis).

of the dataset lasted about two weeks on two workstations, each equipped with a single NVIDIA Quadro P5000 GPU with 16GB of memory. In terms of quantity of frames, IDDA is 2 orders of magnitude larger than GTAV [7] and SYNTHIA [8] and 5 order of magnitude larger than semantically annotated images in KITTI [9]. Most importantly, IDDA features many scenarios spanning different cities, weather conditions and viewpoints, so as to support the development and evaluation of single or multi-source DA techniques applied to SemSeg.

A. Data Diversity

The 105 scenarios composing IDDA (examples in Fig. 2) are obtained by varying three aspects of the simulation.

Towns. The frames of the dataset are collected across seven different towns. Town1 (T01) and Town2 (T02) are characterized respectively by 2.9 km and 1.4 km of drivable roads with buildings, bridges, vegetation, terrain, traffic signs and various kinds of infrastructures. Town3 (T03), Town4 (T04), Town5 (T05) and Town6 (T06) are characterized by a complex urban scene with multi-lane roads, tunnels, roundabouts, freeways and connection ramps. Lastly, Town7 (T07) stands out from the rest because it depicts a bucolic countryside with narrow roads, fewer traffic lights and lots of non-signalized crossings. We believe that this entirely different domain is one important novelty provided by our dataset with regards to the autonomous driving task. All the seven cities are populated by vehicles and pedestrians.

Weather Conditions. We considered three weather settings that differ significantly from each other: Clear Noon (CN), characterized by bright daylight, Clear Sunset (CS), with the sun low above the horizon and pink/orange hues, and Hard Rain Noon (HRN), with a cloudy sky, intense rain and puddles that cause reflections on the floor.

Viewpoints. The third parameter that is varied to create the scenarios is the player vehicle. For each vehicle we positioned the sensor system approximately at the height of the rear-view mirror. We used five player vehicles that differ significantly in their height and shape, i.e, an Audi TT (A), a Ford Mustang (M), a Jeep Wrangler (J), a Volkswagen T2 (V) and a Bus (B). This choice guarantees not only that the resulting images have distinct perspectives, but also that the hood of the player vehicle, if visible[1], is dissimilar in both shape and color. To the best of our knowledge, the inclusion of images not only from the perspective of cars but also jeeps, vans and buses is a a unique feature of IDDA and it adds a whole new dimension of variability.

We use tSNE [16] to visually examine and evaluate the diversity of all the 105 available scenarios. To do so, we train a ResNet101 [17] from scratch, using 1000 samples from each scenario, with the sole task of classifying the domain of origin for each frame. Then, for each scenario, we compute its mean feature vector using 500 samples randomly taken from its validation set. Finally, we apply PCA, take the first 50 principal components and project them into a more intelligible 2D embedding. Fig. 3 presents a drawing of this embedding that intuitively shows the inherent domain shift that exists among the different scenarios. There is a strict correlation between the gap observed in the distribution of the domains in Fig. 3 and the results in terms of mIoU. Even if at a different scale, the similarity with Fig. 5, in which tSNE is computed only for handpicked sub-domains, is clearly discernible. In the experiments section we will demonstrate that this shift is strictly related to the results.

B. Semantic Segmentation

One of our goals in the creation of IDDA was to build a competitive dataset in terms of the range of recognizable items within a scene. In particular, we wanted to increase the default number of semantic classes provided by the simulator and get it as close as possible to the ones in Cityscapes or in GTA V. In order to achieve this result we made changes in the 3D maps themselves and we modified and rebuilt the source code of the simulator so that each static and dynamic element would be identified and tagged the moment before being spawned inside the virtual world. This strategy allowed us to increase the number of tags provided by the simulator

[1]The hood is not visible in the case of the Bus and the Volkswagen T2.

from the original 13 to a total number of 24 semantic classes. The distribution of classes in the IDDA dataset is analyzed in Fig. 4. It is clearly distinguishable that the predominant classes are building, road, vehicle, vegetation, terrain and sky. Other useful statistics are synthesized in the Tab. I.

