Plugging Self-Supervised Monocular Depth Into Unsupervised Domain Adaptation for Semantic Segmentation

Plugging Self-Supervised Monocular Depth into Unsupervised Domain Adaptation for Semantic Segmentation

Adriano Cardace Luca De Luigi Pierluigi Zama Ramirez Samuele Salti Luigi Di Stefano
Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
{adriano.cardace2, luca.deluigi4, pierluigi.zama}@unibo.it

Abstract

Although recent semantic segmentation methods have made remarkable progress, they still rely on large amounts of annotated training data, which are often infeasible to collect in the autonomous driving scenario. Previous works usually tackle this issue with Unsupervised Domain Adaptation (UDA), which entails training a network on synthetic images and applying the model to real ones while minimizing the discrepancy between the two domains. Yet, these techniques do not consider additional information that may be obtained from other tasks. Differently, we propose to exploit self-supervised monocular depth estimation to improve UDA for semantic segmentation. On one hand, we deploy depth to realize a plug-in component which can inject complementary geometric cues into any existing UDA method. We further rely on depth to generate a large and varied set of samples to Self-Train the final model. Our whole proposal allows for achieving state-of-the-art performance (58.8 mIoU) in the GTA5→CS benchmark. Code is available at https://github.com/CVLAB-Unibo/d4-dbst.

1. Introduction

Semantic segmentation is the task of classifying each pixel of an image. Nowadays, Convolutional Neural Networks can achieve impressive results in this task but require huge quantities of labelled images at training time [44, 3, 34, 41]. A popular trend to address this issue concerns leveraging on computer graphics simulations [42] or game engines [40] to obtain automatically synthetic images endowed with per-pixel semantic labels. Yet, a network trained on synthetic data only will perform poorly in real environments due to the so called *domain-shift* problem. In the last few years, many Unsupervised Domain Adaptation (UDA) techniques aimed at alleviating the domain-shift problem have been proposed in literature. These approaches try to minimize the gap between the labeled source domain (e.g. synthetic images) and the unlabeled target domain (e.g. real images) by either hallucinating input images, manipulating the learned features space or imposing statistical constraints on the predictions [58, 8, 65, 18].

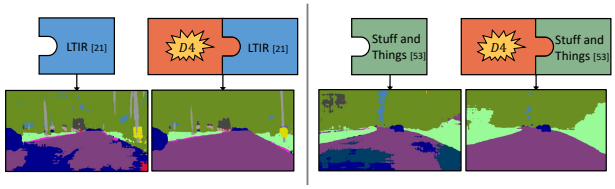

Figure 1. D4 can be plugged seamlessly into any existing method to improve UDA for Semantic Segmentation. Here we show how the introduction of D4 can ameliorate the performance of two recent methods like LTIR [22] and Stuff and Things [55].

At a more abstract level, UDA may be thought of as the process of transferring more effectively to the target domain the knowledge from a task solved in the source domain. This suggests that it may be possible to improve UDA by transferring also knowledge learned from *another task* to improve performance in the real domain. In fact, the existence of tightly related representations within CNNs trained for different tasks has been highlighted since the early works in the field [60], and it is nowadays standard practice to initialize CNNs deployed for a variety of diverse tasks, such as, e.g., object detection [46], semantic segmentation [4] and monocular depth estimation [14], with weights learned on Imagenet Classification [11]. The notion of *transferability* of representations among CNNs trained to solve different visual tasks has been formalized computationally by the Taskonomy proposed in [63]. Later, [38] has shown that it is possible to train a CNN to hallucinate deep features learned to address one task into features amenable to another task related to the former.

Inspired by these findings, we argue that monocular depth estimation could be an excellent task in order to gather additional knowledge useful to address semantic seg-

mentation in UDA settings. First of all, a monocular depth estimation network makes predictions based on 3D cues dealing with the appearance, shape, relative sizes and spatial relationships of the stuff and things observed in the training images. This suggests that the network has to predict geometry by implicitly learning to understand the semantics of the scene. Indeed, [37, 21, 24, 15] show that a monocular depth estimation network obtains better performances if forced to learn jointly a semantic segmentation task. We argue, though, the correlation between geometry and semantics to hold bidirectionally, such that a semantic segmentation network may obtain useful hints from depth information. This intuition is supported by [38], which shows that it is possible to learn a mapping in both directions between features learned to predict depth and per-pixel semantic labels. It is also worth observing how depth prediction networks tend to extract accurate information for regions characterized by repeatable and simple geometries, such as roads and buildings, which feature strong spatial and geometric priors (e.g. the road is typically a plane in the bottom part of the image) [13, 14, 47, 57]. Therefore, on one hand predicting accurately the semantics of such regions from depth information alone should be possible. On the other, a semantic network capable of reasoning on the geometry of the scene should be less prone to mistakes caused by appearance variations between synthetic and real images, the key issue in UDA for semantic segmentation.

Despite the above observations, injection of geometric cues into UDA frameworks for semantic segmentation has been largely unexplored in literature, with the exception of a few proposals, which either assume availability of depth labels in the real domain [56], a very restrictive assumption, or can leverage on depth information only in the synthetic domain due to availability of cheap labels [53, 27, 6]. In this respect, we set forth an additional consideration: nowadays effective self-supervised procedures allow for training a monocular depth estimation network without the need of ground-truth labels [14, 12, 70].

Based on the above intuitions and considerations, in this paper we propose the first approach that, thanks to self-supervision, allows for deploying depth information from both synthetic and *unlabelled real* images in order to inject geometric cues in UDA for semantic segmentation. Purposely, we adapt the knowledge learned to pursue depth estimation into a representation amenable to semantic segmentation by the feature transfer architecture proposed in [38]. As the geometric cues learned from monocular images yield semantic predictions that are often complementary to those attainable by current UDA methods, we realize our proposal as a depth-based add-on, dubbed D4 (Depth For), which can be plugged seamlessly into any UDA method to boost its performances, as illustrated in Fig. 1.

A recent trend in UDA for semantic segmentation is Self-Training (ST), which consists in further fine-tuning the trained network by its own predictions [72, 73, 68, 29, 33, 30]. We propose a novel Depth-Based Self-Training (DBST) approach which deploys once more the availability of depth information for real images in order to build a large and varied dataset of plausible samples to be deployed in the Self-Training (ST) procedure[1].

Our framework can improve many state-of-the-art methods by a large margin in two UDA for semantic segmentation benchmarks, where networks are trained either on GTA5 [40] or SYNTHIA VIDEO SEQUENCES [42] and tested on Cityscapes [10]. Moreover, we show that our DBST procedure enables to distill the whole framework into a single ResNet101 [16] and achieve state-of-the-art performance. Our contributions can be summarized as follows:

- We are the first to show how to exploit self-supervised monocular depth estimation on real images to pursue semantic segmentation in a domain adaptation settings.
- We propose a depth-based module (D4) which can be plugged into any UDA for semantic segmentation method to boost performance.
- We introduce a new protocol (DBST) that exploits depth predictions to synthesize augmented training samples for the final self-training step deployed oftentimes in UDA for semantic segmentation pipelines.
- We show that leveraging on both D4 and DBST allows for achieving 58.8 mIoU in the popular GTA5→CS UDA benchmark, i.e., to the best of our knowledge, the new state-of-the-art.

2. Related Work

Domain Adaptation. Domain Adaptation is a promising way of solving semantic segmentation without annotations. Pioneering works [17, 58, 2, 9, 31, 66, 62, 28, 22] rely on CycleGANs [71] to convert source data into the style of target data, reducing the low-level visual appearance discrepancy among domains. Other works exploit adversarial training to enforce domain alignment [49, 50, 54, 67, 59, 36, 1, 52]. [55] extended this idea by aligning differently objects with low and high variability in terms of appearance. Few works tried to exploit depth information to boost UDA for semantic segmentation. [53], for example, proposes a unified depth-aware UDA framework that leverages the knowledge of depth maps in the source domain to perform feature space alignment. [43] extends this idea by modelling explicitly the relation between different visual semantic classes and depth ranges. [7], instead, considers depth as a way to obtain adaptation at both the input and output level. [56] is the first work to consider depth

[1] See also [19] for concurrent work that proposes a similar idea.

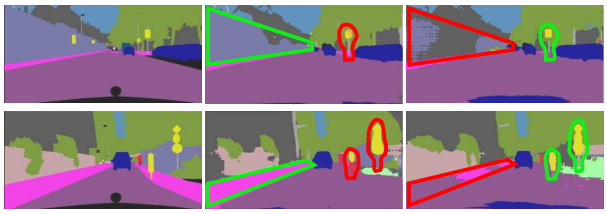

Figure 2. From left to right: ground truth, semantics from depth, semantics by LTIR [22]. The semantic labels predicted from depth are more accurate than those yielded by UDA methods in regularly-shaped objects (such as the *wall* in the top image and the *sidewalk* in the bottom one), whilst UDA approaches tend to perform better on small objects (see the *traffic signs* in both images).

in the target domain, although assuming supervision to be available. Conversely, we show how to deploy depth in the target domain without availability of ground-truth depths.

Self-Training. More recently, a new line of research focuses on self-training [26], where a semantic classifier is fine-tuned directly on the target domain, using its own predictions as pseudo-labels. [72, 73, 30] cleverly set class-confidence thresholds to mask wrong predictions. [69, 33, 68] propose to use pseudo-labels with different regularization techniques to minimize both the inter-domain and intra-domain gap. [64] instead, estimates the likelihood of pseudo-labels to perform online correction and denoising during training. Differentely, [48] synthesizes new samples for the target domain by cropping objects from source images using ground truth labels and pasting them onto target images. Inspired by this work, we propose a novel algorithm for generating new samples to perform self-training on the target domain. In contrast to [48], our strategy is applied to target images only and relies on the availability of depth maps obtained through self-supervision.

Task Adaptation. All existing approaches tackle independently task adaptation or domain adaptation. [51] was the first paper to propose a cross-tasks and cross-domains adaptation approach, considering two image classification problems as different tasks. UM-Adapt [25] employs a cross-task distillation module to force inter-task coherency. Differently, [38], directly exploits the relationship among tasks to reduce the need for labelled data. This is done by learning a mapping function in feature space between two networks trained independently for two separate tasks, a pretext and target one. We leverage on this intuition but, unlike [38], our approach does not require supervision to solve the pretext task in the target domain.

3. Method

In Unsupervised Domain Adaptation (UDA) for semantic segmentation one wishes to solve semantic segmentation in a target domain, $\mathcal{D}_\mathcal{T}$, though labels are available only in another domain, referred to as source domain $\mathcal{D}_\mathcal{S}$. In the following we describe the two ingredients of our proposal to better tackle this problem. In Sec. 3.1 we show how to transfer information from self-supervised monocular depth to semantic segmentation and merge this knowledge with any UDA method (D4-UDA, Depth For UDA). Then, in Sec. 3.2 we introduce a Depth-Based Self-Training strategy (DBST) to further improve semantic predictions while distilling the whole framework into a single CNN.

