TADA: TAXONOMY ADAPTIVE DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditional domain adaptation addresses the task of adapting a model to a novel target domain under limited or no additional supervision. While tackling the input domain gap, the standard domain adaptation settings assume no domain change in the *output* space. In semantic prediction tasks, different datasets are often labeled according to different semantic taxonomies. In many real-world settings, the target domain task requires a different taxonomy than the one imposed by the source domain. We therefore introduce the more general taxonomy adaptive domain adaptation (TADA) problem, allowing for inconsistent taxonomies between the two domains. We further propose an approach that jointly addresses the image-level and label-level domain adaptation. On the label-level, we employ a bilateral mixed sampling strategy to augment the target domain, and a relabelling method to unify and align the label spaces. We address the image-level domain gap by proposing an uncertainty-rectified contrastive learning method, leading to more domain-invariant and class-discriminative features. We extensively evaluate the effectiveness of our framework under different TADA settings: open taxonomy, coarse-to-fine taxonomy, and partially-overlapping taxonomy. Our approach outperforms previous state-of-the-art by a large margin, while capable of adapting to target taxonomies.

1 INTRODUCTION

Approaches for domain adaptation (Ganin & Lempitsky, 2015; Tzeng et al., 2017; Hoffman et al., 2018; Long et al., 2015; Chen et al., 2018; Tsai et al., 2018; Liu et al., 2020) typically focus on *image level* domain gap, which can include visual style, weather, lighting conditions, *etc.*. However, these methods are restricted by the assumption of having consistent taxonomies between source and target domains, *i.e.,* each source domain class can be unambiguously mapped to one target domain class (see Fig. 1 (a-c)), which is often not the case in practice. In many applications, the label spaces of the source and target domains are inconsistent, due to different application scenarios, changeable requirements, inconsistent annotation practices, or the strive towards an increasingly fine-grained taxonomy (Neuhold et al., 2017; Lambert et al., 2020; Cordts et al., 2016).

The aforementioned observations motivate us to consider the *label level* domain gap problem. Even though recent open/universal/class-incremental/zero-shot domain adaptation works (Panareda Busto & Gall, 2017; You et al., 2019; Bucher et al., 2020) touched upon the label level domain gap, they only focus on unseen classes in the target domain. However, label level domain gap in practical scenarios is far more complicated, rather than restricted to unseen classes. We therefore formulate and explore the label level domain gap problem in a more general and complete setting. We identify three typical categories of label taxonomy inconsistency. i) *Open taxonomy*: some classes, *e.g.,* "terrain" in Fig. 1(d), appear in the target domain, but are unlabeled or unseen in the source domain. ii) *Coarse-to-fine taxonomy*: some classes in the source domain, *e.g.,* "person", are divided into several sub-classes in the target domain, *e.g.,* "pedestrain" and "rider' (Fig. 1(e)). iii) *Partially-overlapping taxonomy*: for a certain class in the source domain, one or more of its sub-classes are merged into other classes in the target domain. For example, there exists taxonomic conflict between {"vehicle", "bicycle"} in the source domain and {"car", "cycle"} in the target domain (Fig. 1(f)).

We therefore introduce a more general and challenging domain adaptation problem, namely *taxonomy adaptive domain adaptation* (TADA). In traditional unsupervised domain adaptation (UDA), the goal is to transfer a model learned on a labelled source domain to an unlabelled target domain, under the consistent taxonomy assumption. In contrast, TADA allows for inconsistent taxonomies between a labeled source domain and a few-shot/partially labeled target domain, where the inconsistent classes

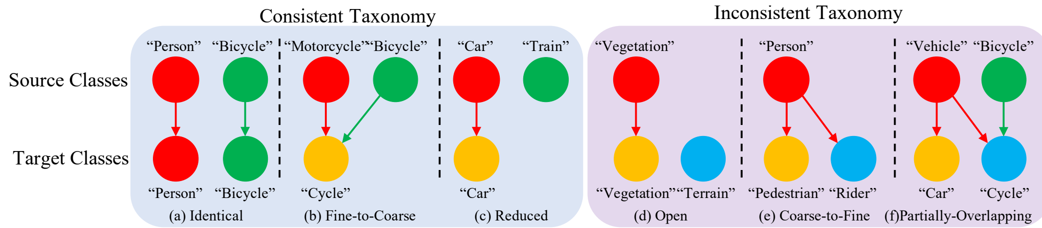

Figure 1: Consistent *v.s.* inconsistent taxonomy. In (a)-(f), the upper row is the source domain classes, and the lower row is the target domain classes. Circles represent classes while an arrow represents a mapping from the source domain class to the target domain class. The (a)-(c) and (d)-(f) are the examples of the consistent and inconsistent taxonomy, respectively.

on the target domain are exemplified by a few labeled samples. Thus TADA approaches domain adaptation over both image and label levels, under the few-shot/partially labeled setting. Such task setting is realistic and involves practical challenges. On the one hand, TADA allows methods to take the full use of the labeled source domain without annotation cost in the target domain for consistent classes. On the other hand, methods are allowed to conduct taxonomy adaptation, with very limited supervision in the target domain side, *i.e.,* only a few samples from the inconsistent classes in the target domain are labeled. In this article, semantic segmentation, raising particular interest in domain adaptation field due to its great potential in autonomous driving, is set as the exemplar task of TADA.