V. Experiments

We demonstrate the main features and potential applications of IDDA with two experiments. In the first one we want to verify that the scenarios available in IDDA are an effective tool to validate and benchmark how well SemSeg methods can adapt to domain shifts in driving applications. To do so, we selected several state-of-the-art networks, both with and without DA, and we looked at the performance degradation when the train and test sets are taken from different scenarios. With the second experiment we use the scenarios available in IDDA to investigate how different data distributions in the synthetic source domain affect the performance of a network on a real target domain. For this purpose we use the same networks from the first experiment but test them on Cityscapes, BDD100K, Mapillary Vistas and A2D2.

Evaluated methods. For the experiments we use eight state-of-the-art SemSeg architectures. Four of these networks do not implement DA, i.e. PSPNet [18], that introduces a Pyramid Scene Parsing module, PSANet [19], that proposes a point-wise spatial attention network to gather information from all the positions in the feature maps and DeepLab V3+ [20], that implements the Atrous Spatial Pyramid Pooling module. The fourth SemSeg architecture included in our experiments is DeepLab V2 [21] with a ResNet-101 [17] as backbone, because this is the main building block for all the chosen DA methods.

The remaining four architectures are some of the best performing unsupervised DA models: ADVENT [22], DISE [23], CLAN [24], and DADA [25]. Each approach achieves its goal in a different way with respect to the others: both ADVENT and DADA use an entropy minimization technique with the help of an adversarial task, but the latter also takes advantage of depth information, DISE unravels images into domain-invariant structure and domain-specific texture representations, allowing for label transferring, and CLAN takes into account the local semantic consistency when pursuing the global alignment of the distributions, reducing the negative transfer side effect, that is the misalignment of features that were already aligned well prior to the mapping.

Experimental setup. For each network we used the hyperparameters reported in its original paper, so as to obtain a fair evaluation of the performances. For all the DA architectures the official implementation provided by the authors is used, whereas for the SemSeg-only part of the experiments re-implementations in PyTorch are used.

To better compare with Cityscapes, since it is the main real dataset for benchmarking SemSeg for autonomous driving, we excluded from our experiments those classes that were either ambiguous (dynamic, static, other) or not present in the reference dataset (road line). We ended up considering the 16 labels in Fig. 4.

TABLE II
Scenarios Distances

	Case		
Distance Function	Viewpoint Change	Weather Change	City Change
Euclidean	2.7604	5.6555	6.4551
Cosine	0.2590	1.2633	1.0586
Bhattacharyya	0.0149	0.0337	0.0426

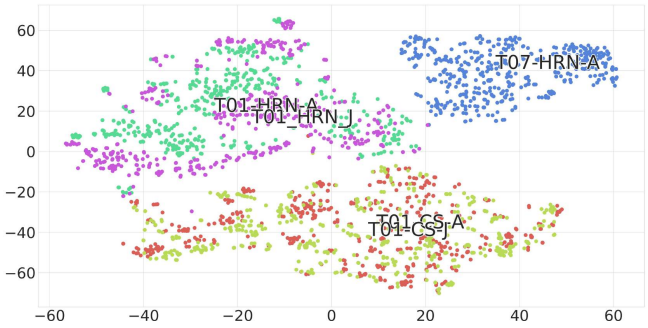

Fig. 5. The tSNE representation of the 5 chosen scenarios to assess IDDA.

To quantify the distance between source and target domain (similar to 3.1 in [26]) we extract, using ResNet-101 [17] pretrained on ImageNet, the features of the first 500 samples of each domain and we reduce the dimensionality (using PCA) taking the first 50 principal components. Then we proceed in two directions: in one case we compute the mean-feature vector for each domain and we measure the Euclidean and the Cosine distances, in the other case we compute the feature-wise Bhattacharyya distance.

Lastly, in all the experiments the performance is measured using the mean Intersection over Union (mIoU) metric.

A. Assessing IDDA

We test the ability of the selected networks to adapt to a new domain by considering three cases that cover the three variability factors:

- the first, tests the viewpoint change (from A as source to J as target), fixing background and weather (T01, CS);
- the second, tests the weather shifting (from CS as source to HRN as target), fixing viewpoint and background (J, T01)
- the third, considers two scenarios that take place in different environments (T01 as source and T07 as target), fixing viewpoint and weather (A, HRN);

We used the method detailed in the section V to measure numerically the distance and the difficulty of the three cases (see Tab. II). As a visual confirmation, we use tSNE to project the features extracted with the ResNet-101 into a more comprehensible 2D space (see Fig. 5).