3.1. D4 (Depth For UDA)

Semantics from Depth. The main intuition behind our work is that semantic segmentation masks obtained exploiting depth information have peculiar properties that make them suitable to improve segmentation masks obtained with standard UDA methods. However, predicting semantics from depth is an arduous task. Indeed, we experiment several alternatives (see Sec. 4.4 *Alternative strategies to exploit depth*) and find out that the most effective way is a procedure similar to the one proposed in [38], which we adapt to the UDA scenario. The pipeline works as follows: train one CNN to solve a first task on $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$, train another CNN to solve a second task on $\mathcal{D}_\mathcal{S}$ only (i.e. the only domain where ground truth labels for the second task are available) and, finally, train a *transfer* function to map deep features extracted by the first CNN into deep features amenable to the second one. As the second CNN has been trained only on $\mathcal{D}_\mathcal{S}$, also the transfer function can be trained only on $\mathcal{D}_\mathcal{S}$ but, interestingly, it can generalize to $\mathcal{D}_\mathcal{T}$. As a consequence, at inference time one can solve the second task in $\mathcal{D}_\mathcal{T}$ based on the features transferred from the first task. We refer to [38] for further details.

Hence, if we assume the first and second task to consist in depth estimation and semantic segmentation, respectively, the idea of transferring features might be deployed in a UDA scenario since it gives the possibility to solve the second task on $\mathcal{D}_\mathcal{T}$ without the need of ground truth labels. However, the learning framework delineated in [38] assumes availability of ground-truth labels for the first task (depth estimation in our setting) also in $\mathcal{D}_\mathcal{T}$ (real images). As pointed out in Sec. 1, this assumption does not comply with the standard UDA for semantic segmentation problem formulation, which requires availability of semantic labels for source images ($\mathcal{D}_\mathcal{S}$) alongside with unlabelled target images only ($\mathcal{D}_\mathcal{T}$). To address this issue we propose to rely on *depth proxy-labels* attainable from images belonging to both $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$ without the need of any ground-truth information. In particular, we propose to deploy one of the recently proposed deep neural networks, such as [14], that can be trained to perform monocular depth estimation based on a self-supervised loss that requires availability of raw image sequences only, i.e. without ground-truth depth labels. Thus, in our method we introduce the following protocol.

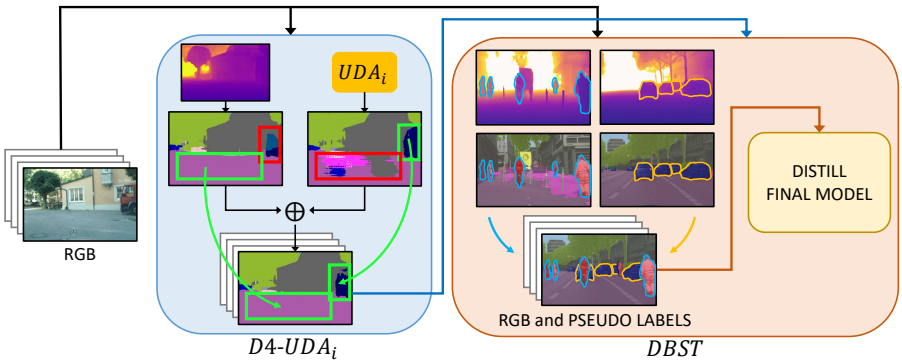

Figure 3. Overview of our proposal. RGB images are first processed by two different segmentation engines to produce complementary predictions that are then combined by a weighted sum which accounts for the relative strengths of the two engines (Eq. 3). During the next step (DBST), predictions from D4-UDA$_i$ are used to synthesize augmented samples by mixing portions of different images according to depth and semantics. The augmented samples are used to train a final model, so as to distill the whole pipeline into a single network.

First, we train a self-supervised monocular depth estimation network on both $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$. Then, we use this network to generate *depth proxy-labels* for both domains. We point out that we use such network as an off-the-shelf algorithm without the aim of improving depth estimation. Finally, according to [38], we train a first CNN to predict depth from images on both domains by the previously computed *depth proxy-labels*, a second CNN to predict semantic labels on $\mathcal{D}_\mathcal{S}$ and a transfer network which allows for predicting semantic labels from depth features in $\mathcal{D}_\mathcal{T}$. In the following, we will refer to such predictions as "semantics from depth" because they concern semantic information extracted from features amenable to perform monocular depth estimation.

Combine with UDA. Fig. 2 compares semantic predictions obtained from depth by the protocol described in the previous sub-section and from a recent UDA method. The reader may observe a clear pattern: predictions from depth tend to be smoother and more accurate on objects with large and regular shapes, like *road*, *sidewalk*, *wall* and *building*. However, they turn out often imprecise in regions where depth predictions are less informative, like thin things partially overlapped with other objects or fine-grained structures in the background. As UDA methods tend to perform better on such classes (see Fig. 2), our D4 approach is designed to *combine* the semantic knowledge extracted from depth with that provided by any chosen UDA method in order to achieve more accurate semantic predictions. Depth information helps on large objects with regular shapes, which usually account for the majority of pixels in an image. On the contrary, UDA methods perform well in predicting semantic labels for categories that typically concern much smaller fractions of the total number of pixel in an image, like e.g. the *traffic signs* in Fig. 2. This orthogonality suggests that a simple yet effective way to combine the semantic knowledge drawn from depth with that provided by UDA methods consists in a weighted sum of predictions, with weights computed according to the frequency of classes in $\mathcal{D}_\mathcal{S}$ (the domain where semantic labels are available). As weights given to UDA predictions ($\mathbf{w}_{uda}$) should be larger for rarer classes, they can be computed as:

$$\mathbf{w}_{uda} = [w^1_{uda}, \dots, w^C_{uda}] \quad \text{where} \quad w^i_{uda} = \frac{1}{ln(\delta + f^i)} \tag{1}$$

where C denotes the number of classes and $f^i = \frac{n^i}{n^{tot}}$ denotes their frequencies at the pixel level, *i.e.* the ratio between the number n^i of pixels labelled with class i in $\mathcal{D}_\mathcal{S}$ and the total number n^{tot} of labelled pixels in $\mathcal{D}_\mathcal{S}$. Eq. 1 is the standard formulation introduced in [34] to compute bounded weights inversely proportional to the frequency of classes. We set δ in Eq. 1 to 1.02, akin to [34].

Accordingly, weights applied to semantic predictions drawn from depth ($\mathbf{w}_{dep}$) are given by:

$$\mathbf{w}_{dep} = [w^1_{dep}, ..., w^C_{dep}] \quad \text{where} \quad w^i_{dep} = 1 - w^i_{uda}. \tag{2}$$

Thus, at each pixel of a given image we propose to combine semantics from depth and predictions yielded by any chosen UDA method as follows:

$$\widehat{\mathbf{y}}_f = \mathbf{w}_{dep} \cdot \phi_T(\tilde{\mathbf{y}}_{dep}) + \mathbf{w}_{uda} \cdot \phi_T(\tilde{\mathbf{y}}_{uda}), \tag{3}$$

where $\widehat{\mathbf{y}}_f$ is the final prediction, $\tilde{\mathbf{y}}_{dep}$ and $\tilde{\mathbf{y}}_{uda}$ are the logits associated with semantics from depth and the selected UDA method, respectively, ϕ_T denotes the *softmax* function with a temperature term T that we set to 6 in our experiments.

As illustrated in Fig. 3, the formulation presented in Eq. 3 and symbolized as $\bigoplus$ can be used seamlessly to plug semantics obtained from self-supervised monocular depth into any existing UDA method. We will refer to the combination of a given UDA method with our D4 with the expression D4-UDA. Experimental results (Sec. 4.3) show that all recent s.o.t.a. UDA methods do benefit significantly from the complementary geometric cues brought in by D4.

Figure 4. The rightmost column is synthesized by copying pixels from the left column into the central one. Pixels are chosen according to their semantic class (second row) and stacked according to their depths (third row). The white pixels in the depth maps represent areas too far from the camera that cannot be selected.

3.2. DBST (Depth-Based Self-Training)

We describe here our proposal to further improve semantic predictions and distill the knowledge of the entire system into a single network easily deployable at inference time. First, we predict semantic labels for every image in $\mathcal{D}_\mathcal{T}$ by our whole framework (i.e. D4 alongside a selected UDA method, referred to as D4-UDA); then, we use these labels to train a new model on $\mathcal{D}_\mathcal{T}$. This procedure, also known as self-training [26], has become popular in recent UDA for semantic segmentation literature [72, 73, 68, 29, 33, 30] and consists in training a model by its own predictions, referred to as *pseudo-labels*, sometimes through multiple iterations. On the other hand, we only perform one iteration, and the novelty of our approach concerns the peculiar ability to leverage on the depth information available for the images in $\mathcal{D}_\mathcal{T}$ to generate plausible new samples.

Running D4-UDA on $\mathcal{D}_\mathcal{T}$ yields semantic pseudo-labels for every image in $\mathcal{D}_\mathcal{T}$. Yet, as described in Sec. 3.1 (*Semantics from Depth*), each image in $\mathcal{D}_\mathcal{T}$ is also endowed with a depth prediction, provided by a self-supervised monocular depth estimation network. We can take advantage of this information to formulate a novel depth-aware data augmentation strategy whereby portion of images and corresponding pseudo-labels are *copied* onto others so as to synthesize samples for the self-training procedure. The crucial difference between similar approaches presented in [32, 48] and ours consists in the deployment of depth information to steer the data augmentation procedure towards more plausible samples. Indeed, a first intuition behind our method deals with semantic predictions being less accurate for objects distant from the camera: as such predictions play the role of labels in self-training, we prefer to pick closer rather than distant objects in order to generate training samples. Moreover, we reckon certain kinds of objects, like persons, vehicles and traffic signs, to be more plausibly transferable across different images as they tend to be small and less bound to specific spatial locations. On the contrary, it is quite unlikely to merge seamlessly a piece of road or building from a given image into a different one.

Given N randomly selected images x^n from $\mathcal{D}_\mathcal{T}$, with $n \in \{1, \dots, N\}$, paired with semantic pseudo-labels s^n and depth predictions d^n, we augment x^1, by copying on it pixels from the set $\mathcal{X}^{src} = \{x^2, \cdots, x^N\}$. For each pixel of the augmented image we have N possible candidates, one from x^1 itself and $N-1$ from the images in $\mathcal{X}^{src}$. We filter such candidates according to two criteria: the predicted depth should be lower than a threshold t and the semantic prediction should belong to a predefined set of classes, C^*. Hence, we define the set of depths of the filtered candidates at each spatial location p as:

$$D_p = \{d_p^n \mid d_p^n < t \wedge s_p^n \in C^*\} \quad n \in \{1, \dots, N\}. \quad (4)$$

In our experiments, for each image the depth threshold t is set to the 80^{th} percentile of the depth distribution, so as to avoid selecting pixels from the farthest objects in the scene. C^* contains all *things* classes (e.g. *person*, *car*, *traffic light*, etc.), which include foreground elements that can be copied onto other images without altering the plausibility of the scene, while excluding all the *stuff* classes, which include background elements that cannot be easily moved across scenes. This categorization is similar to the one proposed in [55] and we consider it easy to replicate in other datasets.

Then, we synthesize a new image x^z and corresponding pseudo-labels s^z, by assigning at each spatial location p the candidate with the lowest depth, so that objects from different images will overlap plausibly into the synthesized one:

$$x_p^z = x_p^k \quad s_p^z = s_p^k \quad (5)$$

$$k = \begin{cases} 1, & D_p = \emptyset \\ n \text{ s.t. } d_p^n = \min D_p, & D_p \neq \emptyset \end{cases} \quad (6)$$

In Fig. 4 we depict our depth-based procedure to synthesize new training samples, considering, for the sake of simplicity, the case where N is 2.

Hence, with the procedure detailed above, we synthesize an augmented version of $\mathcal{D}_\mathcal{T}$, used to distill the whole D4-UDA framework into a single model by a self-training process. This dataset is much larger and exhibits more variability than the original $\mathcal{D}_\mathcal{T}$. Due to its reliance on depth information, we dub our novel technique as DBST (Depth-Based Self-Training). The results reported in Sec. 4.3 prove its remarkable effectiveness, both when used as the final stage following D4 as well as when deployed as a standalone self-training procedure applied to any other UDA method.

