

Multi-Target Unsupervised Domain Adaptation for Semantic Segmentation without External Data

Yonghao Xu^{1,2}, Pedram Ghamisi^{1,3}, Yannis Avrithis¹

¹Institute of Advanced Research in Artificial Intelligence (IARAI), Austria

²Computer Vision Laboratory, Linköping University, Sweden

³Helmholtz Institute Freiberg for Resource Technology, HZDR, Germany

yonghao.xu@liu.se, p.ghamisi@hzdr.de, yannis@avrithis.net

Abstract

Multi-target unsupervised domain adaptation (UDA) aims to learn a unified model to address the domain shift between multiple target domains. Due to the difficulty of obtaining annotations for dense predictions, it has recently been introduced into cross-domain semantic segmentation. However, most existing solutions require labeled data from the source domain and unlabeled data from multiple target domains concurrently during training. Collectively, we refer to this data as “external”. When faced with new unlabeled data from an unseen target domain, these solutions either do not generalize well or require retraining from scratch on all data. To address these challenges, we introduce a new strategy called “multi-target UDA without external data” for semantic segmentation. Specifically, the segmentation model is initially trained on the external data. Then, it is adapted to a new unseen target domain without accessing any external data. This approach is thus more scalable than existing solutions and remains applicable when external data is inaccessible. We demonstrate this strategy using a simple method that incorporates self-distillation and adversarial learning, where knowledge acquired from the external data is preserved during adaptation through “one-way” adversarial learning. Extensive experiments in several synthetic-to-real and real-to-real adaptation settings on four benchmark urban driving datasets show that our method significantly outperforms current state-of-the-art solutions, even in the absence of external data. Our source code is available online (<https://github.com/YonghaoXu/UT-KD>).

1. Introduction

Among many other computer vision tasks, progress in deep learning has significantly advanced semantic segmentation [1]. Nevertheless, the particular difficulty of seman-

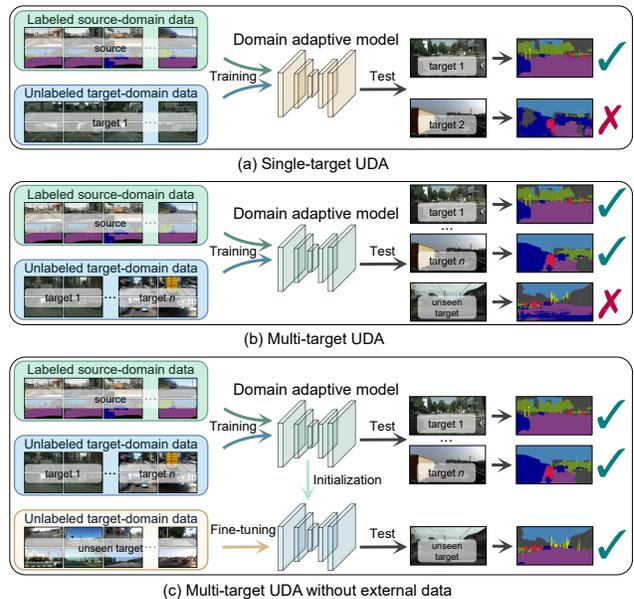


Figure 1. Different strategies in cross-domain semantic segmentation. (a) *Single-target unsupervised domain adaptation (UDA)*: the segmentation model cannot generalize well to unseen domains. (b) *Multi-target UDA*: target domains are still predetermined at training and the model needs to be retrained from scratch on all data when a new unseen target domain is given, or else it will suffer from the same problem. (c) Our new strategy, *multi-target UDA without external data*: the pre-trained model is quickly adapted to a new unseen target domain without accessing any external data from the original source or target domains.

tic segmentation, being a dense prediction task, is that training a deep learning-based model usually requires a large amount of pixel-level annotations, which are laborious and time-consuming to collect. To address this challenge and mitigate the insufficiency of labeled data, *unsupervised domain adaptation (UDA)* algorithms have been recently developed for *cross-domain semantic segmentation* [10, 15].

The latter aims to learn a domain-adaptive segmentation model by training on *labeled* source-domain data and *unlabeled* target-domain data [16].

So far, most of the existing UDA methods are designed for a *single target* domain [29]. The main limitation of such methods is that the segmentation model may perform well in the target domain they are trained on, but can hardly generalize well to other unseen domains [11]. Consequently, they are unable to perform effectively, for example, on urban driving data from different cities with diverse visual styles and imaging environments, as demonstrated in **Figure 1**(a). A natural idea to address this problem is to extend to *multi-target* UDA [14, 24]. This involves adapting a labeled source domain to multiple unlabeled target domains, as illustrated in **Figure 1**(b). However, since the target domains are predetermined during training, the model either does not generalize well to unseen domains or requires retraining from scratch on all data to do so. This significantly increases the cost of applying these approaches to new domains and renders them unsuitable when the original external data is inaccessible.

To address these challenges, we introduce a new strategy called *multi-target UDA without external data* for semantic segmentation. In particular, the segmentation model is first trained on labeled source-domain and unlabeled target-domain data from multiple targets. This data is collectively referred to as *external*. Then, the pre-trained segmentation model is adapted to a new unseen target domain without accessing any external data, as shown in **Figure 1**(c). This strategy leverages the knowledge of the pre-trained model to eliminate the dependency on external data. Therefore, it is more scalable compared to existing multi-target UDA approaches and remains applicable even when external data is inaccessible.

Our contributions are summarized as follows:

1. We introduce a new multi-target UDA strategy for semantic segmentation, where the segmentation model is adapted quickly to an unseen domain using unlabeled data of this domain alone, *without external data*.
2. To exhibit this strategy, we design a simple method called *multi-target knowledge distillation* (MT-KD), which uses self-distillation and adversarial learning and achieves new state-of-the-art performance on multi-target UDA for semantic segmentation.
3. As a second step, we modify MT-KD by removing access to labeled data and supervision. This new method, called *unseen target knowledge distillation* (UT-KD), directly adapts a pre-trained MT-KD model to a new unseen target domain by a novel *one-way* adversarial learning strategy, without external data. To the best of our knowledge, we are the first to do so.
4. To further boost performance, we perform visual style

transfer across multiple domains. We parameterize the style of each domain by a single vector, thus decoupling it from the style transfer process itself. The latter is performed by a *multi-target style transfer network* (MT-STN), which is shared across all domains.

2. Related work

2.1. Single-target unsupervised domain adaptation

Early research for cross-domain semantic segmentation primarily focuses on the adaptation of source-domain knowledge to a specific target domain. The prevailing approach commonly employed involves acquiring domain-invariant representations through the use of *adversarial learning*. This adaptation process can occur within various spaces, such as the intermediate feature space [18], the output feature space [26], or within the realm of fine-grained categorical features [19]. Given that the primary discrepancy among different domains lies in their visual appearances, an alternative strategy involves the application of *visual style transfer* to directly mitigate the domain disparity [30]. Nonetheless, these methodologies, while effective within their intended single target domain, tend to exhibit limited generalization capabilities across unseen domains [11].

2.2. Multi-target unsupervised domain adaptation

To address the limitation of single-target UDA, Isobe *et al.* [11] propose the first *multi-target UDA* approach for semantic segmentation. In particular, they first train an expert model for each source-target pair and then conduct collaborative learning with each expert model to achieve adaptation between different target domains. Saporta *et al.* [24] further extend adversarial learning into the multi-target UDA setting, where one discriminator for each target domain aims to discriminate that domain from all other target domains. To achieve more efficient multi-target UDA, Lee *et al.* [14] directly adapt a single model to multiple target domains without training multiple domain-specific expert models. However, since the aforementioned multi-target UDA methods are trained on predetermined multiple target domains, the entire model still needs to be retrained from scratch on all data when a new unseen target domain is given; otherwise, it will suffer from the same limitation of single-target UDA: it will not generalize well. This makes it difficult to apply these approaches to unseen domains.

2.3. Source-free domain adaptation

While UDA approaches typically necessitate access to labeled data from the source domain and unlabeled data from the target domain, the practical application of these approaches might be hindered by potential privacy issues that could undermine the availability of source data. In such

Table 1. Characteristics of diverse problem settings in cross-domain semantic segmentation. X_s : source data; $\mathcal{X}_t = \{X_{t_n}\}_{n=1}^N$: target data; X_u : “unseen” data used as target at inference, possibly after fine-tuning. EXT: using external data (X_s or \mathcal{X}_t) while training on X_u , either at pre-training or fine-tuning. Single-target and multi-target UDA have to “see” X_u at pre-training. Source-free DA and domain generalization do not use any domain other than X_s and X_u . Example: G: GTA5; C: CityScapes; I: IDD.