The results of the experiments are reported in Tab. III. As expected from the distance computation, the shift across cities, made even more challenging by the rainy condition, produces a higher degradation in performance than the other two experiments. In the town shift the SemSeg networks

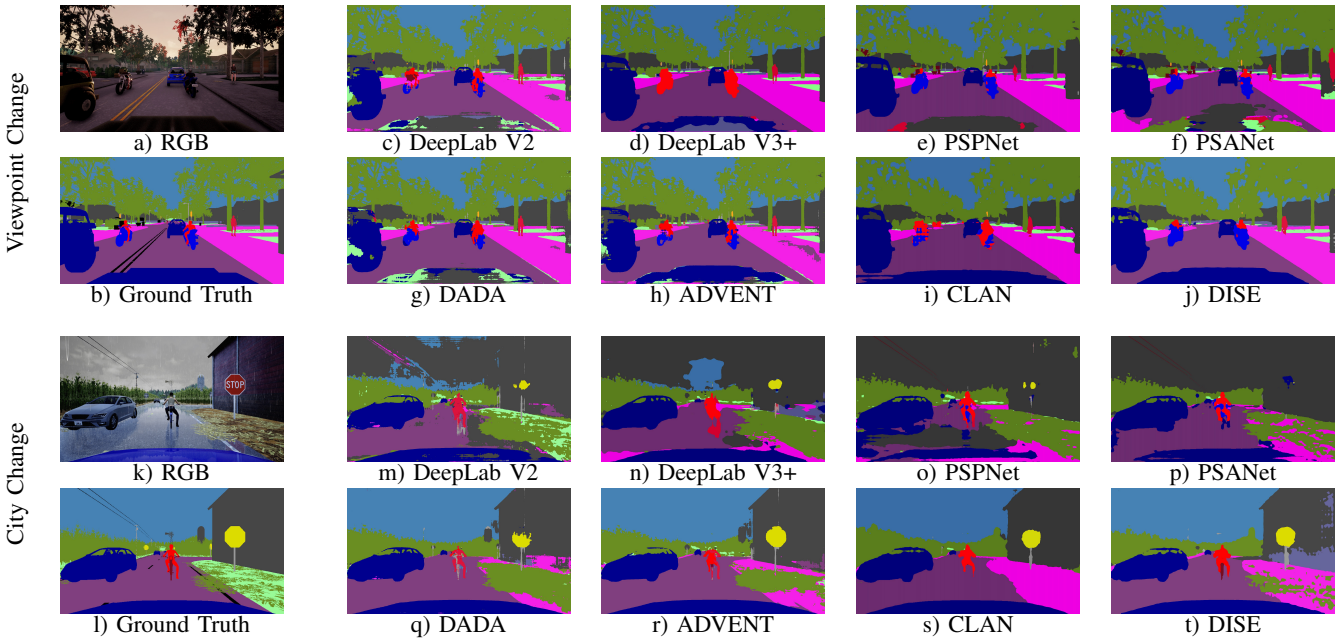

Fig. 6. Qualitative results for the viewpoint and background change experiment. Note the more severe side effect caused by the domain shift across different cities.

TABLE III
ASSESSING IDDA EXPERIMENT RESULTS

Semantic Segmentation Networks		Scenarios (% mIoU)		
		Viewpoint Change	Weather Change	City Change
	Source:	T01 CS A	T01 CS J	T01 HRN A
	Target:	T01 CS J	T01 HRN J	T07 HRN A
w/o DA	DeepLab V2	62.60	40.24	21.65
	DeepLab V3+	64.93	33.93	14.27
	PSPNet	67.32	29.65	14.64
	PSANet	66.88	33.60	15.52
	DeepLab V2 (source=target)	79.13	78.31	78.02
w/ DA	DADA	66.42	55.87	36.48
	ADVENT	68.43	61.13	39.30
	CLAN	70.30	65.52	41.18
	DISE	73.64	71.91	46.71

struggle to correctly classify the scene and their accuracy drops as low as 14%. In this case DA produces a considerable boost with an averaged accuracy of 40%. This trend is repeated within the shift across weathers but, since the gap among source and target domain is smaller, the resulting average mIoU is of 34%. In this case DA performs quite well, giving as outcome an averaged 63%. Lastly, the viewpoint change proves to be the best performing set of experiments, so the addition of DA increases the average accuracy by only 4%. Among all of the DA networks, DISE proves to be the most capable while the depth information exploited by DADA does not seem to improve the performance.