4. Experiments

4.1. Implementation Details

Network Architectures. We use Monodepth2 [14] to generate depth proxy-labels for the procedure described in

Method	Road	Sidewalk	Building	Walls	Fence	Pole	T-light	T-sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorbike	Bicycle	**mIoU**	**Acc**
AdaptSegNet [49]	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.6	32.5	35.4	3.9	30.1	28.1	42.4	85.6
D4-AdaptSegNet + DBST	93.1	53.0	85.1	42.8	27.3	35.8	43.9	18.5	85.9	39.0	89.9	63.0	31.6	86.6	39.8	36.7	0	42.4	35.0	**50.0**	**90.3**
MaxSquare [5]	88.1	27.7	80.8	28.7	19.8	24.9	34.0	17.8	83.6	34.7	76.0	58.6	28.6	84.1	37.8	43.1	7.2	32.2	34.5	44.3	86.9
D4-MaxSquare + DBST	92.9	51.2	84.7	43.5	22.2	35.7	42.5	20.0	86.2	42.0	90.0	63.7	33.0	86.9	45.5	50.9	0	42.2	41.4	**51.3**	**90.3**
BDL [28]	88.2	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5	89.2
D4-BDL + DBST	93.2	52.6	86.4	44.1	31.2	36.5	42.4	36.1	86.3	41.0	89.8	63.3	37.4	86.3	42.8	57.8	0	40.3	37.9	**52.9**	**90.7**
MRNET [69]	90.5	35.0	84.6	34.3	24.0	36.8	44.1	42.7	84.5	33.6	82.5	63.1	34.4	85.8	32.9	38.2	2.0	27.1	41.8	48.3	88.3
D4-MRNET + DBST	93.2	51.6	86.1	45.9	24.5	37.9	47.4	40.4	85.3	37.5	89.6	64.7	39.8	85.8	41.1	53.2	8.9	17.1	33.4	**51.7**	**90.0**
Stuff and things* [55]	90.2	43.5	84.6	37.0	32.0	34.0	39.3	37.2	84.0	43.1	86.1	61.1	29.9	81.6	32.3	38.3	3.2	30.2	31.9	48.3	88.8
D4-Stuff and things + DBST	93.3	54.0	86.5	46.4	32.3	37.7	45.2	39.5	85.5	39.4	90.0	63.7	32.8	85.5	32.0	39.5	0	37.7	35.5	**51.4**	**90.5**
FADA [54]	92.5	47.5	85.1	37.6	32.8	33.4	33.8	18.4	85.3	37.7	83.5	63.2	39.7	87.5	32.9	47.8	1.6	34.9	39.5	49.2	88.9
D4-FADA + DBST	93.9	58.2	86.4	45.9	29.6	36.9	44.6	27.0	86.3	39.4	90.0	64.9	41.0	85.8	34.6	51.2	9.9	24.2	37.3	**52.0**	**90.7**
LTIR [22]	92.9	55.0	85.3	34.2	31.1	34.4	40.8	34.0	85.2	40.1	87.1	61.1	31.1	82.5	32.3	42.9	3	36.4	46.1	50.2	90.0
D4-LTIR + DBST	94.2	59.6	86.9	43.9	35.3	36.9	45.7	36.1	86.2	40.6	90.0	65.9	38.2	84.4	33.3	52.4	13.7	46.2	51.7	**54.1**	**91.0**
ProDA [64]	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5	89.1
D4-ProDA + DBST	94.3	60.0	87.9	50.5	43.0	42.6	50.8	51.3	88.0	45.9	89.7	68.9	41.8	88.0	45.8	63.8	0	50.0	55.8	**58.8**	**92.1**

Table 1. Results on GTA5→CS. When available, checkpoints provided by authors are used. * denotes method retrained by us.

Method	Sky	Building	Road	Sidewalk	Fence	Vegetation	Pole	Car	T-Sign	Person	Bicycle	T-Light	**mIoU**	**Acc**
AdaptSegNet* [49]	75.6	78.0	89.7	28.5	3.4	76.0	28.5	85.1	27.2	55.3	46.6	0	49.5	86.9
D4-AdaptSegNet + DBST	87.7	80.1	94.0	61.8	66.0	81.1	32.2	85.4	31.3	59.0	52.3	0	**55.9**	**90.2**
MaxSquare* [5]	72.4	79.2	89.2	36.0	4.6	75.7	31.5	84.9	30.7	55.8	45.8	8.6	51.2	87.3
D4-MaxSquare + DBST	87.5	80.0	93.7	61.8	7.3	80.8	33.2	84.6	35.1	58.1	48.1	8.2	**56.5**	**90.1**
MRNET* [69]	84.6	79.7	93.9	56.3	0	80.5	35.4	88.9	27.2	59.4	56.3	0	54.5	90.0
D4-MRNET + DBST	88.3	79.9	93.9	63.0	6.3	81.3	35.5	84.3	31.3	59.5	47.9	0	**55.9**	**90.2**

Table 2. Results on SYNSEQ→CS. * denotes method retrained by us.

Sec. 3.1. We adapt the general framework presented in [38] to our setting by deploying the popular Deeplab-v2 [3] for depth estimation and semantic segmentation networks. Both networks consist of a backbone and an ASPP module [3], which substitute, respectively, the encoder and decoder used in [38]. The backbone is implemented as a dilated ResNet50 [61]. We also remove the downsampling and upsampling operations used in [38] when learning the transfer function between depth and semantics. More precisely, in our architecture the transfer function is realized as a simple 6-layers CNN with kernel size 3×3 and Batch Norm [20]. Following the recent trend in UDA for semantic segmentation [49, 5, 28, 69, 55, 54, 22], during DBST we train a single Deeplab-v2 [3] model, with a dilated ResNet101 pre-trained on Imagenet [11] as backbone.

Training Details. Our pipeline is implemented using PyTorch [35] and trained on one NVIDIA Tesla V100 GPU with 16GB of memory. In every training and test phase we resize input images to 1024×512, with the exception of DBST, when we first perform random scaling and then random crop with size 1024×512. During DBST we use also color jitter to avoid overfitting on the pseudo-labels. In our version of [38], the depth and the transfer network are optimized by Adam [23] with batch size 2 for 70 and 40 epochs, respectively, while the semantic segmentation network is trained by SGD with batch size 2 for 70 epochs.The final model obtained by DBST is trained again with SGD, batch size 3 and for 30 epochs. We adopt the One Cycle learning rate policy [45] in every training, setting the maximum learning rate to 10^{-4} but in DBST, where we use 10^{-3}.

4.2. Datasets

We briefly describe the datasets adopted in our experiments, pointing to the Suppl. Mat. for additional details. We follow common practice [49, 22, 28] and test our framework in the synthetic-to-real case using GTA5 [39, 40] or SYNTHIA [42] as synthetic datasets. The former consists in synthetic images captured with the game Grand Theft Auto V, while the latter is composed of images generated by rendering a virtual city. Since our method requires video sequences to train Monodepth2 [14], we use the split SYNTHIA VIDEO SEQUENCES (SYNTHIA-SEQ) in the experiments involving the SYNTHIA dataset. As for real images, we leverage the popular Cityscapes dataset [10], which consists in a large collection of video sequences of driving scenes from 50 different cities in Germany.

4.3. Results

We report here experimental results obtained in two domain adaptation benchmarks, which show how the combination with our D4 method allows to boost performance of recent UDA for semantic segmentation approaches.

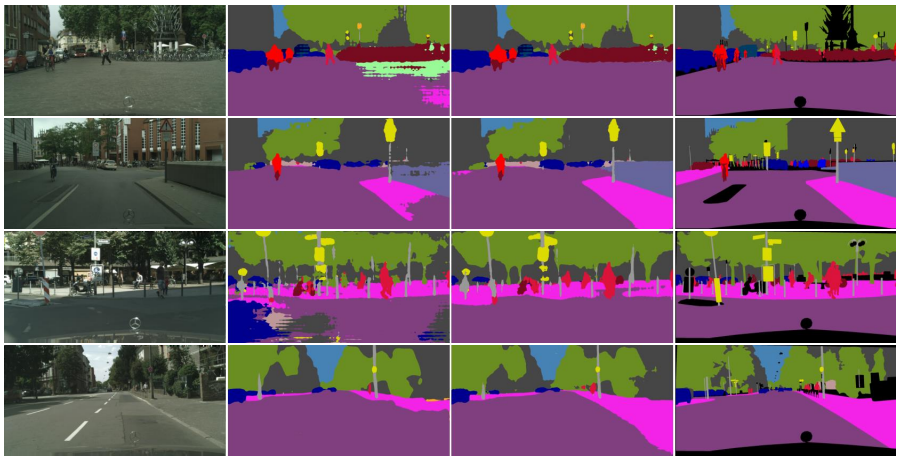

Figure 5. From left to right: RGB image, prediction from UDA method, prediction from D4-UDA + DBST, GT. The top two rows deal with GTA5→CS, the other two with SYNSEQ→CS. Selected methods are, from top to bottom: LTIR [22], BDL [28], MaxSquare [5] and MRNET [69]. In all these examples our proposal can ameliorate dramatically the output of the given stand-alone method, especially on classes featuring large and regular shapes, like *road* in rows 1-3, *sidewalk* in rows 2-4 and *wall* in row 2.

GTA5→CS. Tab. 1 reports results on the most popular UDA benchmark for semantic segmentation, i.e. GTA5→CS, where methods are trained on GTA5 and tested on Cityscapes. We selected the most relevant UDA approaches proposed in the last years [49, 5, 28, 69, 55, 54, 22, 64], using checkpoints provided by authors when available. We report per-class and overall results in terms of mean intersection over union (mIoU) and pixel accuracy (Acc), when each method is either used stand-alone or deployed within our proposal (i.e. D4 + DBST). The reader may notice how every UDA method does improve considerably if combined with our proposal, despite the variability of their stand-alone performances. Indeed, AdaptSegNet [49], which yields about 42 in terms of mIoU, reaches 50 when embedded into our framework. Likewise, ProDA, currently considered the s.o.t.a. UDA method, improves in mIoU from 57.5 to 58.8. Moreover, we can observe in Tab. 1 that our method produces a general improvement for all classes, although we experience a certain performance variability for some of them (such as *train*, *motorbike* and *bicycle*), probably due to noisy pseudo-labels used during DBST. Conversely, our method yields consistently a significant gain on classes characterized by large and regular shapes, namely *road*, *sidewalk*, *building*, *wall* and *sky*. This validates the effectiveness of a) the geometric cues derivable from depth to predict the semantics of these kind of objects and b) the methodology we propose to leverage on these additional cues in UDA settings. This behavior is also clearly observable from qualitatives in Fig. 5. We point out that, to the best of our knowledge, the performance obtained by D4-ProDA + DBST, i.e. 58.8 mIoU (last row of Tab. 1) establishes the new state-of-the-art for GTA5→CS.