We further put forward the first approach for TADA, addressing both of the label and image level domain gap. For the former, we aim to remedy the gap in label space using pseudo-labelling techniques. First, a *bilateral mixed sampling* strategy is proposed to augment unlabeled images by mixing them with both labeled source-domain and target-domain samples. Second, we map inconsistent source domain labels with a *stochastic label mapping* strategy, which encourages a more flexible taxonomy adaptation during the earlier learning phase. Third, a *pseudo-label based relabeling* strategy is proposed to replace the inconsistent classes in the source-domain according to model's predictions, to further enforce taxonomy adaptation during the training process. To tackle image level domain gap, we introduce an *uncertainty-rectified contrastive learning* scheme that facilitates the learning of class-discriminative and domain-invariant features, under the uncertainty-aware guidance of predicted pseudo-labels. Our complete approach for TADA demonstrates strong results in the different inconsistent taxonomy settings (*i.e.,* open, coarse-to-fine, and partially-overlapping). Moreover, our suggested mixed-sampling and contrastive-learning based scheme outperforms current state-of-the-art methods by a large margin, under traditional UDA setting.

To summarize, the contributions of this paper are three-fold:

- A new problem – *taxonomy adaptive domain adaptation* (TADA) – is proposed and opens a new venue for domain adaptation, *i.e.,* address both image and label level domain gap.
- A generic solution for domain adaptation is proposed, where a set of mixed sampling and pseudo-labelling techniques are developed to reduce the label level domain gap, and a uncertainty-rectified contrastive learning scheme is presented to enable robust cross-domain representation learning.
- Extensive experiments are conducted under the traditional UDA and different TADA settings.

2 Related Work

Domain adaptation: The traditional unsupervised domain adaptation (UDA) (Tsai et al., 2018; Zhang et al., 2017; Hoffman et al., 2016; Ganin & Lempitsky, 2015; Zou et al., 2018; Long et al., 2015) considers the case when the source and target domain share the same label space and where the target domain is unlabeled. However, this setting does not conform with many practical applications. Some recent works have therefore explored alternative settings. The open/universal domain adaptation (Panareda Busto & Gall, 2017; Saito et al., 2018; You et al., 2019) aims at recognizing the new unseen classes in the target domain together as the "unknown" class. The class-incremental/zero-shot domain adaptation (Kundu et al., 2020; Bucher et al., 2020) are proposed to recognize the new unseen classes explicitly and separately on the target domain under the source domain free setting and in the zero-shot segmentation way, resp. These works touch upon the specific case of open taxonomy setting in TADA. However, the above works only consider the case where the unseen classes are absent in the source domain. In contrast, the open taxonomy setting in TADA also allows for the unseen classes to exist in the source domain, where they are unlabelled. Besides, the above works do not consider about the coarse-to-fine and partially-overlapping taxonomy problems, which are covered by the more general TADA formulation. Recent few-shot/semi-supervised domain adaptation works (Teshima et al., 2020; Motiian et al., 2017; Zhang et al., 2019) aim at improving the domain adaptation performance by introducing the few-shot fully labeled target domain samples. However, they still assume the consistent taxonomy between the source and target domain.

Contrastive learning: Recently, the contrastive learning (Chen et al., 2020a; Grill et al., 2020; Chen et al., 2020b; He et al., 2020; Chen et al., 2020c) is proven to be successful for unsupervised image classification. Benefiting from the strong representation learning ability, contrastive learning has been applied to different applications, including semantic segmentation (Wang et al., 2021), image translation (Park et al., 2020), object detection (Xie et al., 2021) and domain adaptation (Kang et al., 2019). In Kang et al. (2019), contrastive learning is exploited to minimize the intra-class discrepancy and maximize the inter-class discrepancy for the domain adaptive image classification task. However, since the approach is designed for the images classification task, it utilizes the contrastive learning between the whole feature vector of the different image samples, which is not applicable to dense prediction tasks, such as semantic segmentation. Instead, we develop a pseudo-label guided and uncertainty-rectified pixel-wise contrastive learning, to distinguish between positive and negative pixel samples to learn more robust and effective cross-domain representations.