Fig. 6 illustrates some qualitative results of our experiments. Looking at the output produced we can highlight two interesting problems that seem to affect the SemSeg networks and their generalization capability. Considering the viewpoint change, all the SemSeg models without DA struggle to classify well the portion of the image occupied by the hood of the vehicle, improperly classifying it as a building. Moreover, when changing the scenario and moving to a countryside scene with vegetation in place of roadside and sidewalks ("city change" case), we observe that, during training, all the networks (with the only exception of DeepLab V2) learned and memorized the pattern "building-sidewalk-road" of the source scenario. Therefore, when moving to the target environment they are not able to adapt and tend to incorrectly classify the terrain as sidewalk.

The diversity of our dataset and the possibility to simulate various kind of real scenario has made it possible to gain this kind of insight. Furthermore, we have demonstrated the limitations of the actual state-of-the-art SemSeg networks and how IDDA could be a powerful tool to validate the adaptation performances to a domain shift in driving applications.

B. Synthetic vs. real scenarios

In the second experiment we test how well the networks trained on a synthetic dataset can adapt to a real one. In particular, we consider two cases, each using a source domain obtained from a combination of several scenarios in IDDA:

- the first, called "best case", is a mixture of samples with similar environmental conditions to the target domains, counting a total of 29,952 elements sampled in a stratified fashion and taken only from urban environments (T01-T06), with a car-like point-of-view (A or M) and clear weather conditions at noon (CN);
- the second, called "worst case", has a higher visual discrepancy w.r.t. the target samples and it counts 40,128 samples taken from the previously excluded countryside town (T07), with a hooded and a non-hooded point of views (J and B) and rainy conditions at noon (HRN).

Tab. IV and Fig. 7 show numerically and visually the distance among the dataset distributions. When evaluating

TABLE IV

DISTANCES BETWEEN IDDA AND REAL DATASETS

	Distance Function	Dataset			
		Cityscapes	BDD100K	Mapillary	A2D2
Best Case	Euclidean	7.4419	7.6177	5.4493	6.3874
	Cosine	1.3582	1.6209	1.2924	1.0589
	Bhattacharyya	0.0552	0.0502	0.0106	0.0447
Worst Case	Euclidean	8.2360	7.7618	4.9548	7.0150
	Cosine	1.5465	1.5526	0.9147	1.1849
	Bhattacharyya	0.0498	0.0381	0.0267	0.0387

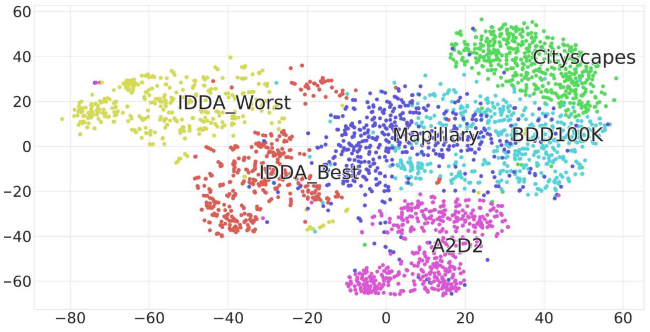

Fig. 7. The tSNE representation of the distributions of synthetic and real datasets.

the performance on the target datasets, we ignored all labels not included in Fig. 4 and labeled all four-wheeled vehicles as our semantic class "vehicle". For A2D2 we considered only 13 labels due to its labeling inconsistencies with IDDA, e.g. the complete absence of class "rider" and "wall" and the union of "vegetation" with "terrain". Results are in Tab. V. With the SemSeg-only architectures we can measure a drop in performance of 30.97% on average in the best case (not considering the A2D2 experiments since a fair comparison in terms of mIoU cannot be done due to the different evaluation setup). As it can be seen in Fig. 8c (best case), the network struggles to disambiguate between building, road and sidewalk, though it does an acceptable job at recognizing pedestrians. Among the DA approaches, DISE proves to be the most effective. Nonetheless, the gap with the baseline is still remarkable and the improvements introduced by DA are not enough to guarantee acceptable performance. Interestingly, it seems that the additional depth information exploited by DADA is helpful only in Mapillary Vistas.