SETTING	PRE-TRAINING		FINE-TUNE	EXT	KNOWLEDGE FLOW	EXAMPLE
	SOURCE	TARGET				
Single-target UDA	X_s	X_u	–	✓	$X_s \rightarrow X_u$	G→I
Multi-target UDA	X_s	$\mathcal{X}_t \cup \{X_u\}$	–	✓	$X_s \rightarrow \mathcal{X}_t \cup \{X_u\}$	G→{C, I}
Source-free DA	X_s	–	X_u	✗	$X_s \rightarrow X_u$	G→I
Domain generalization	X_s	–	–	✗	$X_s \rightarrow X_u$	G→I
Multi-target UDA w/o external data (ours)	X_s	\mathcal{X}_t	X_u	✗	$(X_s \rightarrow \mathcal{X}_t) \rightarrow X_u$	(G→C)→I

scenarios, an alternative strategy is to directly transfer the knowledge from a segmentation model pre-trained on the source domain to the target domain. This setting is known as source-free domain adaptation [23]. Liu *et al.* [17] propose the first source-free domain adaptation approach for semantic segmentation. Specifically, their approach involves self-supervised learning on the target domain with both pixel- and patch-level optimization objectives. Huang *et al.* [9] further propose a historical contrastive learning framework using a historical source hypothesis to compensate for absent source data. Kundu *et al.* [13] use a multi-head generalization framework with self-training. All of these methods solely draw knowledge from a single source domain, as they operate under the assumption that only the pre-trained segmentation model from the source domain is at their disposal. Consequently, the transfer of knowledge from both the source domain and other known target domains remains a non-trivial challenge.

2.4. Domain generalization

In contrast to DA, domain generalization aims to enhance a segmentation model’s ability to perform effectively in new, unseen domains. This improvement is achieved without utilizing data from the target domain during training; instead, one or more source domains are employed. Common strategies employed for domain generalization include learning domain-agnostic feature representations [2, 15] and style augmentation [31]. Despite the simplicity of these methods, their performance is relatively restricted as they neglect to incorporate any data from the target domain during the training phase.

3. Problem formulation

Formally, let X_s and $\mathcal{X}_t = \{X_{t_n}\}_{n=1}^N$ denote the labeled *source-domain* data and the unlabeled *target-domain* data, respectively, where N is the number of target domains. This data is collectively called *external*. The source domain data X_s contains pairs of the form (x, y) , where $x \in [0, 1]^K$ is an input gray-scale image and $y \in \mathbb{R}^{K \times C}$ is the correspond-

ing dense one-hot encoded class label; K is the number of pixels and C is the number of classes in the segmentation task. The target domain data \mathcal{X}_t contains only unlabeled images $x \in [0, 1]^K$. Let X_u denote the new *unseen target-domain* data used at inference, consisting of unlabeled images $x \in [0, 1]^K$.

Table 1 summarizes the characteristics of different problem settings in cross-domain semantic segmentation. Single-target UDA, multi-target UDA, and domain generalization are one-stage methods, while source-free DA and the proposed multi-target UDA without external data include a second stage of fine-tuning on X_u after pre-training. We use EXT to refer to using external data (X_s or \mathcal{X}_t) while training on X_u either at pre-training or fine-tuning. The detailed formulation of each setting follows.

Single-target UDA aims to learn a domain adaptive segmentation model F using the labeled source-domain data X_s and the unlabeled target-domain data X_t . Since we aim to conduct inference on the unseen target-domain data X_u in this study, the target-domain data X_t will become X_u for single-target UDA methods. The knowledge flow is thereby from the source domain X_s to the “unseen” target domain X_u , which has to be available at pre-training with X_s .

Multi-target UDA is trained with X_s and multiple target-domain data $\mathcal{X}_t = \{X_{t_n}\}_{n=1}^N$. To adapt to the new unseen target domain, X_u will be regarded as the $(N + 1)$ -th target domain for multi-target UDA methods and the complete target domain data used at pre-training will become $\mathcal{X}_t \cup \{X_u\}$. Accordingly, the knowledge flow is from the source domain to multiple target domains: $X_s \rightarrow \mathcal{X}_t \cup \{X_u\}$. Again, X_u has to be available at pre-training with X_s and \mathcal{X}_t .

Source-free DA involves two stages. In the first stage, a segmentation model F is pre-trained on the source domain X_s . In the second stage, F is fine-tuned on the unseen target domain X_u . Thus, the knowledge flow is solely from the source domain to the unseen target domain: $X_s \rightarrow X_u$. The transfer of knowledge from other known target domains (*i.e.*, \mathcal{X}_t) is not feasible in this case.

Domain generalization aims to enhance a segmentation

model’s generalization ability to other unseen domains by training with X_s alone without seeing any target-domain data. The knowledge flow is also solely from the source domain to the unseen target domain: $X_s \rightarrow X_u$, in this case without even adapting to X_u .

The proposed new strategy, *multi-target UDA without external data*, involves two stages. In the first pre-training stage, it learns a domain adaptive segmentation model F on X_s and \mathcal{X}_t . After pre-training, it is expected that the obtained model F inherits knowledge from both X_s and \mathcal{X}_t . In the second stage, only the pre-trained model F and X_u are available. The aim is to distill the knowledge in F and adapt it to X_u without accessing any external data from X_s and \mathcal{X}_t . Thus, the knowledge flow is first from the source domain to multiple known target domains, then to the new unseen target domain: $(X_s \rightarrow \mathcal{X}_t) \rightarrow X_u$.

In this sense, the new strategy is similar to multi-target UDA in using multiple targets \mathcal{X}_t , thus acquiring as much knowledge as is available, and to source-free DA in fine-tuning on X_u without access to external data, thus quickly adapting to new unseen domains.

4. Methodology

Here, we introduce our methodology and the particular implementation of our new strategy, *multi-target UDA without external data*. We first describe our *multi-target knowledge distillation* (MT-KD) method in detail, which uses self-distillation and adversarial learning for multi-target UDA. We then simplify it to derive our new *unseen target knowledge distillation* (UT-KD) method, which quickly adapts a pre-trained MT-KD model to an unseen target domain, without accessing any external data from the original source or any other target domain. Finally, we introduce a new *multi-target style transfer network* (MT-STN) to achieve visual style transfer across multiple domains, which can serve as an add-on component for style augmentation.

4.1. Multi-target knowledge distillation

As shown in **Figure 2**, the key idea of *multi-target knowledge distillation* (MT-KD) is to conduct self-distillation and adversarial learning across multiple target domains, so that the knowledge from the labeled source domain is distilled and adapted to multiple target domains.

Formally, we aim to learn a *student network* F_S , using a *teacher network* F_T of the same architecture, whose parameters ϕ_T^i at iteration i are obtained by exponential moving average (EMA) [25] on the parameters of the student ϕ_S^i :

$$\phi_T^i = \alpha \phi_T^{i-1} + (1 - \alpha) \phi_S^i, \quad (1)$$

where α is a decay parameter. Both networks are functions of the form $F : \mathbb{R}^{K \times 3} \rightarrow \mathbb{R}^{K \times C}$, which map an input image $x \in [0, 1]^K$ to a predicted segmentation map

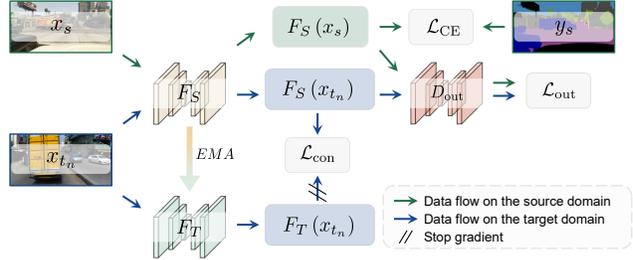


Figure 2. Illustration of our *multi-target knowledge distillation* (MT-KD). Given a set of labeled images X_s from the source domain and unlabeled images $\mathcal{X}_t = \{X_{t_n}\}_{n=1}^N$ from multiple target domains, the student network F_S is trained by cross-entropy \mathcal{L}_{CE} on the source domain, consistency loss \mathcal{L}_{con} on the target domains and adversarial loss \mathcal{L}_{out} in the output space. The teacher network F_T is obtained by the exponential moving average (EMA) of F_S parameters. Only one target domain is shown for brevity.

$F(x) \in \mathbb{R}^{K \times C}$. The vector $F(x)^{(k)} \in \mathbb{R}^C$ is a distribution of predicted class probabilities at pixel k and $F(x)^{(k,c)} \in \mathbb{R}$ is the predicted probability for class c at pixel k .

On the labeled source domain data X_s , we define the supervised dense cross-entropy loss

$$\mathcal{L}_{CE}(X_s, F_S) = \mathbb{E}_{(x,y) \sim X_s} \ell_{CE}(y, F_S(x)) \quad (2)$$

$$\ell_{CE}(y, q) = -\frac{1}{K} \sum_{k=1}^K (y^{(k)})^\top \log q^{(k)}. \quad (3)$$

To distill knowledge from the labeled source domain to multiple unlabeled target domains, we apply the consistency regularization to the student predictions on unlabeled examples from multiple target domains by minimizing their mean squared error (MSE) from the teacher predictions:

$$\mathcal{L}_{con}(\mathcal{X}_t, F_S) = \sum_{n=1}^N \mathbb{E}_{x \sim X_{t_n}} \ell_{con}(\mathcal{A}(x), F_S) \quad (4)$$

$$\ell_{con}(x, F_S) = \frac{1}{K} \sum_{k=1}^K \left\| F_S(x)^{(k)} - F_T(x)^{(k)} \right\|^2, \quad (5)$$

where \mathcal{A} is an input transformation for data augmentation. In practice, we adopt CutMix [6] along with the proposed style transfer network MT-STN. See **Appendix C** for more details on the effect of different choices.