SYNSEQ→CS. Akin to common practice in literature we present results also on the popular SYNTHIA dataset. As our pipeline requires video sequences to train the self-supervised monocular depth estimation network, we select the SYNTHIA VIDEO SEQUENCES split for training and the Cityscapes dataset for testing. We will call this setting SYNSEQ→CS. To address it, we re-trained the UDA methods for which the code is available and the training procedure is more affordable in terms of memory and run-time requirements, namely AdaptSegNet [49], MaxSquare [5] and MRNET [69]. The results in Tab. 2 show that all the selected UDA approaches exhibit a substantial performance gain when coupled with our proposal, with a general improvement in all classes. In particular, similarly to the results obtained in GTA5→CS, we observe a consistent improvement for classes related to objects with large and regular shapes (as depicted also in Fig. 5), with the only exception of a slight performance drop for the class *building* when using MRNET [69] (last row of Tab. 2). We argue that our approach is relatively less effective with MRNET [69] as, unlike AdaptSegNet [49] and MaxSquare [5], it yields already satisfactory results in those classes which are usually improved by the geometric clues injected by D4.

In the Suppl. Mat. we show that it is also possible to exploit the depth ground-truths provided by the SYNTHIA dataset as an additional source of supervision during the training of Monodepth2 [14], obtaining a small improvement in the performances of the overall framework.

4.4. Analysis

We report here the most relevant analysis concerning our work. Additional ones can be found in the Suppl. Mat..

Ablation studies. In Tab. 3, we analyze the impact on the performance of our two main contributions, i.e. injec-

tion of geometric cues into UDA methods by D4 and DBST. Purposely, we select the GTA5→CS benchmark and, for the top performing UDA methods, we report the mIoU figures obtained by using the stand-alone UDA method (column *UDA*), combining it with D4 (column *D4-UDA*), applying DBST directly on the stand-alone method (column *UDA + DBST*) and embedding the method into our full pipeline (column *D4-UDA + DBST*). We can observe that each of our novel contributions improves the performance of the most recent UDA methods by a large margin, which is even more remarkable considering that the selected methods already include one or more step of self-training. Moreover, D4 and DBST further enhance the performances of any selected method when deployed jointly, as shown in the column *D4-UDA + DBST*, suggesting that they are complementary. In order to further assess the effectiveness of DBST, in the column *D4-UDA + ST* we report results obtained by D4-UDA in combination with a baseline self-training procedure, which consists in simply fine-tuning the model by its own predictions on the images of the target domain. As the only difference between this procedure and our DBST is the dataset employed for fine-tuning, the results prove the effectiveness of DBST in generating a varied set of plausible samples more amenable to self-training than the original images belonging to the target domain.

Alternative strategies to exploit depth. As explained in Sec. 3.1 *Semantics from Depth*, we rely on the mechanism of transferring features across tasks and domains from [38] to inject depth cues into semantic segmentation. To validate our choice, we explore two possible alternatives, namely DeepLabV2-RGBD and DeepLabV2-Depth. Both consist in the popular DeepLabV2 [3] network, with RGBD images in input in the first case and depth maps (no RGB) in the second (more details in the Suppl. Mat.). Tab. 4 compares the performance of these alternatives with our method, either when used standalone (rows 2, 3, and 4) or when combined with LTIR [22] according to the strategy presented in Sec. 3.1 *Combine with UDA*. Results allow us to make some important considerations. First, our intuition on the possibility of exploiting depth to improve semantics is correct since also simple approaches improve over the baseline (reported in the first row of the table). Nonetheless, these naive methods produce a significantly smaller improvement compared to our approach, showing that our decision to adapt [38] to the UDA scenario is not obvious. Moreover, [38] requires only RGB images at test time. Finally, when combined with LTIR [22], a stronger depth-to-semantic model provides better results, validating our choice once again.

Impact of video sequences. As described in Sec. 3.1, we obtain depth proxy-labels with a self-supervised depth estimation network [14], that we train using the raw video sequences (just RGB images) provided by the datasets involved in our experiments. In order to validate that using video sequences from the target domain doesn't provide any advantage to our framework, we train AdaptSegNet [49] on GTA5→CS using the whole training split available for Cityscapes (i.e. 83300 images with temporal consistency). We choose AdaptSegNet [49] since it can be considered the building block of many UDA methods. We observe a drop in performances from 42.4 to 41.9 mIoU, showing that using video sequences does not boost semantic segmentation in a UDA setting, probably because of the similarity between consecutive frames, and that the improvement produced by our framework is provided by the effective strategy that we adopt to exploit depth.

Method	UDA	D4-UDA	UDA + DBST	D4-UDA + DBST	D4-UDA + ST
BDL [28]	48.5	49.6	51.7	**52.9**	50.1
MRNET [69]	48.3	49.6	50.0	**51.7**	50.3
Stuff and Things* [55]	48.3	49.1	50.4	**51.4**	49.4
FADA [54]	49.3	49.9	51.4	**52.0**	50.0
LTIR [22]	50.2	51.1	53.1	**54.1**	51.5
ProDa [64]	57.5	57.6	58.0	**58.8**	56.8

Table 3. Impact on performance of the two components of our proposal (D4, DBST) when applied separately or jointly to selected UDA methods on GTA5→CS. * indicates that the method was retrained by us. Results are reported in terms of mIoU.

Method	mIoU
DeepLabV2-RGB	34.5
DeepLabV2-RGBD	35.5
DeepLabV2-Depth	36.5
Semantics from depth (Sec. 3.1)	**43.1**
DeepLabV2-RGBD $\bigoplus$ LTIR [22]	47.7
DeepLabV2-Depth $\bigoplus$ LTIR [22]	49.3
D4-LTIR	**51.1**

Table 4. Comparison between alternative methods to infer semantics from depth. DeepLabV2-RGB, DeepLabV2-RGBD and DeepLabV2-Depth stand for DeepLabV2 [3] trained on $\mathcal{D}_S$, using respectively RGB images, RGBD images or depth proxy-labels as input, while "Semantics from depth" is the approach described in Sec. 3.1 *Semantics from Depth*. The symbol $\bigoplus$ represents the merge operation described in Sec. 3.1 *Combine with UDA*. Results are reported in terms of mIoU on the Cityscapes dataset.

5. Conclusion

We have shown how to exploit self-supervised monocular depth estimation in UDA problems to obtain accurate semantic predictions for objects with strong geometric priors (like road and buildings). As all recent UDA approaches lack such geometric knowledge, we build our D4 method as a depth-based add-on, pluggable into any UDA method to boost performances. Finally, we employed self-supervised depth estimation to realize an effective data augmentation strategy for self-training. Our work highlights the possibility of exploiting auxiliary tasks learned by self-supervision to better tackle UDA for semantic segmentation, paving the way for novel research directions.

References

[1] Matteo Biasetton, Umberto Michieli, Gianluca Agresti, and Pietro Zanuttigh. Unsupervised domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2

[2] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. 2

[3] 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, Apr 2018. 1, 6, 8

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[5] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 6, 7

[6] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2019. 2

[7] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. 2

[8] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 1

[9] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 2

[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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 6

[11] J. Deng, W. Dong, R. Socher, L. 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, 2009. 1, 6

[12] Ravi Garg, Vijay Kumar B.G., Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. *Lecture Notes in Computer Science*, page 740–756, 2016. 2

[13] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 2

[14] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 1, 2, 3, 5, 6, 7, 8

[15] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR)*, 2020. 2

[16] 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)*, pages 770–778, 2016. 2

[17] Judy Hoffman, E. Tzeng, T. Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 2

[18] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation, 2016. 1

[19] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Koring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11130–11140, 2021. 2

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. 6

[21] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[22] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. 1, 2, 3, 6, 7, 8

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. 6

[24] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision*, pages 582–600. Springer, 2020. 2

[25] Jogendra Nath Kundu, Nishank Lakkakula, and R Venkatesh Babu. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1436–1445, 2019. 3

[26] D. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013. 3, 5

[27] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. Spigan: Privileged adversarial learning from simulation. In *International Conference on Learning Representations*, 2019. 2

[28] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. 2, 6, 7, 8

[29] Qing Lian, Lixin Duan, Fengmao Lv, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 2, 5

[30] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *The European Conference on Computer Vision (ECCV)*, August 2020. 2, 3, 5

[31] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 2

[32] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning, 2020. 5

[33] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. 2, 3, 5

[34] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *CoRR*, abs/1606.02147, 2016. 1, 4

[35] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*, 2017. 6

[36] Fabio Pizzati, Raoul de Charette, Michela Zaccaria, and Pietro Cerri. Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[37] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *Asian Conference on Computer Vision*, pages 298–313. Springer, 2018. 2

[38] Pierluigi Zama Ramirez, Alessio Tonioni, Samuele Salti, and Luigi Di Stefano. Learning across tasks and domains. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 1, 2, 3, 4, 6, 8

[39] 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. 6

[40] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. *Lecture Notes in Computer Science*, page 102–118, 2016. 1, 2, 6

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, page 234–241, 2015. 1

[42] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 6

[43] Suman Saha, Anton Obukhov, Danda Pani Paudel, Menelaos Kanakis, Yuhua Chen, Stamatios Georgoulis, and Luc Van Gool. Learning to relate depth and semantics for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8197–8207, 2021. 2

[44] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, Apr 2017. 1

[45] Leslie N. Smith and Nicholay Topin. Super-convergence: very fast training of neural networks using large learning rates. *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, May 2019. 6

[46] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020. 1

[47] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019. 2

[48] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1379–1389, January 2021. 3, 5

[49] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 2, 6, 7, 8

[50] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 2

[51] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015. 3

[52] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. 2

[53] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez Perez. Dada: Depth-aware domain adaptation in semantic segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 2

[54] Haoran Wang, Tong Shen, Wei Zhang, Lingyu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *The European Conference on Computer Vision (ECCV)*, August 2020. 2, 6, 7, 8

[55] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. 1, 2, 5, 6, 7, 8

[56] Kohei Watanabe, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Multichannel semantic segmentation with unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2

[57] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2162–2171, 2019. 2

[58] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gökhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S. Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. *Lecture Notes in Computer Science*, page 535–552, 2018. 1, 2

[59] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12613–12620, Apr 2020. 2

[60] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014. 1

[61] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 6

[62] Pierluigi Zama Ramirez, Alessio Tonioni, and Luigi Di Stefano. Exploiting semantics in adversarial training for image-level domain adaptation. In *2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, pages 49–54, 2018. 2

[63] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 1

[64] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. *arXiv preprint arXiv:2101.10979*, 2021. 3, 6, 7, 8

[65] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[66] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 2

[67] Y. Zhang and Zilei Wang. Joint adversarial learning for domain adaptation in semantic segmentation. In *AAAI*, 2020. 2

[68] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision (IJCV)*, 2020. 2, 3, 5

[69] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Jul 2020. 3, 6, 7, 8

[70] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 2

[71] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[72] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018. 2, 3, 5

[73] Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 5

Supplementary Material For Plugging Self-Supervised Monocular Depth into Unsupervised Domain Adaptation for Semantic Segmentation

Adriano Cardace Luca De Luigi Pierluigi Zama Ramirez Samuele Salti Luigi Di Stefano
Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
{adriano.cardace2, luca.deluigi4, pierluigi.zama}@unibo.it

1. Additional Implementation Details

As stated in Sec. 3.1 of the main paper, we obtain depth proxy-labels by deploying a self-supervised method for solving monocular depth estimation from video sequences. Specifically, we train Monodepth2 [6] following the training protocol and hyper-parameters used in the original paper. We train it for 20 epochs using mixed mini-batch of size 6, composed of 3 real and 3 synthetic images. We resize samples at resolution 1024×512 for training and testing. It is important to train the network on both domains jointly because we want depth predictions to be *aligned* across domains. Self-supervised depth methods typically estimates depth maps up to a scale-factor. Thus, we train on both domains simultaneously to force the network to yield predictions from the two domains that share the same range and scale. When $\mathcal{D}_\mathcal{S}$ is synthetic, we can collect depth ground-truth labels with minimum effort. In such case, we could exploit these labels to provide an additional source of supervision to Monodepth2. SYNTHIA-SEQ provides much less images with smaller variability with respect to GTA5, but provides depth ground-truth labels. Thus, in the SYNSEQ→CS setting, we could train Monodepth2 by adding a L_1 loss between predictions and ground-truths of SYNTHIA-SEQ to the set of Monodepth2 losses, so as to achieve better pseudo-labels results. Nevertheless, the availability of ground-truth labels is not crucial to improve the performance of the considered UDA method. Indeed, in Tab. 1 we can observe that the use of synthetic depth ground-truth labels provides just a slight performance improvement (i.e. 1% mIoU or less).