As expected, in the worst case the domain shift is much more severe, with a maximum drop of 46.08% when tested on Cityscapes. In this case, the SemSeg-only network fails to even identify the road, confusing it with the "terrain" (see Fig. 8c, worst case). This can be imputable to the relevant textural differences of source and target domains. The impact of DA is visually high, yet numerically we observe how even the best performing architecture does not get close to the baseline. We also note that in both Cityscapes and

TABLE V

SYNTHETIC VS REAL EXPERIMENT RESULTS

*CONSIDERING ONLY 13 LABELS

Source	Networks	Target			
		Cityscapes	BDD100K	Mapillary Vistas	A2D2*
Same as target (baseline)	DeepLab V2	62.89	52.71	67.63	65.43
Best case	DeepLab V2	32.66	24.18	36.09	32.10
	DADA	33.13	29.58	37.29	38.57
	ADVENT	35.32	33.18	36.97	42.56
	CLAN	39.26	33.47	39.42	44.31
	DISE	42.07	40.09	41.70	46.73
Worst case	DeepLab V2	16.81	17.48	27.09	29.80
	DADA	23.68	23.45	32.57	36.18
	ADVENT	23.83	27.04	30.26	38.57
	CLAN	25.75	30.70	30.88	42.71
	DISE	31.25	31.37	33.72	45.49

BDD100K the best performing DA (DISE) almost doubles the performances of the SemSeg-only architecture, but has a much lower increase of performance in the case of Mapillary. This suggests that the higher the performance of the SemSeg-only networks, the lower the impact of DA. Overall the gap remains of 28.96% on average, showing once again how DA techniques still have to work to achieve adequate results.

Looking at the A2D2 results, the SemSeg-only architecture shows a drop in performance close to 30% both in the Best and Worst cases. The domain shift in the Worst case is a little less severe than Cityscapes and BDD100K, and more similar to Mapillary. This can be imputed to an higher presence of roads out of town, decreasing the difference among A2D2 and the Worst case distribution.

VI. CONCLUSION

This paper presents IDDA, a synthetic database explicitly designed for supporting research in domain adaptive semantic segmentation for autonomous driving. With 105 different domains, it is the largest existing dataset supporting this research. As shown in the experiment section, it lends itself well to the benchmark of wide range of domain adaptation study cases, due to the domain gap that exists both among scenarios inside IDDA and with respect to a real dataset. Furthermore, the constant development of the simulator allows for a further expansion of the currently available scenarios, for instance by adding a night view, new environments or sensor types (such as LIDAR).

REFERENCES

[1] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128 837–128 868, 2019.

[2] I. Ulku and E. Akagunduz, "A survey on deep learning-based architectures for semantic segmentation on 2d images," 2019.

[3] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135 – 153, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231218306684

[4] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *ECCV (1)*, 2008, pp. 44–57.