To encourage the F_S to yield domain-invariant segmentation maps, we further introduce a discriminator D_{out} with a DCGAN architecture [21] to perform adversarial learning in the output space across the source and multiple target domains. In particular, the adversarial loss is defined as

$$\mathcal{L}_{out}(X_s, \mathcal{X}_t, F_S, D_{out}) = \mathcal{L}^+(X_s, F_S, D_{out}) + \sum_{n=1}^N \mathcal{L}^-(X_{t_n}, F_S, D_{out}), \quad (6)$$

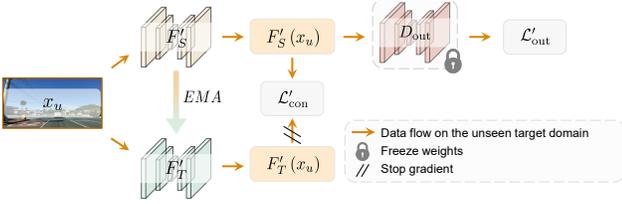


Figure 3. Illustration of our *unseen target knowledge distillation* (UT-KD). Given a set of unlabeled images X_u from an unseen target domain, UT-KD distills and adapts the knowledge from a pre-trained MT-KD model by self-distillation and one-way adversarial learning. Both student and teacher networks F'_S, F'_T are initialized from the pre-trained model. Same for the discriminator D_{out} , which remains frozen.

where the two terms

$$\mathcal{L}^+(X, F, D) = \mathbb{E}_{x \sim X} \log(1 - D(F(x))) \quad (7)$$

$$\mathcal{L}^-(X, F, D) = \mathbb{E}_{x \sim X} \log D(F(x)) \quad (8)$$

respectively represent the loss for original examples in each domain that are treated as positive for the discriminator of that domain, and the loss for the examples from other domains that are treated as negative accordingly.

Similar to previous adversarial learning work [26], we optimize (6) through a min-max criterion, where F_S aims to fool D_{out} by maximizing the probability of the target-domain predictions (segmentation maps) being classified as source-domain, while D_{out} aims to discriminate a source domain prediction from predictions of all target domains. The complete objective function is thus

$$\min_{F'_S} \max_{D_{\text{out}}} \mathcal{L}_{\text{CE}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}} + \lambda_{\text{out}} \mathcal{L}_{\text{out}}, \quad (9)$$

with factors λ_{con} and λ_{out} controlling the balance between the two terms.

4.2. Unseen target knowledge distillation

Most multi-target UDA approaches for cross-domain semantic segmentation use a predetermined set of target domains [11, 14, 24]. Thus, the learned model still needs to be retrained from scratch on all data when a new unseen target domain is given, which makes it difficult to apply these approaches to new datasets.

To address this challenge, we introduce a *unseen target knowledge distillation* (UT-KD) method that quickly adapts a pre-trained MT-KD model to a new unseen target domain without accessing any external data from the source or other target domains. As shown in Figure 3, this method is a simplified version of MT-KD, where the source-domain data and the supervised loss are removed. The key idea is to perform self-distillation and adversarial learning directly on the new unseen target domain so that the knowledge from the pre-trained MT-KD model is distilled and adapted.

To achieve this goal, there are again a student network F'_S and a teacher network F'_T . We initialize F'_T according to the pre-trained MT-KD model while training F'_S from scratch. At each iteration, F'_T is again obtained by EMA on the parameters of F'_S . As in subsection 4.1, we perform self-distillation on the unseen target domain data using a consistency loss that minimizes the MSE between the student and teacher predictions

$$\mathcal{L}'_{\text{con}}(X_u, F'_S) = \mathbb{E}_{x \sim X_u} \ell'_{\text{con}}(\mathcal{A}(x), F'_S) \quad (10)$$

$$\ell'_{\text{con}}(x, F'_S) = \frac{1}{K} \sum_{k=1}^K \left\| F'_S(x)^{(k)} - F'_T(x)^{(k)} \right\|^2, \quad (11)$$

where $\mathcal{A}(x)$ is data augmentation, as in (4). Although there are no labels in X_u , this loss allows the student F'_S to self-train, guided by the teacher F'_T .

More importantly, we now also have a pre-trained discriminator D_{out} from MT-KD that can discriminate segmentation maps between the source and multiple target domains. Considering that the new unseen target domain may be distinctly different from the source domain, the pre-trained D_{out} should tend to classify predictions for input examples $x \in X_u$ as the target domain. Since our goal is to make the UT-KD model yield domain-invariant segmentation maps on the unseen target domain, a natural idea is to perform adversarial learning to fool the pre-trained D_{out} by maximizing the probability of the unknown target-domain predictions being classified as the source-domain. Accordingly, the adversarial loss is

$$\mathcal{L}'_{\text{out}}(X_u, F'_S) = \mathcal{L}^-(X_u, F'_S, D_{\text{out}}), \quad (12)$$

where the negative loss \mathcal{L}^- is defined in (8). Since there is no source data, there is no positive term as in (6). Thus, this is *one-way* adversarial learning. According to our knowledge, we are the first to introduce such an approach in domain adaptation. In addition, we keep the discriminator D_{out} fixed, as pre-trained by MT-KD. This is what prevents the segmentation model F'_S from forgetting the knowledge acquired from the external data while it is being adapted. There is thus no maximization as in (9), and the complete objective function becomes

$$\min_{F'_S} \lambda_{\text{con}} \mathcal{L}'_{\text{con}} + \lambda_{\text{out}} \mathcal{L}'_{\text{out}}. \quad (13)$$

4.3. Multi-target style transfer network

To further mitigate the visual appearance shift between the source and multiple target domains and boost the performance of MT-KD, we introduce a *multi-target style transfer network* (MT-STN). As shown in Figure 4, the main idea is to simultaneously learn the style of each domain. A shared

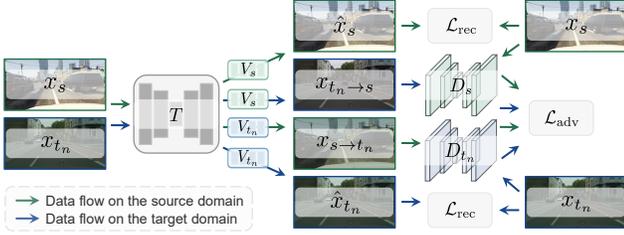


Figure 4. Illustration of our *multi-target style transfer network* (MT-STN). Given a set of labeled images X_s from the source domain and unlabeled images $\mathcal{X}_t = \{X_{t_n}\}_{n=1}^N$ from multiple target domains, the style transfer network T learns to either reconstruct, guided by the reconstruction loss \mathcal{L}_{rec} , or transfer the style of the input image to another domain, guided by the adversarial loss \mathcal{L}_{adv} , depending on the style parameters V that are plugged into T as shown in Figure 5. There is one discriminator $D_s, \mathcal{D}_t = \{D_{t_n}\}_{n=1}^N$ and one set of learnable style parameters $V_s, \mathcal{V}_t = \{V_{t_n}\}_{n=1}^N$ for each domain. We use $x_{a \rightarrow b}$ to denote the transferred image from domain a to b . Learning is unsupervised. Only one target domain is shown for brevity.

network can then transfer the style from one domain to another, simply by plugging in the target style, while maintaining the content of the original image.

Formally, we represent a style as $V = \{\gamma, \beta\}$, where $\gamma, \beta \in \mathbb{R}^d$ are scaling and shifting parameters in a feature space of dimension d . We denote by V_s the source domain style and by $\mathcal{V}_t = \{V_{t_n}\}_{n=1}^N$ the target domain styles. Given a style V , the style transfer network T maps an image $x \in [0, 1]^K$ to another image $T(x, V) \in [0, 1]^K$. We write $T_V(x) = T(x, V)$ for brevity. Figure 5 illustrates the architecture of T , containing a series of *conditional instance normalization* (CIN) layers [5], all taking the same style as input. Given an intermediate feature map f of T , the CIN operation with style $V = \{\gamma, \beta\}$ is defined as

$$\text{CIN}(f, V) = \gamma \left(\frac{f - \mu(f)}{\sigma(f)} \right) + \beta, \quad (14)$$

where $\mu(f)$ and $\sigma(f)$ are the mean and standard deviation over spatial dimensions independently for each channel in f , and all operations are element-wise. Importantly, the parameters $V = \{\gamma, \beta\}$ used in (14) are independent of the network T , which can transfer from one style to another simply by switching V . The way we learn $\{\gamma, \beta\}$ differs from CIN, which learns each style from a single image, using a style loss on that image [5]. Instead, we aim to learn each style from all training images of one domain, and we achieve this by an adversarial loss.