As regards the training of semantics prediction from depth features, we follow the protocol explained in [9]. We train the depth network simultaneously on $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$, by minimizing the mean absolute error (i.e. L_1 loss) between predicted depth maps and depth proxy-labels, previously generated for both domains. Then, we train the semantic network only on $\mathcal{D}_\mathcal{S}$, using a weighted Cross Entropy loss with weights computed as in [18]. The weights of the two networks are pre-initialized on ImageNet, and, following a common protocol [14, 18, 7], all Batch Normalization layers are frozen both at training and test time to use ImageNet statistics. Differently from [9], we deploy the more performant DeepLabV2 [1] architecture for both networks: as the framework requires to split the network into an encoder and a decoder, we consider the backbone as the encoder and the ASPP module as the decoder. Hence, the transfer function in D4 is learned by minimizing the mean squared distance (i.e. L_2 loss) between the semantic features extracted by the semantic network encoder and the ones hallucinated by the transfer function itself starting from the depth encoder. Finally, during DBST, the final distilled model is obtained by minimizing a standard Cross Entropy loss on $\mathcal{D}_\mathcal{T}$ and exploiting only the pseudo-labels, as explained in Sec. 3.2 of the main paper.

2. Additional Datasets Details

Cityscapes. The Cityscapes dataset [4] provides a large collection of video sequences of driving scenes from 50 different European cities. The dataset is composed of 150000 video-sequence images, of which 83300 are used for training. A subset of 5000 images from Cityscapes is commonly used as benchmark for semantic segmentation, as these images are annotated with high-quality pixel-level semantic labels (19 classes). This subset is split into train, validation and test with 2975, 500 and 1525 images respectively. In our experiments we train Monodepth2 [6] on the 83300 training sequences. For training D4 and DBST we use the 2975 train images (without their semantic labels) and, following the protocol adopted in recent works [14, 2, 8, 18, 17, 16, 7], we evaluate our final model on the validation split. The augmented dataset obtained during DBST starting from the 2975 images accounts for 7500 samples.

GTA5. The GTA5 dataset [10, 11] consists in synthetic

images captured while playing the video-game Grand Theft Auto V. It consists of 120000 video-sequence images that we use in the Monodepth2 [6] training procedure. Moreover, the dataset provides 24966 samples with fine semantic annotations (same 19 classes as Cityscapes). We train the depth network of D4 on only 3000 randomly sampled images among the 24966 to keep the training balanced with the 2975 images of Cityscapes. Finally, we train the semantic and transfer network of D4 on the whole 24966 synthetic images.

SYNTHIA VIDEO SEQUENCES. The SYNTHIA dataset [12] is composed of images generated by rendering a virtual city created with the Unity development platform. Since our method requires video sequences to train Monodepth2 [6], we use the split SYNTHIA VIDEO SEQUENCES, selecting sub-sequences *Spring*, *Summer*, *Fall*, *Winter*, *Dawn* and *Fog*. We collect thus a total of 26948 images, paired with fine-grained semantic labels (12 classes in common with Cityscapes). In particular, we train on *sky*, *building*, *road*, *sidewalk*, *fence*, *vegetation*, *pole*, *car*, *traffic sign*, *person*, *bicycle*, *traffic light*. It is worth noticing that to make the Cityscapes dataset consistent with SYNTHIA VIDEO SEQUENCES, it is necessary to map the Cityscapes class *rider* into *bicycle* and collapse *bus* and *truck* into *car*. We use only 3000 randomly sampled images to train the depth, semantic and transfer network of D4, as well as for the training of the other considered methods which were retrained by us (* in Tab. 2 of the main paper) due to the authors not providing their results on SYNTHIA VIDEO SEQUENCES.

3. Semantics From Depth

In this section, we evaluate alternative ways to predict semantics in the target domain by exploiting also the depth cues available once depth proxy-labels have been computed as discussed in sec 3.1 (*Semantics from depth*) of the main paper. This study motivates our choice to rely on the mechanism of transferring features across tasks and domains [9], with the improvements and modifications discussed in Sec. 4.1 of the main paper and Sec. 1 of this supplementary document. As we have semantic labels only for the source domain $\mathcal{D}_\mathcal{S}$, all approaches are trained only on $\mathcal{D}_\mathcal{S}$, and their ability to generalize is assessed on the target domain $\mathcal{D}_\mathcal{T}$.

We investigate two possible alternatives, namely:

- a semantic segmentation network that processes RGB-D images, where the proxy depth of each image is stacked as an additional channel
- a semantic segmentation network that processes directly proxy depths, without using RGB information.

We realize both options by training the popular DeepLabV2 [1] architecture to perform semantic segmentation on $\mathcal{D}_\mathcal{S}$, initializing the network with ImageNet [5] pre-trained weights. Moreover, in the first case, we add a convolutional layer at the beginning of the architecture, to reduce the input RGBD channels from 4 to 3, while in the second case we obtain 3-channels input images by stacking three times the proxy depth map. In the following, we will call DeepLabV2-RGBD the first network and DeepLabV2-Depth the second one. We also consider as baseline the performance of DeepLabV2 trained only on RGB images, referred to as DeepLabV2-RGB.

Method	mIoU
AdaptSegNet* [14]	49.5
D4-AdaptSegNet + DBST (w/o synthetic GT)	55.9
D4-AdaptSegNet + DBST (w/ synthetic GT)	56.9
MaxSquare* [2]	51.2
D4-MaxSquare + DBST (w/o synthetic GT)	56.5
D4-MaxSquare + DBST (w/ synthetic GT)	57.4
MRNET* [18]	54.5
D4-MRNET + DBST (w/o synthetic GT)	55.9
D4-MRNET + DBST (w/ synthetic GT)	56.3

Table 1. Results on the SYNSEQ→CS benchmark with or without synthetic ground-truths. * denotes method retrained by us.

In Tab. 2 we report mIoU results obtained on Cityscapes (i.e. our target domain) by DeepLabV2-RGB, DeepLabV2-RGBD, DeepLabV2-Depth, and our method. We observe that the RGBD and the Depth versions yield slightly better results compared to the RGB baseline. Interestingly, DeepLabV2-Depth provides better results than DeepLabV2-RGB and DeepLabV2-RGBD, which supports our intuition about semantic cues extracted from depth alone being more effectively transferable across different domains due to their reliance on geometry rather than appearance. Yet, the ability to overcome the domain shift by DeepLabV2-RGBD and DeepLabV2-Depth is limited, as performance is low for both variants. On the contrary, by tackling the problem with the method proposed in the main paper, we can improve the baseline by 8.6% in terms of mIoU.

Moreover, we evaluate DeepLabV2-RGBD and DeepLabV2-Depth also in combination with an UDA method, as proposed in Sec. 3.1 (*Combine with UDA*) of the main paper. In the last three rows of Tab. 2, we report mIoU results obtained by such combinations (row 5 and 6), compared to our proposal (last row), while considering one of the best performing UDA methods, namely LTIR [7]. As intuitively expected, we observe that a better depth-based semantic model leads to a better combination with the selected UDA method, motivating once again the need for an approach robust to domain-shift in order to infer semantics from depth cues in UDA settings.

Rather than relying on self-supervised depth on both do-

Method	mIoU
DeepLabV2 RGB	34.5
DeepLabV2-RGBD	35.5
DeepLabV2-Depth	36.5
Semantics from depth (sec 3.1)	**43.1**
DeepLabV2-RGBD $\bigoplus$ LTIR [7]	47.7
DeepLabV2-Depth $\bigoplus$ LTIR [7]	49.3
D4-LTIR (*i.e.* Semantics from depth $\bigoplus$ LTIR)	**51.1**

Table 2. Comparison between alternative methods to infer semantics with the aid of depth cues. DeepLabV2-RGB, DeepLabV2-RGBD and DeepLabV2-Depth stand for DeepLabV2 [1] trained on $\mathcal{D}_\mathcal{S}$, using respectively RGB images, RGBD images or depth proxy-labels as input, while "Semantics from depth" is the approach described in the subsection with the same name of sec 3.1 in the main paper. The symbol $\bigoplus$ represents the merge operation described in subsection *Combine with UDA* of Sec. 3.1 of the main paper. Results are reported in terms of mIoU on the Cityscapes dataset.

mains as done for the previous cases, one may try to use just the depth provided by synthetic source dataset. To the best of our knowledge, only two works [15, 3] proposed to exploit depth in a UDA context for outdoor scenes segmentation. We compare here our D4 module with [15], the only publicly available framework, to show that the additional information for the target domain is a key component for Domain Adaptation. We retrained [15] with the same hyper-parameters, and changed only the training split (i.e. SYNTHIA-SEQ instead of SYTNHIA-RAND-CITYSCAPES). As Tab 3 shows, D4 surpasses by a large margin (3.6%) [15], suggesting that self-supervised information for the target domain can be used to boost performance in Domain Adaptation.

Method	mIoU
DADA [15]	42.3
D4 (ours)	45.9

Table 3. Comparison between depth-based frameworks.

4. DBST vs DACS [13]

In Tab. 4 we compare our DBST with the method presented in DACS [13], as they share some similarities. In particular, both approaches generate training samples by *copying* portions of images onto other images. However, they differ in three main aspects:

- [13] copies portions of images from $\mathcal{D}_\mathcal{S}$ onto images from $\mathcal{D}_\mathcal{T}$, while in our DBST we use exclusively images from $\mathcal{D}_\mathcal{T}$.
- In our proposal, we copy only image patches whose semantic predictions belong to a predefined set of classes that we deem as more amenable to be moved across images, like, e.g., *person*, *car* and *pole*; conversely, in [13] no semantic filter is applied to select the patches that will be copied across the images.
- Unlike [13], we exploit depth information to plausibly stack objects in the generated sample.

Method	mIoU
D4-LTIR [7]	51.1
D4-LTIR [7] + DACS [13]	52.7
D4-LTIR [7] + DBST	**54.1**

Table 4. Comparison between the approach proposed in [13] (DACS) and our DBST, when applied to our D4 combined with [7]. Results are reported in terms of mIoU in the GTA5→CS benchmark.

Method	mIoU
AdaptSegNet (w/o video) [14]	42.4
AdaptSegNet (w/ video)	41.9

Table 5. AdaptSegNet [14] trained with or without additional unlabeled target images

In addition to these points, in our DBST we further exploit depth information to guide the selection of the patches to be copied by excluding areas of the scene that are too far away from the camera, where semantic predictions are less likely accurate. In Tab. 4 we report results in the GTA5→CS benchmark when applying DBST or [13] to D4 combined with [7]: our DBST outperforms the strategy proposed in [13], though the latter can also yield a notable performance improvement.