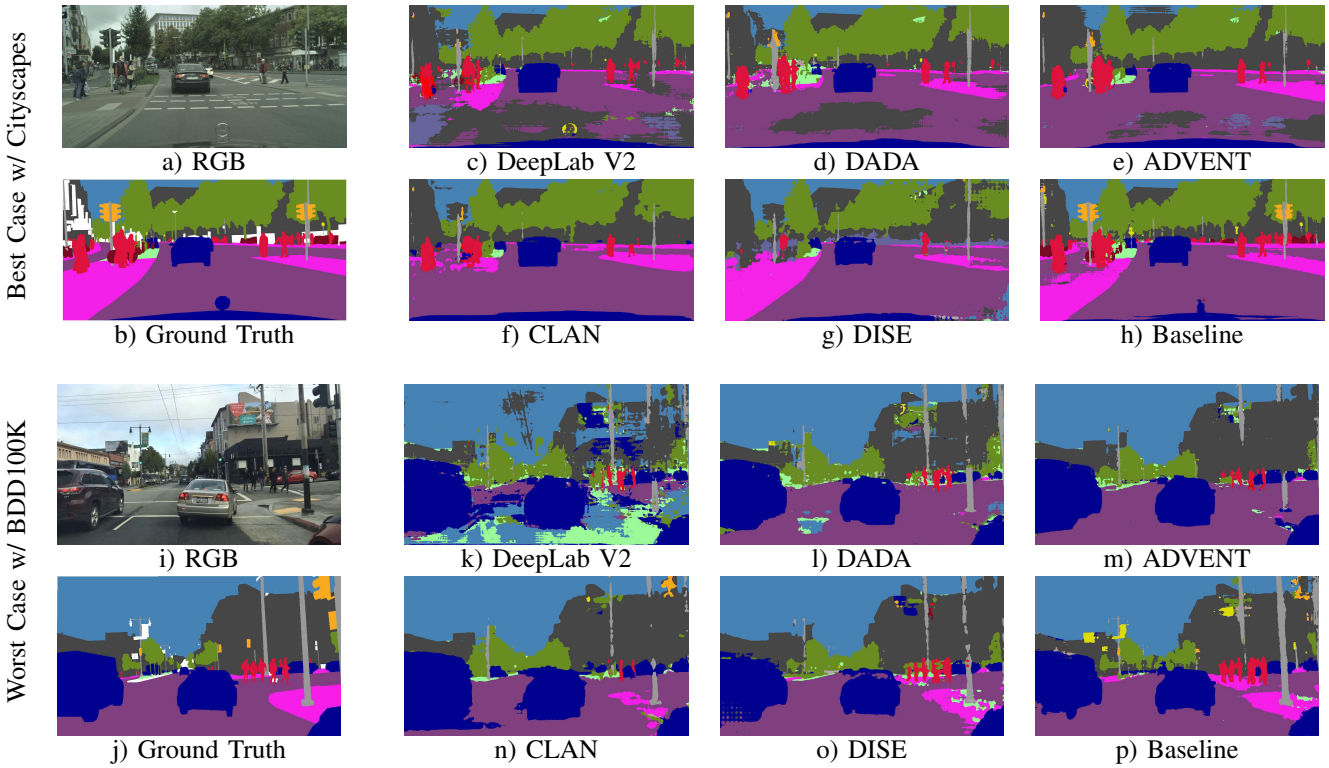

Fig. 8. Examples of the results when training on the best and worst case scenarios of IDDA and testing on real datasets. Baseline refers to DeepLab V2 trained on the target (real) domain.

[5] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, pp. 88–97, 2008.

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[7] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *European Conference on Computer Vision (ECCV)*, ser. LNCS, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer International Publishing, 2016, pp. 102–118.

[8] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez, "The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *CVPR*, 2016.

[9] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[10] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *International Conference on Computer Vision (ICCV)*, 2017. [Online]. Available: https://www.mapillary.com/dataset/vistas

[12] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[13] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. H. Pham, M. Mühlegg, S. Dorn, T. Fernandez, M. Jänicke, S. Mirashi, C. Savani, M. Sturm, O. Vorobiov, M. Oelker, S. Garreis, and P. Schuberth, "A2D2: Audi Autonomous Driving Dataset," 2020. [Online]. Available: https://www.a2d2.audi

[14] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[15] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[16] L. van der Maaten and G. Hinton, "Viualizing data using t-sne," *JMLR*, vol. 9, pp. 2579–2605, 11 2008.

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

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017.

[19] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *ECCV*, 2018, pp. 267–283.

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, p. 834—848, April 2018. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2699184

[22] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *CVPR*, 2019.

[23] W.-L. Chang, H.-P. Wang, W.-H. Peng, and W.-C. Chiu, "All about structure: Adapting structural information across domains for boosting semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[24] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Dada: Depth-aware domain adaptation in semantic segmentation," in *ICCV*, 2019.

[26] B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, and T. Darrell, "Best practices for fine-tuning visual classifiers to new domains," in *ECCV Workshops*, 2016, pp. 435–442.