To maintain the content of the input image for each domain, we define the reconstruction loss

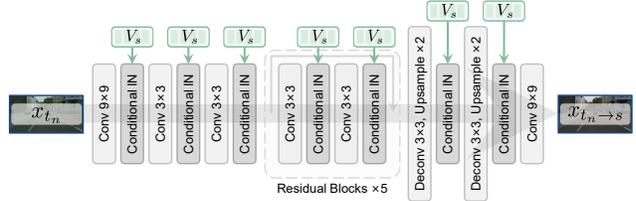


Figure 5. Architecture of style transfer network T in our MT-STN. Domain style parameters V are plugged into T as parameters of a series of *conditional instance normalization* (CIN) layers. Here, input image x_{t_n} from target domain X_{t_n} is transferred to the style V_s of source domain X_s , denoted as $x_{t_n \rightarrow s} = T(x_{t_n}, V_s)$. More examples shown in Figure 4.

$$\mathcal{L}_{rec}(X_s, \mathcal{X}_t, T, V_s, \mathcal{V}_t) = \mathbb{E}_{x \sim X_s} \ell_{rec}(x, T_{V_s}) + \sum_{n=1}^N \mathbb{E}_{x \sim X_{t_n}} \ell_{rec}(x, T_{V_{t_n}}), \quad (15)$$

where, given an image x and a mapping F ,

$$\ell_{rec}(x, F) = \|x - F(x)\|_1. \quad (16)$$

To achieve style transfer between different domains, we define a discriminator for each domain. We denote by D_s the discriminator for the source domain and by $\mathcal{D}_t = \{D_{t_n}\}_{n=1}^N$ the discriminators for the target domains. We then formulate an adversarial loss across domains

$$\mathcal{L}_{adv}(X_s, \mathcal{X}_t, T, V_s, \mathcal{V}_t, D_s, \mathcal{D}_t) = \mathcal{L}^+(X_s, \text{id}, D_s) + \sum_{n=1}^N \mathcal{L}^-(X_{t_n}, T_{V_s}, D_s) + \sum_{n=1}^N (\mathcal{L}^+(X_{t_n}, \text{id}, D_{t_n}) + \mathcal{L}^-(X_s, T_{V_{t_n}}, D_{t_n})), \quad (17)$$

where the positive and negative loss $\mathcal{L}^+, \mathcal{L}^-$ are defined in (7), (8) and id is the identity function. That is, original images of a domain are treated as positive by the discriminator of that domain (first and third term), while images with style transferred to a domain are treated as negative by the discriminator of that domain (second and fourth term). The complete objective function is

$$\min_{T, V_s, \mathcal{V}_t} \max_{D_s, \mathcal{D}_t} \mathcal{L}_{rec} + \lambda_{adv} \mathcal{L}_{adv}, \quad (18)$$

where λ_{adv} is the weighting factor for the adversarial loss.

5. Experiments

5.1. Datasets and metrics

Four benchmark urban driving datasets are adopted in our experiments, including one synthetic dataset

(GTA5 [22]) and three real-world datasets (CityScapes [3], Indian Driving (IDD) [27], and Mapillary [20]).

GTA5 contains 24,966 high-quality labeled frames from the realistic open-world computer games Grand Theft Auto V (GTA5). Each frame is generated from the fictional city Los Santos, based on Los Angeles in Southern California.

CityScapes contains real-world vehicle-ego-centric images collected from 50 cities in Germany and its surrounding countries. It is split into training and validation sets of 2,975 and 500 examples respectively.

IDD is a diverse street-view dataset that captures unstructured traffic on roads in India. It is split into training and validation sets of 6,993 and 981 examples respectively.

Mapillary is a street-view dataset containing high-resolution images collected from all over the world and diverse imaging devices. It is split into training and validation sets of 18,000 and 2,000 examples respectively.

For fair comparisons, we follow the same label mapping protocol used in [14, 24] and standardize the label set with 7 shared super classes among all four datasets, including *flat*, *construction*, *object*, *nature*, *sky*, *human*, and *vehicle*. When CityScapes, IDD, or Mapillary are used as target domains, only unlabeled images are used at training according to the UDA setting, while the evaluation is conducted with the corresponding labeled validation set.

We report quantitative segmentation results using per-class IoU, mean intersection-over-union (mIoU) over the 7 shared super classes, and the average mIoU over different target domains.

5.2. Implementation details

Following [14, 24, 28], we use DeepLab-v2 [1] with the ResNet-101 [8] network pre-trained on ImageNet [4] as the segmentation model for both the student F_S and the teacher F_T for fair comparisons. The discriminator D_{out} has a DC-GAN [21] architecture with 5 convolutional layers of kernel 4×4 and stride of 2. The EMA parameter α in (1) is set to 0.999. The loss factors λ_{con} and λ_{out} in (9) and (13) are set to 100 and 10^{-3} , respectively.

The data augmentation function \mathcal{A} in (4) is implemented with the CutMix [6] strategy and the proposed visual style transfer network T . For each target-domain input image $x_{t_n} \in X_{t_n}$, we first transfer its visual style to the source domain with $x_{t_n \rightarrow s} = T(x_{t_n}, V_s)$. Then, we use CutMix to generate a mixed example from two transferred target-domain examples. Finally, the teacher predictions for the original two target-domain examples are mixed to produce a pseudo label for the student prediction of the mixed example. For function \mathcal{A} in (10), the implementation is the same except that we do not perform visual style transfer on the unseen target domain.

Table 2. Quantitative cross-domain semantic segmentation results from GTA5 (G) to CityScapes (C) and IDD (I) datasets.

METHOD	FLOW	TARGET		EXTERN	<i>flat</i>	<i>constr</i>	<i>object</i>	<i>nature</i>	<i>sky</i>	<i>human</i>	<i>vehicle</i>	mIoU	AVG
		C	I										
URMA [23] (source-free)	G→C	C	✗		91.1	78.9	26.1	80.7	74.6	60.9	67.7	68.6	67.1
	G→I	I	✗		93.0	52.9	15.8	78.5	90.4	54.8	74.6	65.7	
AdvStyle [31] (domain gen.)	G→C	C	✗		87.2	71.8	25.5	82.2	81.0	59.9	79.2	69.5	67.2
	G→I	I	✗		88.2	49.9	13.4	77.9	90.9	55.9	78.5	64.9	
AdvEnt [28] (single-target)	G→C	C	✓		93.5	80.5	26.0	78.5	78.5	55.1	76.4	69.8 (*)	66.5
	G→C	I	✗		91.3	52.3	13.3	76.1	88.7	46.7	74.8	63.3 _{11.8}	
	G→I	C	✗		78.6	79.2	24.8	77.6	83.6	48.7	44.8	62.5 _{17.3}	
	G→I	I	✓		91.2	53.1	16.0	78.2	90.7	47.9	78.9	65.1 (*)	
MT-KD (single-target)	G→C	C	✓		95.9	85.5	40.2	84.8	81.4	64.1	82.2	76.3 _{16.5}	72.4
	G→C	I	✗		92.5	58.3	19.2	79.3	91.8	56.9	81.6	68.5 _{13.4}	
	G→I	C	✗		95.3	83.7	35.9	83.9	78.5	64.7	79.9	74.5 _{14.7}	
	G→I	I	✓		94.2	58.3	25.0	82.9	92.8	61.6	85.3	71.4 _{16.3}	
AdvEnt [28] (multi-target)	G→{C,I}	C	✓		93.9	80.2	26.2	79.0	80.5	52.5	78.0	70.0 _{10.2}	67.4
	G→{C,I}	I	✓		91.8	54.5	14.4	76.8	90.3	47.5	78.3	64.8 _{10.3}	
MTKT [24] (multi-target)	G→{C,I}	C	✓		94.5	82.0	23.7	80.1	84.0	51.0	77.6	70.4 _{10.6}	68.2
	G→{C,I}	I	✓		91.4	56.6	13.2	77.3	91.4	51.4	79.9	65.9 _{10.8}	
ADAS [14] (multi-target)	G→{C,I}	C	✓		95.1	82.6	39.8	84.6	81.2	63.6	80.7	75.4 _{15.6}	71.2
	G→{C,I}	I	✓		90.5	63.0	22.2	73.7	87.9	54.3	76.9	66.9 _{11.8}	
MT-KD (multi-target)	G→{C,I}	C	✓		96.2	85.3	40.3	85.1	80.1	65.2	83.6	76.5 _{16.7}	73.8
	G→{C,I}	I	✓		94.1	60.3	23.2	82.7	92.7	60.3	85.3	71.2 _{16.1}	
UT-KD (multi-target)	(G→I)→C	C	✗		97.0	84.7	41.2	85.1	81.8	64.3	85.2	77.0 _{17.2}	73.7
	(G→C)→I	I	✗		92.7	59.1	24.5	79.3	91.9	61.0	85.0	70.5 _{15.4}	

Bold: best IoU (%) over all methods in each target domain. **Green / red**: mIoU gain / loss w.r.t. the corresponding per-target baseline, marked by *****. **EXTERN**: using external data from the source or other target domains.