5. Adding videos to UDA methods

In this section, we empirically demonstrate that using additional raw information is not directly useful for the UDA setting in semantic segmentation. To this purpose, we adopt [14], which makes use of adversarial training and it can be considered as the main building block of many UDA methods proposed in the literature. Moreover, adversarial training is a plausible strategy to exploit additional unlabeled images for the target domain. Driven by this reasoning, we retrained [14] in the GTA5→CS benchmark using the whole training split available in Cityscapes (i.e. 83300 images with temporal consistency). The result reported in Tab. 5 suggests that simply collecting more data is not enough to boost semantic semantic segmentation in a UDA setting, and more advanced techniques as the one proposed in this work are necessary to extrapolate useful data.

6. Qualitative Results

In Fig. 1, 2, 3, 4, 5, 6 we report several qualitative results of our D4 proposal combined with the different UDA methods reported in Tab. 1 and Tab. 2 of the main paper. In every case, we observe an overall improvement in the quality of the predictions. In particular, thanks to the additional information provided by depth maps, the errors in large objects with regular shapes are partially removed (see first and second column of Fig. 1). Moreover, with the proposed merging algorithm (Sec 3.1) and with the DBST algorithm detailed in Sec. 3.2, we also preserve the good performance of the selected UDA method for certain classes. For instance, all the predictions concerning classes such as *pole* and *traffic sign* are always maintained or even improved (see second row of Fig. 2).

7. DBST - Qualitative Results

In Fig. 7 and 8 we show some training samples obtained with our DBST algorithm. As explained in Sec. 3.2 of the main paper, we use multiple images from $\mathcal{D}_\mathcal{T}$ as source, alongside with the corresponding depth maps and predictions (referred to as pseudo-labels), to synthesize new training pairs. We can notice how the newly generated samples contain a lot of patterns that would not be present in the original images, enabling a more effective Self-Training procedure. We also point out how, thanks to the use of depth maps, the generated pairs look realistic. For example, in the third row of Fig. 7, the rider on the left side of the image is pasted in front of the pole since it appears closer in the depth maps of the two images.

8. Depth Proxy-Labels

Fig. 9, 10, 11 report depth proxy-labels obtained in the first step of our pipeline by the self-supervised approach proposed in Monodepth2 [6]. We note how the produced depth maps are smooth and accurate on the static parts of the scene (such as road and buildings), while they tend to be noisy on moving objects (like cars and pedestrians). Despite these imperfections, depth proxy-labels produced by [6] provide a solid base of geometric clues for objects with large and regular shapes, which are extensively exploited in our proposal.

References

[1] 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, Apr 2018. 1, 2, 3

[2] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 1, 2, 6, 8

[3] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. 3

[4] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[5] J. Deng, W. Dong, R. Socher, L. 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, 2009. 2

[6] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 1, 2, 4, 11

[7] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. 1, 2, 3, 7

[8] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. 1, 7

[9] Pierluigi Zama Ramirez, Alessio Tonioni, Samuele Salti, and Luigi Di Stefano. Learning across tasks and domains. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 1, 2

[10] 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. 1

[11] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. *Lecture Notes in Computer Science*, page 102–118, 2016. 1

[12] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[13] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1379–1389, January 2021. 3

[14] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 1, 2, 3, 6, 8

[15] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez Perez. Dada: Depth-aware domain

adaptation in semantic segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 3

[16] Haoran Wang, Tong Shen, Wei Zhang, Lingyu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *The European Conference on Computer Vision (ECCV)*, August 2020. 1

[17] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. 1

[18] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Jul 2020. 1, 2

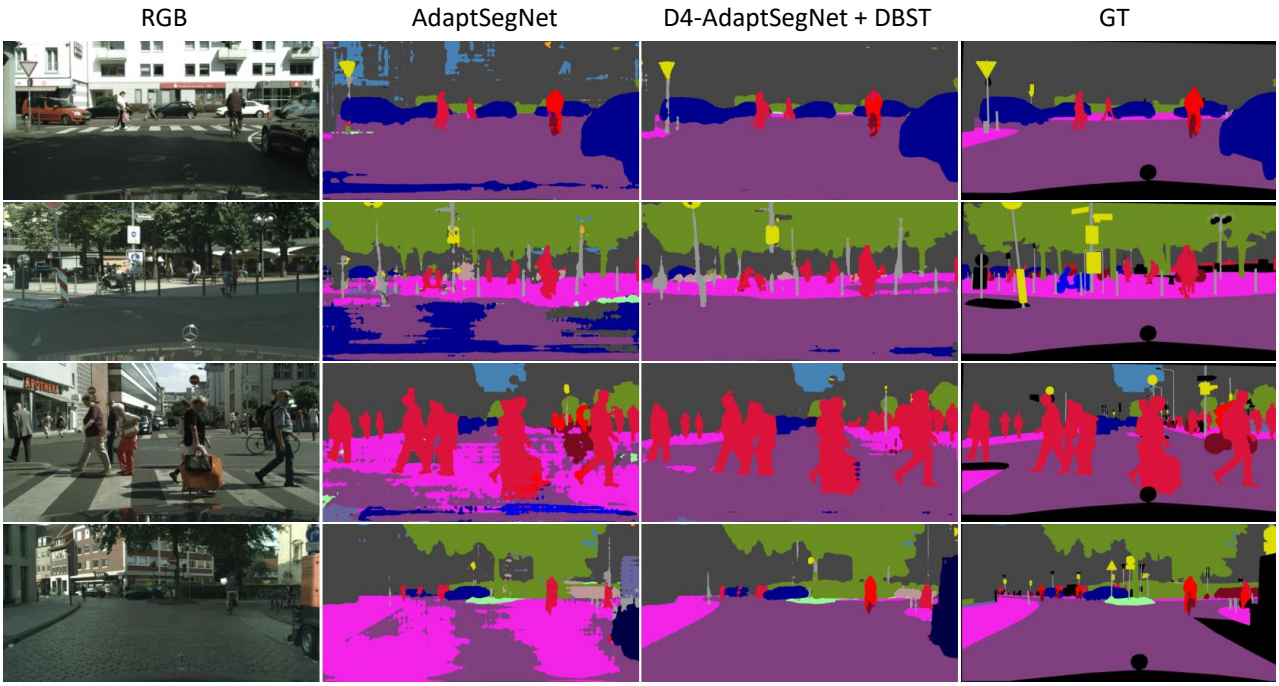

Figure 1. Qualitative results in the GTA5→CS benchmark. From left to right: RGB, prediction from Adaptsegnet [14], prediction from D4-AdaptSegNet + DBST (our proposal), Ground-Truth.

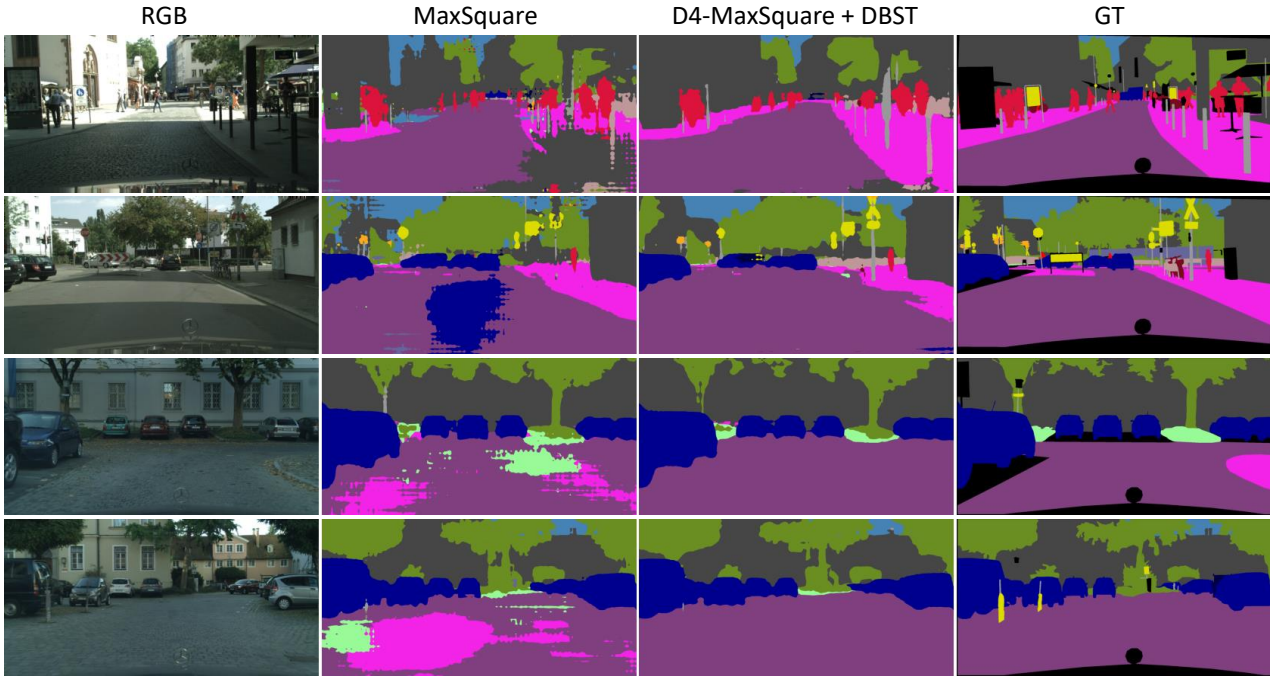

Figure 2. Qualitative results in the GTA5→CS benchmark. From left to right: RGB, prediction from MaxSquare [2], prediction from D4-MaxSquare + DBST (our proposal), Ground-Truth.

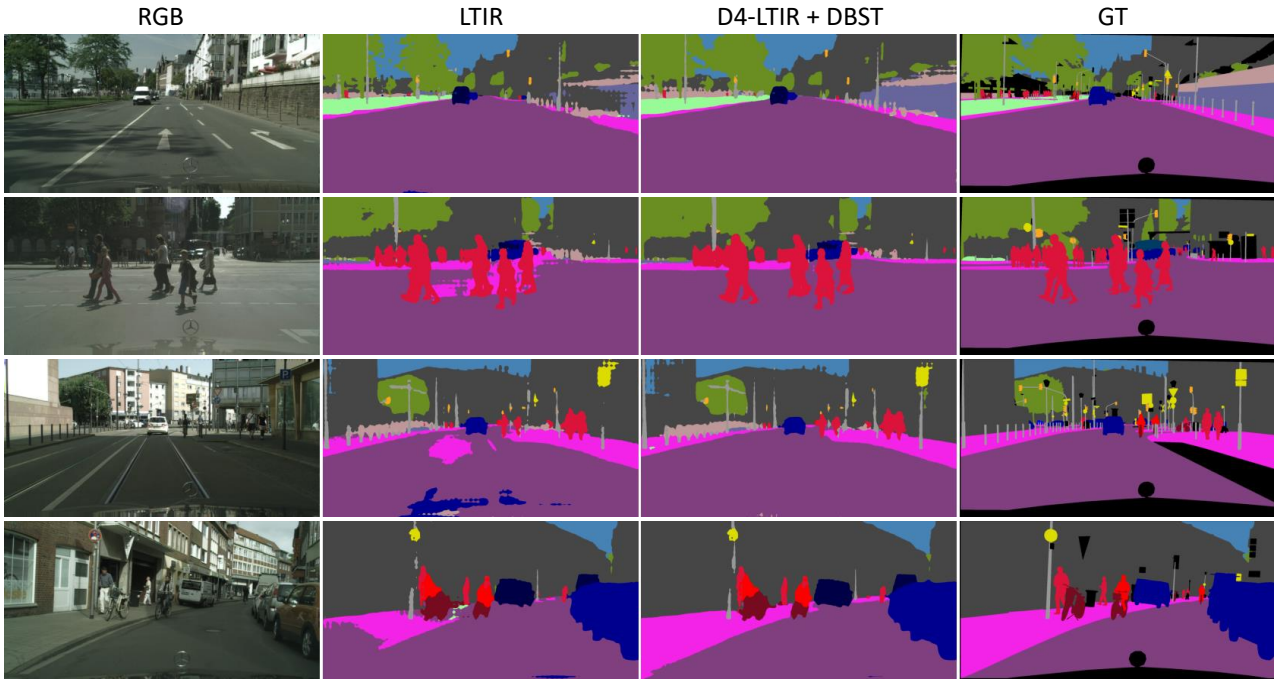

Figure 3. Qualitative results in the GTA5→CS benchmark. From left to right: RGB, prediction from LTIR [7], prediction from D4-LTIR + DBST (our proposal), Ground-Truth.