Table 3. Quantitative cross-domain semantic segmentation results from GTA5 (G) to CityScapes (C) and Mapillary (M) datasets.

METHOD	FLOW	TARGET		EXTERN	<i>flat</i>	<i>constr</i>	<i>object</i>	<i>nature</i>	<i>sky</i>	<i>human</i>	<i>vehicle</i>	mIoU	AVG
		C	M										
URMA [23] (source-free)	G→C	C	✗		91.1	78.9	26.1	80.7	74.6	60.9	67.7	68.6	69.5
	G→M	M	✗		88.3	71.3	39.0	72.9	90.4	56.5	74.5	70.4	
AdvStyle [31] (domain gen.)	G→C	C	✗		87.2	71.8	25.5	82.2	81.0	59.9	79.2	69.5	70.2
	G→M	M	✗		87.5	70.9	33.4	72.8	90.9	62.1	79.1	70.9	
AdvEnt [28] (single-target)	G→C	C	✓		93.5	80.5	26.0	78.5	78.5	55.1	76.4	69.8 (*)	66.6
	G→C	M	✗		86.8	69.0	30.2	71.2	91.5	35.3	59.5	63.4 _{14.2}	
	G→M	C	✗		89.3	79.3	19.5	76.9	84.6	47.7	63.0	65.8 _{14.0}	
	G→M	M	✓		89.5	72.6	31.0	75.3	94.1	50.7	73.8	69.6 (*)	
MT-KD (single-target)	G→C	C	✓		95.9	85.5	40.2	84.8	81.4	64.1	82.2	76.3 _{16.5}	75.7
	G→C	M	✗		89.7	76.2	44.1	75.5	94.1	63.0	83.3	75.1 _{15.5}	
	G→M	C	✗		96.6	84.5	37.7	84.7	80.5	61.8	82.8	75.5 _{15.7}	
	G→M	M	✓		90.0	76.4	47.5	74.1	93.7	60.1	84.6	75.2 _{15.6}	
AdvEnt [28] (multi-target)	G→{C,M}	C	✓		93.1	80.5	24.0	77.9	81.0	52.5	75.0	69.1 _{10.7}	68.9
	G→{C,M}	M	✓		90.0	71.3	31.1	73.0	92.6	46.6	76.6	68.7 _{10.9}	
MTKT [24] (multi-target)	G→{C,M}	C	✓		95.0	81.6	23.6	80.1	83.6	53.7	79.8	71.1 _{11.3}	70.9
	G→{C,M}	M	✓		90.6	73.3	31.0	75.3	94.5	52.2	79.8	70.8 _{11.2}	
ADAS [14] (multi-target)	G→{C,M}	C	✓		96.4	83.5	35.1	83.8	84.9	62.3	81.3	75.3 _{15.5}	73.9
	G→{C,M}	M	✓		88.6	73.7	41.0	75.4	93.4	58.5	77.2	72.8 _{13.0}	
MT-KD (multi-target)	G→{C,M}	C	✓		96.3	85.6	39.8	85.5	82.5	64.5	83.5	76.8 _{17.0}	76.0
	G→{C,M}	M	✓		89.9	76.7	46.3	73.5	93.2	63.8	84.1	75.3 _{15.7}	
UT-KD (multi-target)	(G→M)→C	C	✗		96.6	84.7	43.1	85.4	82.8	62.6	82.9	76.8 _{17.0}	75.9
	(G→C)→M	M	✗		90.1	75.2	46.7	76.2	94.4	60.1	82.9	75.1 _{15.5}	

Bold: best IoU (%) over all methods in each target domain. **Green / red**: mIoU gain / loss w.r.t. the corresponding per-target baseline, marked by *****. **EXTERN**: using external data from the source or other target domains.

5.3. Synthetic-to-real adaptation

We first evaluate the performance of the proposed methods against existing approaches in the synthetic-to-real adaptation scenario, where the labeled GTA5 dataset is adopted as the source domain and the unlabeled CityScapes, IDD, and Mapillary datasets are used as the multi-target domains. Results are reported in Tables 2, 3, and 4. It can be observed that AdvEnt [28] trained with the single-target domain adaptation setting generally yields lower mIoU scores compared to its counterpart in the multi-target domain adap-

Table 4. Quantitative cross-domain semantic segmentation results from GTA5 (G) to CityScapes (C), IDD (I), and Mapillary (M) datasets.

METHOD	FLOW	TARGET EXTERN									mIoU	AVG
			flat	constr.	object	nature	sky	human	vehicle			
URMA [23] (source-free)	G→C	C	✗	91.1	78.9	26.1	80.7	74.6	60.9	67.7	68.6	68.2
	G→I	I	✓	93.0	52.9	15.8	78.5	90.4	54.8	74.6	65.7	
	G→M	M	✗	88.3	71.3	39.0	72.9	90.4	56.5	74.5	70.4	
AdvStyle [31] (domain gen.)	G→C	C	✗	87.2	71.8	25.5	82.2	81.0	59.9	79.2	69.5	68.4
	G→I	I	✓	88.2	49.9	13.4	77.9	90.9	55.9	78.5	64.9	
	G→M	M	✗	87.5	70.9	33.4	72.8	90.9	62.1	79.1	70.9	
AdvEnt [28] (single-target)	G→C	C	✓	93.5	80.5	26.0	78.5	78.5	55.1	76.4	69.8 (*)	65.5
	G→C	I	✗	91.3	52.3	13.3	76.1	88.7	46.7	74.8	63.3 _{11.8}	
	G→C	M	✗	86.8	69.0	30.2	71.2	91.5	35.3	59.5	63.4 _{16.2}	
	G→I	C	✗	78.6	79.2	24.8	77.6	83.6	48.7	44.8	62.3 _{17.3}	65.5
	G→I	I	✓	91.2	53.1	16.0	78.2	90.7	47.9	78.9	65.1 (*)	
	G→I	M	✗	88.5	71.2	32.4	72.8	92.8	51.3	73.7	69.0 _{10.6}	
	G→M	C	✗	89.3	79.3	19.5	76.9	84.6	47.7	63.0	65.8 _{14.0}	66.7
	G→M	I	✓	91.7	54.3	13.0	77.3	92.3	47.4	76.8	64.7 _{10.4}	
	G→M	M	✓	89.5	72.6	31.0	75.3	94.1	50.7	73.8	69.6 (*)	
MT-KD (single-target)	G→C	C	✓	95.9	85.5	40.2	84.8	81.4	64.1	82.2	76.3 _{16.5}	73.3
	G→C	I	✗	92.5	58.3	19.2	79.3	91.8	56.9	81.6	68.5 _{13.4}	
	G→C	M	✗	89.7	76.2	44.1	75.5	94.1	63.0	83.3	75.1 _{15.5}	
	G→I	C	✓	95.3	83.7	35.9	83.9	78.5	64.7	79.9	74.5 _{14.7}	73.4
	G→I	I	✓	94.2	58.3	25.0	82.9	92.8	61.6	85.3	71.4 _{16.3}	
	G→I	M	✗	89.9	75.6	42.9	74.7	93.8	60.8	82.6	74.3 _{14.7}	
	G→M	C	✗	96.6	84.5	37.7	84.7	80.5	61.8	82.8	75.5 _{15.7}	73.6
	G→M	I	✓	94.4	58.1	26.1	81.6	92.2	56.8	81.7	70.1 _{15.0}	
	G→M	M	✓	90.0	76.4	47.5	74.1	93.7	60.1	84.6	75.2 _{15.6}	
AdvEnt [28] (multi-target)	G→{C, I, M}	C	✓	93.6	80.6	26.4	78.1	81.5	51.9	76.4	69.8 -	67.8
	G→{C, I, M}	I	✓	92.0	54.6	15.7	77.2	90.5	50.8	78.6	65.6 _{10.5}	
	G→{C, I, M}	M	✓	89.2	72.4	32.4	73.0	92.7	41.6	74.9	68.0 _{11.6}	
MKT [24] (multi-target)	G→{C, I, M}	C	✓	94.6	80.7	23.8	79.0	84.5	51.0	79.2	70.4 _{10.6}	69.1
	G→{C, I, M}	I	✓	91.7	55.6	14.5	78.0	92.6	49.8	79.4	65.9 _{10.8}	
	G→{C, I, M}	M	✓	90.5	73.7	32.5	75.5	94.3	51.2	80.2	71.1 _{11.5}	
ADAS [14] (multi-target)	G→{C, I, M}	C	✓	95.8	82.4	38.3	82.4	85.0	60.5	80.2	74.9 _{15.1}	71.3
	G→{C, I, M}	I	✓	89.9	52.7	25.0	78.1	92.1	51.0	77.9	66.7 _{11.6}	
	G→{C, I, M}	M	✓	89.9	76.5	46.9	73.4	93.2	56.1	75.4	72.2 _{12.6}	
MT-KD (multi-target)	G→{C, I, M}	C	✓	95.3	85.6	39.7	84.5	82.3	65.5	81.4	76.3 _{16.5}	74.1
	G→{C, I, M}	I	✓	93.9	59.7	22.8	82.1	92.7	60.3	84.6	70.8 _{15.7}	
	G→{C, I, M}	M	✓	89.9	76.5	46.9	73.4	93.2	63.8	84.2	75.4 _{15.8}	
MT-KD [†] (multi-target)	G→{I, M}	C	✗	96.7	84.3	38.2	84.7	78.9	64.6	84.3	75.9 _{16.1}	73.5
	G→{C, M}	I	✗	93.9	58.6	22.7	81.4	91.7	57.7	82.0	69.7 _{14.6}	
	G→{C, I}	M	✗	89.7	76.3	44.1	75.4	94.1	63.0	83.4	75.1 _{15.5}	
UT-KD (multi-target)	(G→{I, M})→C	C	✗	97.0	85.0	41.7	85.5	81.9	65.1	84.9	77.3 _{17.5}	75.0
	(G→{C, M})→I	I	✗	95.0	58.9	30.6	83.8	91.5	60.7	85.0	72.2 _{17.1}	
	(G→{C, I})→M	M	✗	89.8	74.0	46.4	76.6	94.4	64.5	84.2	75.7 _{16.1}	

Bold: best IoU (%) over all methods in each target domain. **Green / red:** mIoU gain / loss w.r.t. the corresponding per-target baseline, marked by ‘*’. **EXTERN:** using external data from the source or other target domains. **MT-KD[†]:** direct transfer from a pre-trained MT-KD model to an unseen target domain.

tation setting, which demonstrates the advantage of incorporating multi-domain data into training.

In all multi-target domain adaptation scenarios, the proposed method MT-KD obtains the highest mIoU scores and outperforms the existing state-of-the-art methods like ADAS [14] by a large margin, more than 2% mIoU. Qualitative adaptation results of MT-KD from GTA5 to CityScapes, IDD, and Mapillary are shown in Figure 6.

Another intriguing finding is that the proposed method UT-KD yields very competitive performance compared to MT-KD, although it does not access any external data. In Table 4 for example, MT-KD yields 74.1% averaged mIoU and UT-KD yields 75.0%, even outperforming MT-KD by 0.9% and ADAS [14] by 3.7%. In Table 2 and Table 3, it is nearly on par with MT-KD, losing only by 0.1%, and still outperforms ADAS [14] by 2.5% and 2% respectively. Considering that in real-world scenarios, it is more common to get access to pre-trained models than to complete street-view datasets collected from other cities because of data privacy, our UT-KD is more flexible and more practi-

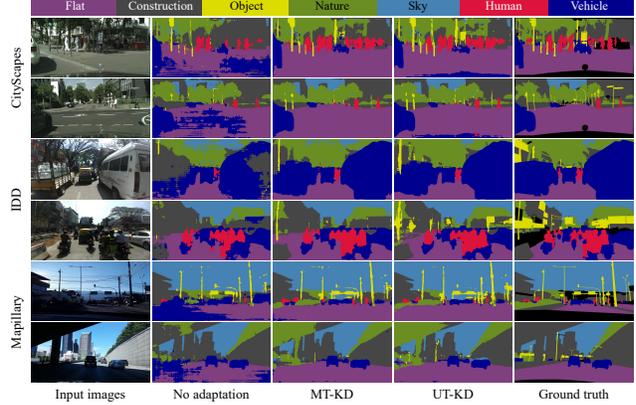


Figure 6. Qualitative cross-domain semantic segmentation results from GTA5 to CityScapes, IDD, and Mapillary datasets.

cal without losing on performance.

6. Conclusion

In this paper, we introduce a new strategy for conducting multi-target unsupervised domain adaptation for semantic segmentation without relying on external data. To implement this idea, we first propose the multi-target knowledge distillation (MT-KD) method, which achieves multi-target UDA for semantic segmentation using adversarial learning and self-distillation, setting new state-of-the-art performance. As a simplified version, we further propose the unseen target knowledge distillation (UT-KD) method, which rapidly adapts a pre-trained MT-KD model to a new unseen target domain through “one-way” adversarial learning, without accessing any external data from the source or other target domains. Despite its simplicity, UT-KD is more scalable than existing multi-target UDA solutions in handling unseen domains, especially under data privacy constraints. It does not compromise performance compared to MT-KD and still outperforms other state-of-the-art methods. To further address the visual appearance shift, we perform visual style transfer across multiple domains by parameterizing the style of each domain through a single vector, thus decoupling it from the style transfer process itself. The latter is accomplished by a multi-target style transfer network (MT-STN), which is shared across all domains.

Although the proposed methods are originally designed for the cross-domain semantic segmentation task, they may also be helpful for solving other cross-domain tasks. We will explore it in future work.

Acknowledgment

The authors would like to thank the Institute of Advanced Research in Artificial Intelligence (IARAI) for its support.

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 Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017. 1, 7
- [2] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. RobustNet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 2021. 3
- [3] 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 *CVPR*, 2016. 7
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 7
- [5] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *ICLR*, 2017. 6
- [6] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*, 2020. 4, 7
- [7] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *ICLR*, 2018. 10
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7
- [9] Jiaying Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *NeurIPS*, 2021. 3
- [10] Jiaying Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. Category contrast for unsupervised domain adaptation in visual tasks. In *CVPR*, 2022. 1
- [11] Takashi Isobe, Xu Jia, Shuaijun Chen, Jianzhong He, Yongjie Shi, Jianzhuang Liu, Huchuan Lu, and Shengjin Wang. Multi-target domain adaptation with collaborative consistency learning. In *CVPR*, 2021. 2, 5
- [12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 10
- [13] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *ICCV*, 2021. 3
- [14] Seunghun Lee, Wonhyeok Choi, Changjae Kim, Minwoo Choi, and Sunghoon Im. ADAS: A direct adaptation strategy for multi-target domain adaptive semantic segmentation. In *CVPR*, 2022. 2, 5, 7, 8, 10
- [15] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. WildNet: Learning domain generalized semantic segmentation from the wild. In *CVPR*, 2022. 1, 3
- [16] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *ICCV*, 2019. 2
- [17] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, 2021. 3
- [18] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *ICCV*, 2019. 2
- [19] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019. 2
- [20] Gerhard Neuhof, Tobias Ollmann, Samuel Rota Buló, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 7
- [21] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 4, 7
- [22] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 7
- [23] Prabhu Teja S and Francois Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *CVPR*, 2021. 3, 7, 8, 10
- [24] Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Multi-target adversarial frameworks for domain adaptation in semantic segmentation. In *ICCV*, 2021. 2, 5, 7, 8, 10
- [25] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 4, 10
- [26] 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. In *CVPR*, 2018. 2, 5
- [27] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *WACV*, 2019. 7
- [28] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 7, 8, 10
- [29] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017. 2
- [30] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *CVPR*, 2018. 2
- [31] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. In *NeurIPS*, 2022. 3, 7, 8, 10

Supplementary Material

In this supplementary material, we provide more training details as well as quantitative and qualitative results.

A. Training details

We use stochastic gradient descent (SGD) with a learning rate of 2.5×10^{-5} to train F_S , while D_{out} is trained by the Adam optimizer [12] with a learning rate of 10^{-5} and $\beta_1 = 0.9$, $\beta_2 = 0.99$. For both optimizers, we set a weight decay of 5×10^{-5} and adopt the ‘‘poly’’ learning rate decay schedule, where the initial learning rate is multiplied by $(1 - i/I)^p$ with $p = 0.9$, where i is the current iteration and I the total number of iterations, set to 50,000.

To train MT-STN, we use the Adam optimizer for 20 training epochs with weight decay 5×10^{-5} and learning rate 2.5×10^{-4} and 10^{-5} for the generator and discriminators, respectively. Each mini-batch consists of one source-domain image and one target-domain image. The loss factor λ_{adv} in (17) is empirically set to 10^{-3} .

At inference, we use the teacher network F_T for MT-KD and F'_T for UT-KD as obtained at the end of training to perform semantic segmentation of input test images.

B. Real-to-real adaptation

We further conduct experiments on a real-to-real adaptation setting, where the labeled dataset CityScapes is adopted as the source domain and the unlabeled datasets IDD and Mapillary are used as the multi-target domains. As can be observed from Table 5, our MT-KD achieves again the best multi-target domain adaptation performance compared to existing approaches. UT-KD yields 74.6% averaged mIoU, which is again higher than MT-KD by 0.7% and ADAS [14] by 1.9%, despite not having access to external data. These results confirm the high practical value of our new UDA strategy without external data.

C. Ablation study

MT-KD Table 6 shows how loss factors λ_{out} and λ_{con} affect performance. MT-KD in general can tolerate a wide range of λ_{out} and is more sensitive to λ_{con} . Based on these results, we empirically set $\lambda_{\text{out}} = 10^{-3}$ and $\lambda_{\text{con}} = 100$.

Table 7 shows the contribution of each component in MT-KD performance. We find that adversarial learning alone cannot bring about satisfactory performance. By contrast, combining adversarial learning and self-distillation brings significant improvement.