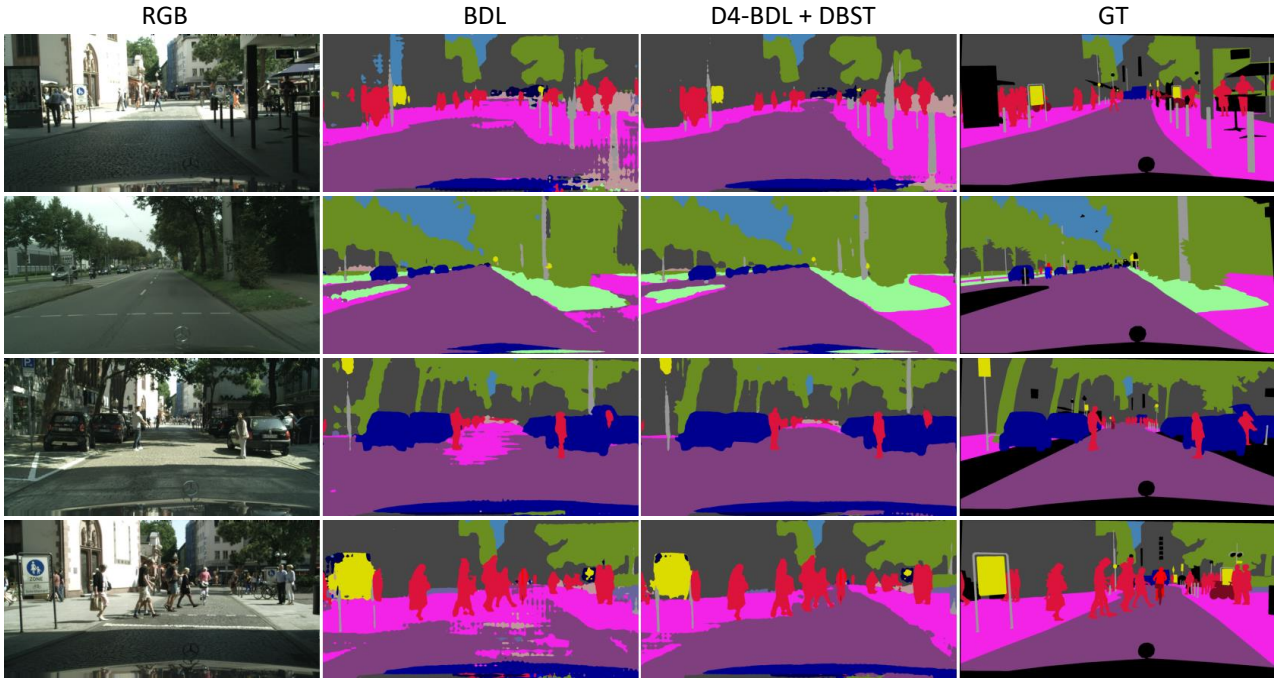

Figure 4. Qualitative results in the GTA5→CS benchmark. From left to right: RGB, prediction from BDL [8], prediction from D4-BDL + DBST (our proposal), Ground-Truth.

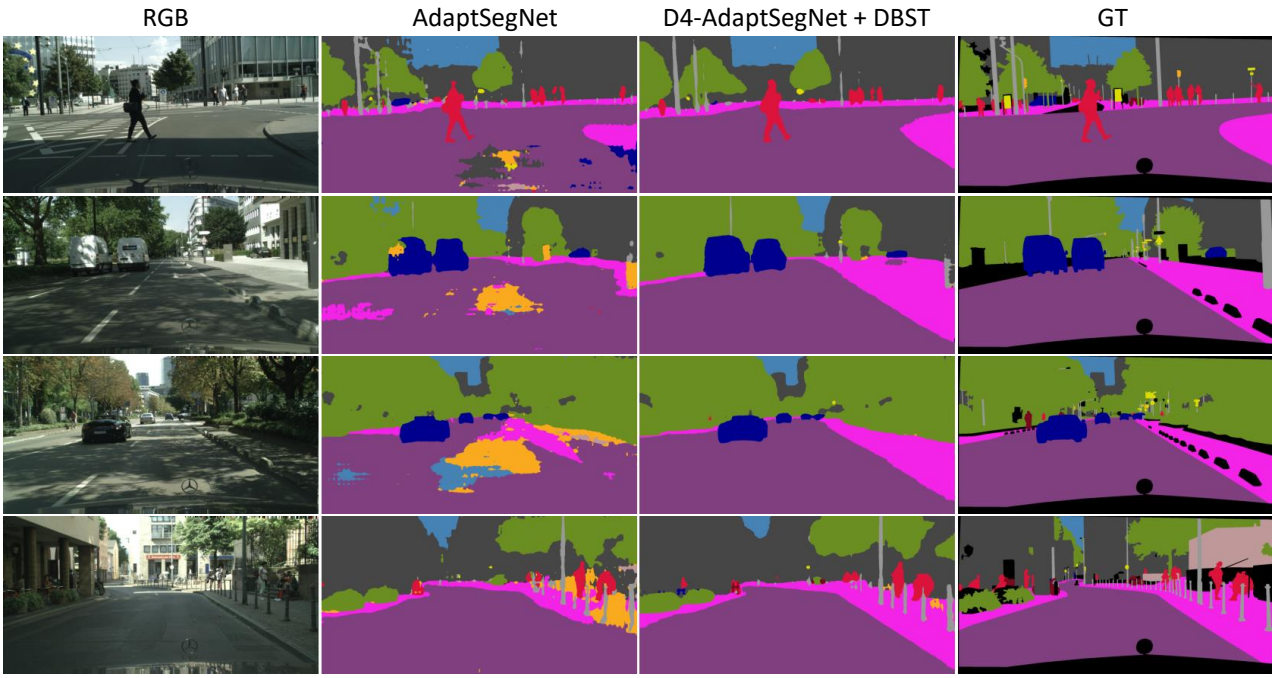

Figure 5. Qualitative results in the SYNSEQ→CS benchmark. From left to right: RGB, prediction from AdaptSegNet [14], prediction from D4-AdaptSegNet + DBST (our proposal), Ground-Truth.

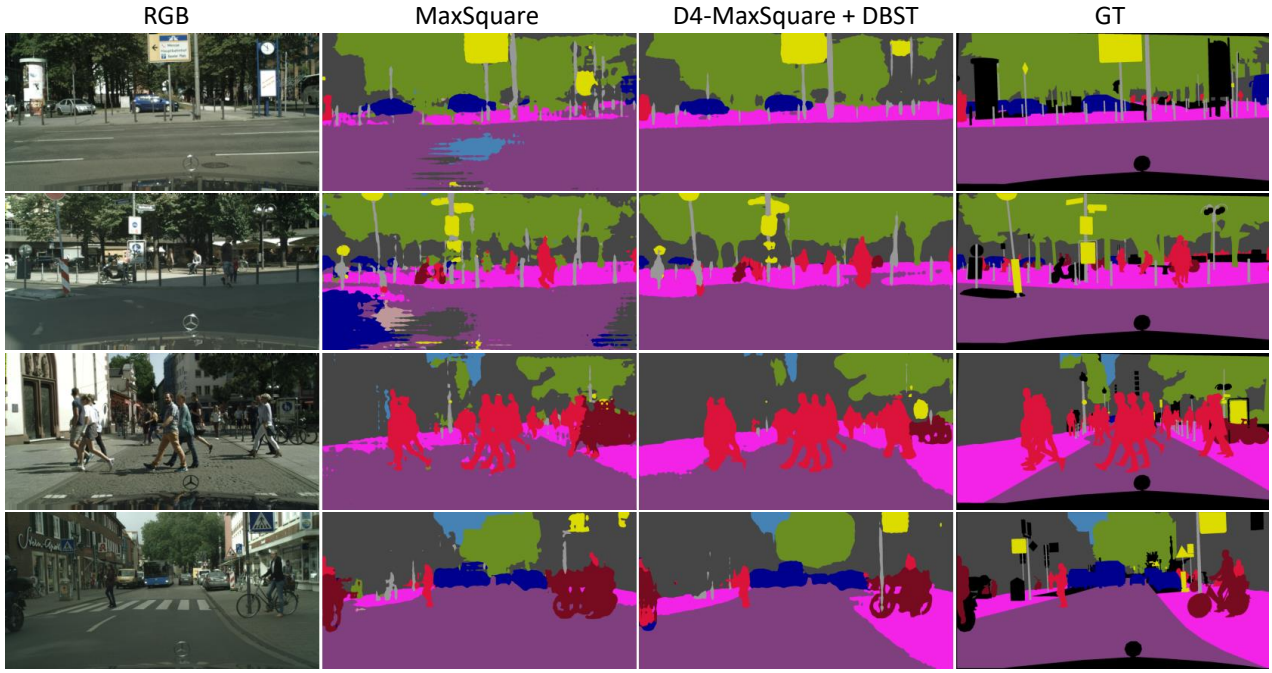

Figure 6. Qualitative results in the SYNSEQ→CS benchmark. From left to right: RGB, prediction from MaxSquare [2], prediction from D4-MaxSquare + DBST (our proposal), Ground-Truth.