Table 8 shows the effect of different augmentation strategies for self-distillation in MT-KD. While Gaussian noise is common [7, 25], we find that CutMix is superior in cross-domain semantic segmentation. In addition, our MT-STN brings further improvement by directly reducing the visual

Table 5. Quantitative cross-domain semantic segmentation results from CityScapes (C) to IDD (I) and Mapillary (M) datasets.

METHOD	FLOW	TARGET	EXTERN	β_{out}	const.	object	nature	sky	human	vehicle	mIoU	AVG
URMA [23] (source-free)	C→I	I	✗	93.9	56.0	23.4	83.7	93.6	52.0	79.2	68.8	68.4
	C→M	M	✗	88.1	71.6	26.5	70.8	92.2	56.5	70.9	68.1	
AdvStyle [31] (domain gen.)	C→I	I	✗	93.9	52.9	18.6	82.9	92.6	51.2	76.9	67.0	68.8
	C→M	M	✗	89.5	70.2	34.4	77.3	93.1	56.6	73.7	70.6	
AdvEnt [28] (single-target)	C→I	I	✓	93.2	53.4	16.5	83.4	93.4	51.4	79.5	67.3 (*)	68.0
	C→I	M	✗	88.2	70.0	28.5	75.4	93.6	49.1	76.7	68.8 ^{+2.2}	
	C→M	I	✗	91.8	52.2	15.9	80.2	91.1	45.7	77.6	65.0 ^{+2.3}	
	C→M	M	✓	87.4	65.9	28.2	72.8	92.1	46.9	72.7	66.6 (*)	
MT-KD (single-target)	C→I	I	✓	93.7	59.2	29.8	83.6	93.3	62.1	85.3	72.4 ^{+5.1}	74.0
	C→I	M	✗	90.3	75.0	46.2	77.6	94.2	63.9	82.3	75.6 ^{+9.0}	
	C→M	I	✗	95.1	58.0	28.7	84.8	92.6	57.7	81.8	71.2 ^{+3.9}	
	C→M	M	✓	89.6	73.4	47.9	75.2	93.5	62.8	84.1	75.2 ^{+8.6}	
AdvEnt [28] (multi-target)	C→(I, M)	I	✓	93.3	53.0	17.2	82.8	92.2	49.3	79.6	66.8 ^{+0.5}	67.0
	C→(I, M)	M	✓	87.7	65.9	29.0	73.2	91.5	47.9	75.7	67.3 ^{+0.7}	
MTKT [24] (multi-target)	C→(I, M)	I	✓	93.6	54.9	18.6	84.0	94.5	53.4	79.2	68.9 ^{+1.0}	69.0
	C→(I, M)	M	✓	88.3	70.4	31.6	75.9	94.4	50.9	77.0	69.8 ^{+3.2}	
ADAS [14] (multi-target)	C→(I, M)	I	✓	-	-	-	-	-	-	-	70.4 ^{+3.1}	72.7
	C→(I, M)	M	✓	-	-	-	-	-	-	-	75.1 ^{+8.5}	
MT-KD (multi-target)	C→(I, M)	I	✓	93.0	60.8	29.4	80.9	92.6	62.3	85.3	72.0 ^{+4.7}	73.9
	C→(I, M)	M	✓	90.3	75.5	48.7	75.3	93.6	63.2	84.7	75.9 ^{+9.3}	
UT-KD (multi-target)	(C→M)→I	I	✗	95.4	59.6	32.7	86.4	94.5	58.3	84.0	72.9 ^{+5.6}	74.6
	(C→I)→M	M	✗	90.5	75.9	47.0	77.9	95.1	63.8	84.7	76.4 ^{+9.8}	

Bold: best IoU (%) over all methods in each target domain. **Green / red:** mIoU gain / loss w.r.t. the corresponding per-target baseline, marked by ‘*’. EXTERN: using external data from the source or other target domains.

Table 6. Parameter analysis of λ_{out} and λ_{con} in MT-KD from GTA5 (G) to CityScapes (C) and IDD (I) datasets.

λ_{out}	10^{-4}	5×10^{-4}	10^{-3}	10^{-2}	λ_{con}	1	10	100	150
C	76.0	76.6	76.5	76.3	C	73.4	72.5	76.5	76.2
I	70.7	70.9	71.2	69.9	I	68.5	69.9	71.2	70.1

Bold: best mIoU (%) scores in each target domain.

Table 7. Ablation study of MT-KD from GTA5 (G) to CityScapes (C), IDD (I), and Mapillary (M) datasets.

METHOD	\mathcal{L}_{CE}	\mathcal{L}_{out}	\mathcal{L}_{con}	C	I	M	Avg.
No adaptation	✓			63.7	64.4	69.4	65.8
Adversarial learning	✓	✓		72.8	67.5	71.9	70.7
Self-distillation	✓		✓	75.7	69.1	74.7	73.1
MT-KD	✓	✓	✓	76.3	70.8	75.4	74.1

Bold: best mIoU (%) scores in each target domain.

Table 8. Comparison of different augmentation strategies for self-distillation in MT-KD from GTA5 (G) to CityScapes (C), IDD (I), and Mapillary (M) datasets.

METHOD	C	I	M	Avg.
No augmentation	73.1	66.7	72.1	70.6
Gaussian noise w/o MT-STN	73.3	66.8	72.7	70.9
Gaussian noise w/ MT-STN	73.2	67.9	73.1	71.4
CutMix w/o MT-STN	76.6	69.4	74.9	73.6
CutMix w/ MT-STN	76.3	70.8	75.4	74.1

Bold: best mIoU (%) scores in each target domain.

appearance shift between different domains. The combination of the two strategies brings an overall improvement of 3.5% average mIoU compared with no augmentation.

Figure 7 and Table 9 show how EMA decay parameter α affects the performance of MT-KD. As the parameter α

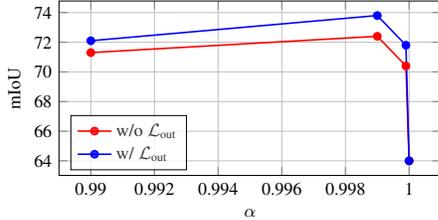


Figure 7. MT-KD average mIoU vs. EMA decay parameter α on GTA5 \rightarrow {CityScapes, IDD}.

Table 9. Effect of EMA decay parameter α on MT-KD from GTA5 (G) to CityScapes (C) and IDD (I) datasets.

α	0	0.5	0.9	0.99	0.999	0.9999	1
w/o \mathcal{L}'_{out}	C	66.9	54.8	74.3	75.1	75.7	73.1
	I	57.2	61.1	66.8	67.5	69.1	67.7
	Avg.	62.0	57.9	70.5	71.3	72.4	70.4
w/ \mathcal{L}'_{out}	C	34.5	32.3	60.4	75.5	76.5	74.2
	I	47.5	21.7	59.5	68.7	71.2	69.4
	Avg.	41.0	27.0	59.9	72.1	73.8	71.8

Bold: best mIoU (%) scores in each target domain.

Table 10. Ablation study of UT-KD from GTA5 (G) to CityScapes (C) and IDD (I) datasets.

METHOD	\mathcal{L}'_{out}	\mathcal{L}'_{con}	(G \rightarrow I) \rightarrow C	(G \rightarrow C) \rightarrow I	Avg.
No adaptation			74.5	68.5	71.5
Adversarial learning	✓		63.9	64.8	64.3
Self-distillation		✓	76.1	69.6	72.8
UT-KD	✓	✓	77.0	70.5	73.7

Bold: best mIoU (%) scores in each target domain.

Table 11. Comparison of different augmentation strategies for self-distillation in UT-KD from GTA5 (G) to CityScapes (C) and IDD (I) datasets.

METHOD	(G \rightarrow I) \rightarrow C	(G \rightarrow C) \rightarrow I	Avg.
No augmentation	75.8	68.8	72.3
Gaussian noise	75.9	68.0	71.9
CutMix	77.0	70.5	73.7

Bold: best mIoU (%) scores in each target domain.

approaches 1, the mIoU values increases and then drops sharply for $\alpha > 0.999$. Based on these results, we empirically set $\alpha = 0.999$.

UT-KD Table 10 shows the contribution of each component in UT-KD performance. An intriguing phenomenon is that adversarial learning alone is harmful. A possible explanation is that it needs the assistance of a more stable loss, as is the case of cross-entropy in MT-KD. By contrast, when combined with self-distillation, it further improves performance by 0.9% average mIoU, reaching a total improvement of 2.2% compared with no adaptation.

Table 11 shows the effect of different augmentation strategies for self-distillation in UT-KD. Again, we find that

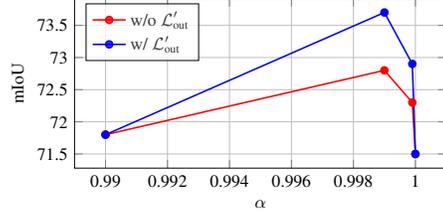


Figure 8. UT-KD average mIoU vs. EMA decay parameter α on GTA5 \rightarrow {CityScapes, IDD}.

CutMix works best.

As described in subsection 4.2, we train UT-KD by performing self-distillation on the unseen target domain data using a consistency loss that minimizes the MSE between the student and teacher predictions, where F'_T is again obtained by EMA on the parameters of F'_S . A simple baseline to achieve this goal is to perform knowledge distillation from a frozen teacher F''_T , as initialized from the pre-trained MT-KD model. Accordingly, we define the *frozen* consistency loss as the MSE between predictions from the student and the frozen teacher

$$\mathcal{L}_{fro}(X_u, F'_S) = \mathbb{E}_{x \sim X_u} \ell''_{con}(\mathcal{A}(x), F'_S) \quad (19)$$

$$\ell''_{con}(x, F'_S) = \frac{1}{K} \sum_{k=1}^K \left\| F'_S(x)^{(k)} - F''_T(x)^{(k)} \right\|^2. \quad (20)$$

Table 12. UT-KD mIoU with and without frozen consistency loss from GTA5 (G) to CityScapes (C) and IDD (I) datasets.

METHOD	\mathcal{L}'_{out}	\mathcal{L}'_{con}	\mathcal{L}_{fro}	(G \rightarrow I) \rightarrow C	(G \rightarrow C) \rightarrow I	Avg.
No adaptation				74.5	68.5	71.5
Adversarial	✓			63.9	64.8	64.3
Self-distillation		✓		76.1	69.6	72.8
UT-KD	✓	✓		77.0	70.5	73.7
Frozen			✓	73.7	66.8	70.2
Adversarial + Frozen	✓		✓	73.9	66.8	70.3
Self-distillation + Frozen		✓	✓	76.0	69.7	72.8
UT-KD + Frozen	✓	✓	✓	76.1	69.9	73.0

Bold: best mIoU (%) scores in each target domain.

Table 12 shows the additional ablation study of UT-KD including this loss. An intriguing phenomenon is that using \mathcal{L}_{fro} (19) alone is harmful, dropping performance by 1.3% average mIoU compared to no adaptation. A possible explanation is that the pseudo label generated by the frozen teacher is not accurate since it is directly initialized with the pre-trained MT-KD model, without refinement from EMA. Another interesting finding is that the adversarial loss \mathcal{L}'_{out} , when combined with \mathcal{L}_{fro} , is not as harmful as when used alone, which confirms its nature as an auxiliary loss. Other than that, all options involving \mathcal{L}_{fro} are inferior to those that do not, and the best option remains $\mathcal{L}'_{out} + \mathcal{L}'_{con}$.

Figure 8 and Table 13 further show how EMA decay parameter α affects the performance of UT-KD. Similar to

Table 13. Effect of EMA decay parameter α on UT-KD from GTA5 (G) to CityScapes (C) and IDD (I) datasets.

α		0	0.5	0.9	0.99	0.999	0.9999	1
w/o \mathcal{L}'_{out}	(G→I)→C	3.3	37.5	68.0	75.7	76.1	75.4	74.5
	(G→C)→I	5.3	24.6	64.9	68.0	69.6	69.2	68.5
	Avg.	4.3	31.0	66.4	71.8	72.8	72.3	71.5
w/ \mathcal{L}'_{out}	(G→I)→C	3.5	35.4	55.3	75.5	77.0	76.2	74.5
	(G→C)→I	5.3	34.1	59.2	68.2	70.5	69.7	68.5
	Avg.	4.4	34.7	57.2	71.8	73.7	72.9	71.5

Bold: best mIoU (%) scores in each target domain.



Figure 9. Visual style transfer results with GTA5 (G), CityScapes (C), IDD (I), and Mapillary (M). Red-boxed images are the original inputs in each domain.

MT-KD, as the parameter α approaches 1, the mIoU values increase and then drop sharply for $\alpha > 0.999$. Thus, we empirically set $\alpha = 0.999$.

MT-STN Figure 9 shows style transfer results between the four datasets using MT-STN. We find that MT-STN can learn the inherent visual style of each domain and perform synthetic-to-real, real-to-synthetic, or real-to-real style transfer between different domains.

D. Additional qualitative results

Figure 10 shows more synthetic-to-real style transfer results from GTA5 to CityScapes, IDD and Mapillary using MT-STN. Figure 11 shows more real-to-real style transfer results from/to CityScapes, IDD, and Mapillary using MT-STN. More qualitative cross-domain semantic segmentation results from GTA5 to CityScapes, IDD, and Mapillary are shown in Figure 12. UT-KD can generally yield competitive or slightly better segmentation results than MT-KD, although it does not access any external data. This is more evident on small objects like traffic signs and poles, as shown in the second row on the IDD dataset.



Figure 10. Synthetic-to-real style transfer results from GTA5 (G) to CityScapes (C), IDD (I), and Mapillary (M). Images in red frames are the original inputs.



Figure 11. Real-to-real style transfer results from/to CityScapes (C), IDD (I), and Mapillary (M). Images in red frames are the original inputs.

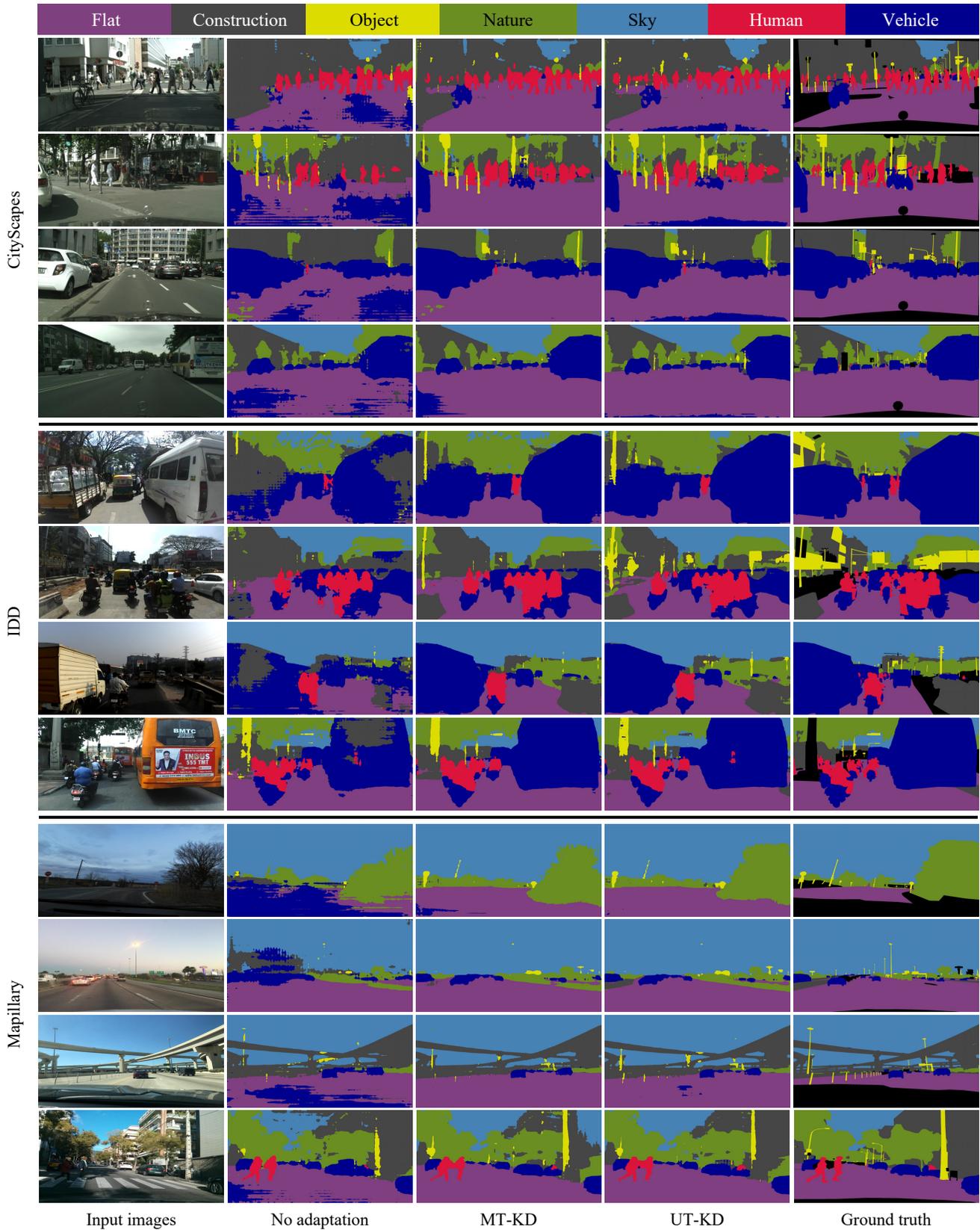


Figure 12. Qualitative cross-domain semantic segmentation results from GTA5 (G) to CityScapes (C), IDD (I), and Mapillary (M) datasets. MT-KD is trained on all three target domains (*i.e.*, $G \rightarrow \{C, I, M\}$), while UT-KD is initialized with the pre-trained MT-KD model on two target domains and then fine-tuned on the third target domain only as unknown (*e.g.*, $(G \rightarrow \{C, I\}) \rightarrow M$ for Mapillary).