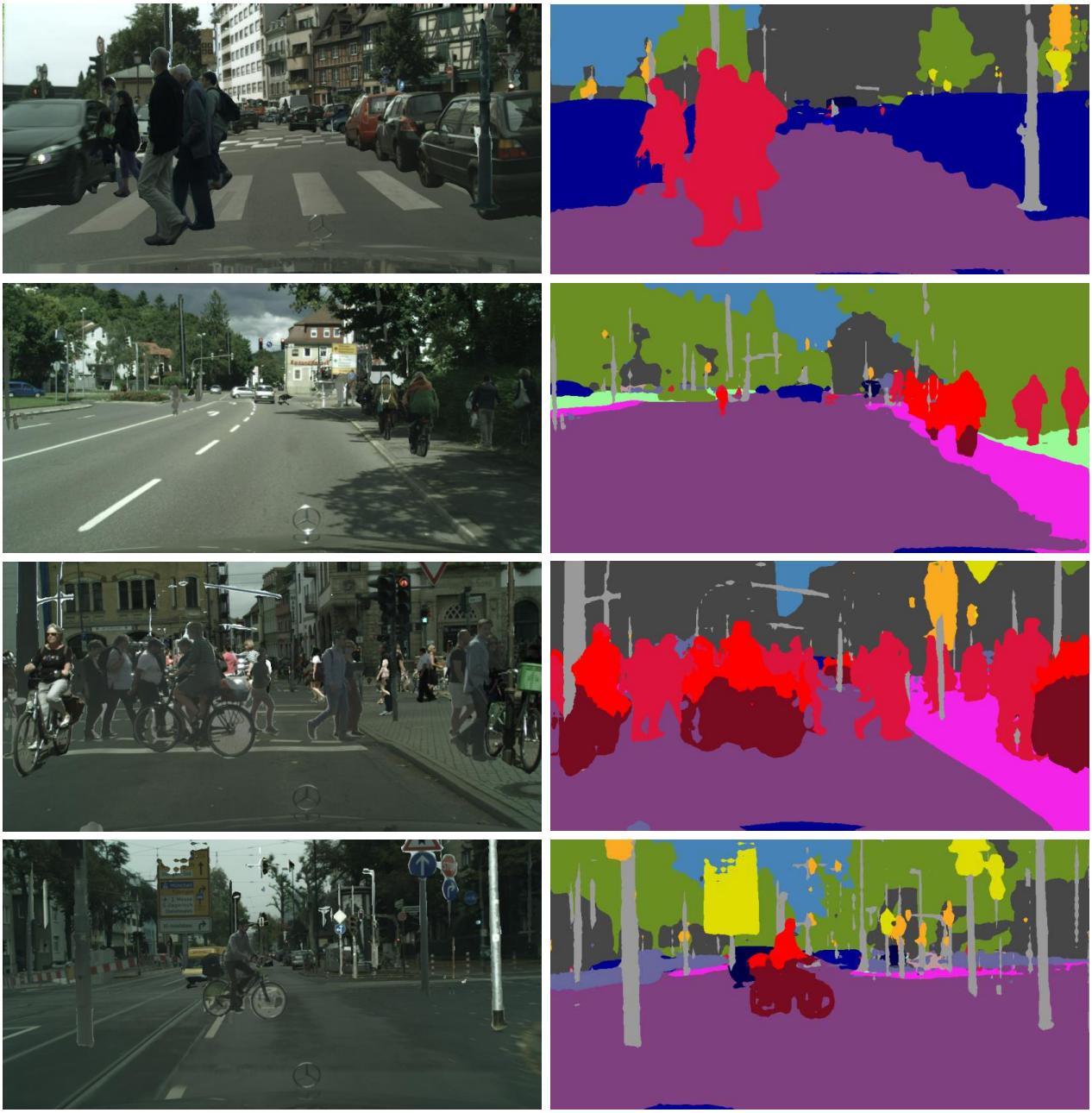

Figure 7. RGB and pseudo-labels generated for our DBST procedure using D4-LTIR in the GTA5→CS benchmark.

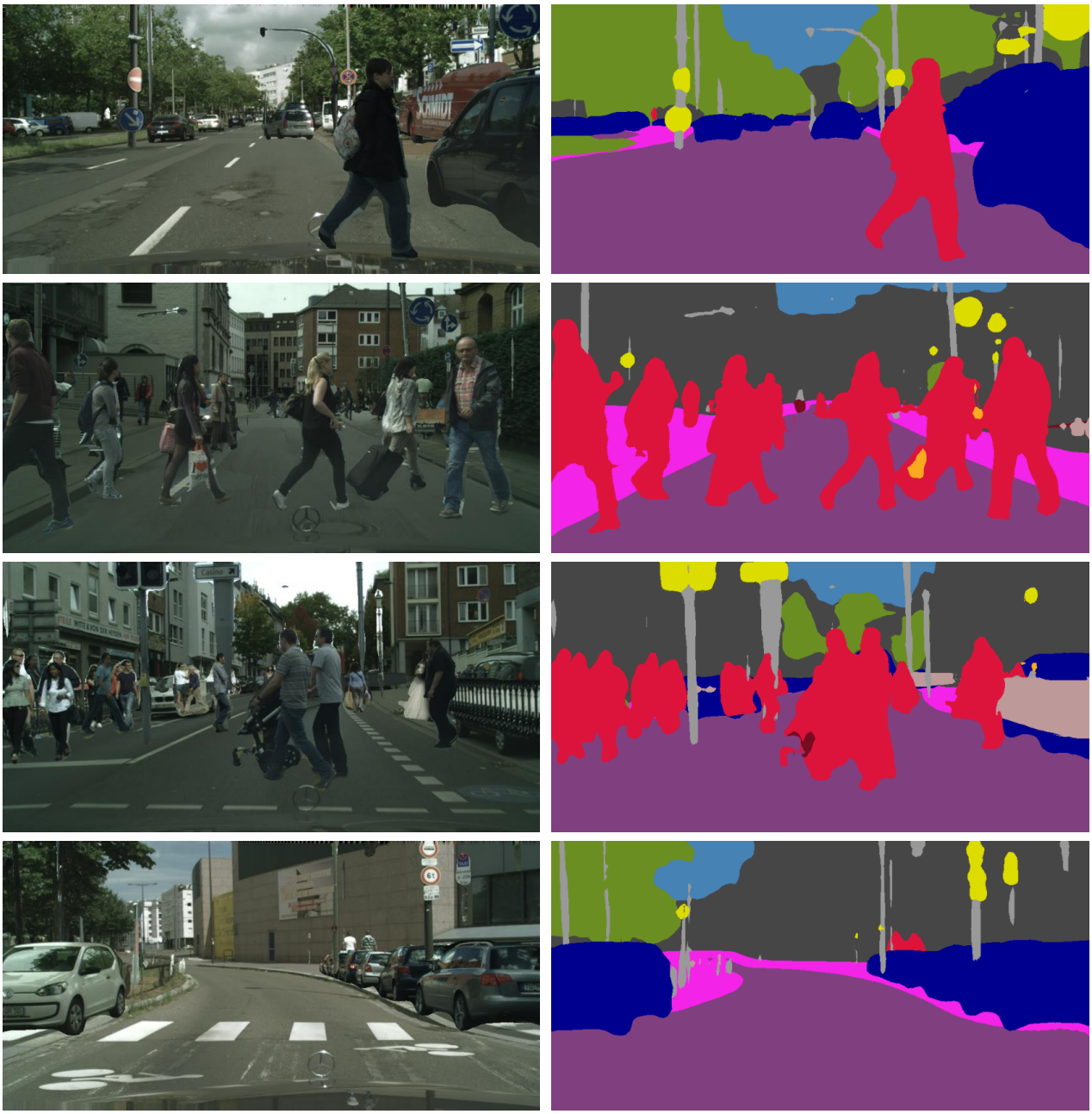

Figure 8. RGB and pseudo-labels generated for our DBST procedure using D4-MRNET in the SYNSEQ→CS benchmark.

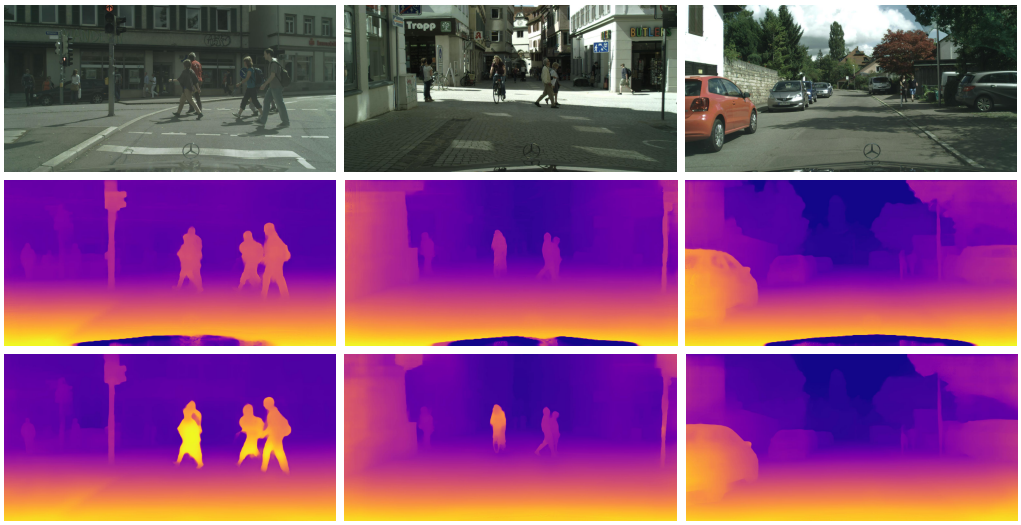

Figure 9. Depth proxy-labels for the Cityscapes dataset obtained with Monodepth2 [6]. From top to bottom: RGB, depth obtained by training Monodepth2 on Cityscapes and GTA5 sequences, depth obtained by training Monodepth2 on Cityscapes and SYNTHIA-SEQ sequences. Depth maps are shown as inverse depth maps for a better visualization.

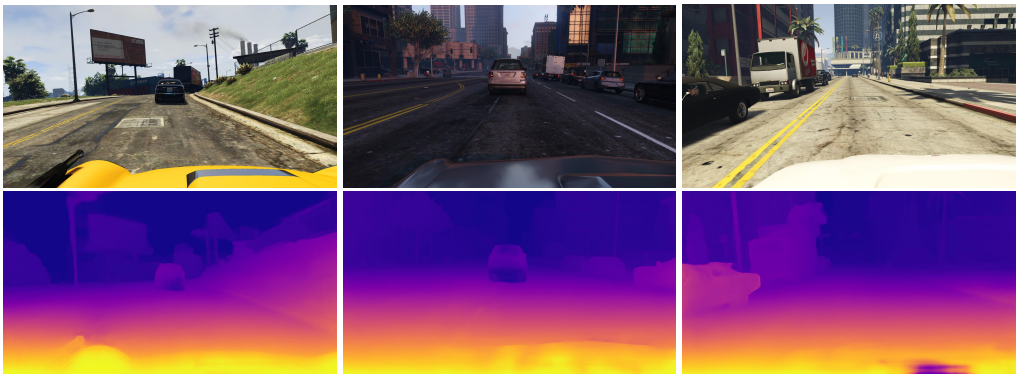

Figure 10. Depth proxy-labels for the GTA5 dataset obtained with Monodepth2 [6]. We show RGB images (first row) and corresponding depth maps (second row), shown as inverse depth maps for a better visualization.

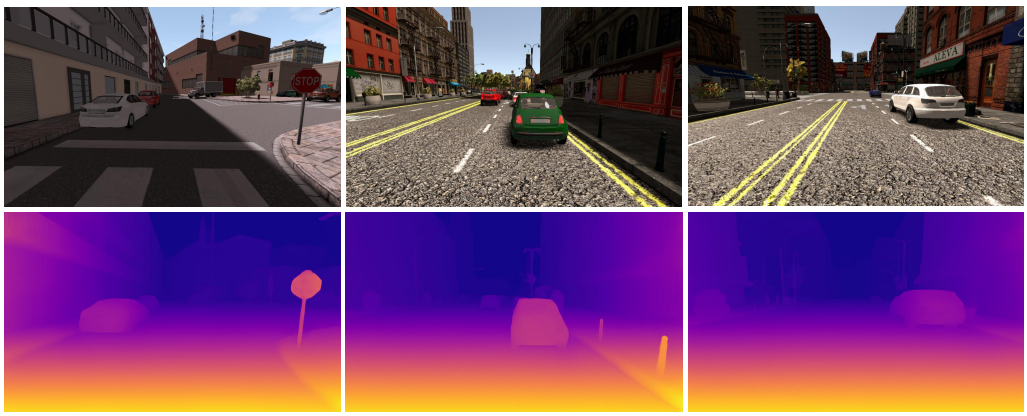

Figure 11. Depth proxy-labels for the SYNTHIA-SEQ dataset obtained with Monodepth2 [6]. We show RGB images (first row) and corresponding depth maps (second row), shown as inverse depth maps for a better visualization.