3 METHOD

3.1 THE TAXONOMY ADAPTIVE DOMAIN ADAPTATION (TADA) PROBLEM

In our introduced taxonomy adaptive domain adaptation (TADA) problem, we are given the labeled source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$, where $\mathbf{x}^s \in \mathbb{R}^{H\times W\times 3}$ is the RGB color image, and $\mathbf{y}^s$ is the associated ground truth C_S-class semantic label map, $\mathbf{y}^s \in \{1, ..., C_S\}^{H\times W}$. In the target domain, we are also given a limited number of labeled samples $\mathcal{D}_t = \{(\mathbf{x}_i^t, \mathbf{y}_i^t)\}_{i=1}^{n^t}$, which we refer to as few-shot or partially labeled target domain samples. We are also given a large set of unlabeled target domain samples $\mathcal{D}_u = \{\mathbf{x}_i^u\}_{i=1}^{n^u}$. The target ground truth $\mathbf{y}^t$ follows the C_T-class semantic label map. Denoting the source and target image samples distributions as P_S and P_T, we have $\mathbf{x}^s \sim P_S$, $\mathbf{x}^t, \mathbf{x}^u \sim P_T$. The source and target image distributions are different, *i.e.,* $P_S \neq P_T$. The label set space of $\mathcal{D}_s$ and $\{\mathcal{D}_t, \mathcal{D}_u\}$ are given by $\mathcal{C}_s = \{\mathbf{c}_1^s, \mathbf{c}_2^s, ..., \mathbf{c}_{C_S}^s\}$ and $\mathcal{C}_t = \{\mathbf{c}_1^t, \mathbf{c}_2^t, ..., \mathbf{c}_{C_T}^t\}$ resp., and $\mathcal{C}_s \neq \mathcal{C}_t$. The inconsistent taxonomy subsets of $\mathcal{C}_s, \mathcal{C}_t$ are denoted as $\overline{\mathcal{C}_s}, \overline{\mathcal{C}_t}$, resp. Our goal is to train the model on $\mathcal{D}_s$, $\mathcal{D}_t$ and $\mathcal{D}_u$, and evaluate on the target domain data in the label sets space $\mathcal{C}_t$.

Inconsistent Taxonomy.[1] Specifically, we consider three different cases of inconsistent taxonomy. 1) The *open taxonomy* considers the case where new classes, unseen or unlabeled in the source domain, appear in the target domain. That is, $\exists \mathbf{c}_j^t \in \mathcal{C}_t$ such that $\mathbf{c}_i^s \cap \mathbf{c}_j^t = \emptyset, \forall \mathbf{c}_i^s \in \mathcal{C}_s$. 2) The *coarse-to-fine taxonomy* considers the case where the target domain has a *finer* taxonomy where source classes have been split into two or more target classes. That is, $\exists \mathbf{c}_i^s \in \mathcal{C}_s, \mathbf{c}_{j_1}^t \in \mathcal{C}_t, \mathbf{c}_{j_2}^t \in \mathcal{C}_t, j_1 \neq j_2$ such that $\mathbf{c}_{j_1}^t, \mathbf{c}_{j_2}^t \neq \mathbf{c}_i^s$ and $(\mathbf{c}_{j_1}^t \cup \mathbf{c}_{j_2}^t) \subseteq \mathbf{c}_i^s$. 3) The *partially-overlapping taxonomy* considers the case where a class in the target domain has a common part with the class in the source domain, but also owns the private part. That is, $\exists \mathbf{c}_i^s \in \mathcal{C}_s, \mathbf{c}_j^t \in \mathcal{C}_t$ such that $\mathbf{c}_j^t \not\subseteq \mathbf{c}_i^s$,$\mathbf{c}_i^s \cap \mathbf{c}_j^t \neq \emptyset$, and $(\mathbf{c}_j^t \setminus (\mathbf{c}_i^s \cap \mathbf{c}_j^t)) \notin \{\emptyset, \mathbf{c}_q^s, q = 1, ..., C_S\}$.

Few-shot/Partially Labeled. In TADA, the $\mathcal{D}_t$ is only few-shot/partially labeled for the inconsistent taxonomy classes, in the class-wise way. More specifically, for each of the class $\mathbf{c}_j^t \in \overline{\mathcal{C}_t}$, we have n^t-shot labeled samples $\{(\mathbf{x}_i^{t_j}, \mathbf{y}_i^{t_j})\}_{i=1}^{n^t}$, where only the class $\mathbf{c}_j^t$ is labeled in $\mathbf{y}_i^{t_j}$. When $n^t \ll n^u$, it is called few-shot labeled. When $n^t \not\ll n^u$, it is named partially-labeled. The sample and corresponding semantic map is written as $\mathbf{x}^{t_j}$ and $\mathbf{y}^{t_j}$.

Technical Challenges. The main technical challenge of TADA is to deal with both of the label-level and image-level domain gap. On the **label level**, there are two main problems: i) The inconsistent taxonomy may induce there is the *one-to-many* mapping from the source domain to the target domain classes. If we purely assign the source class of inconsistent taxonomy to one of the corresponding

[1]With a slight abuse of notation, each class, *e.g.,* $\mathbf{c}_i^s$, is also considered as a set consisting of its domain of definition. The set operations $\cap, \cup, \setminus, \subset$ thus applies to the underlying definition of the class.

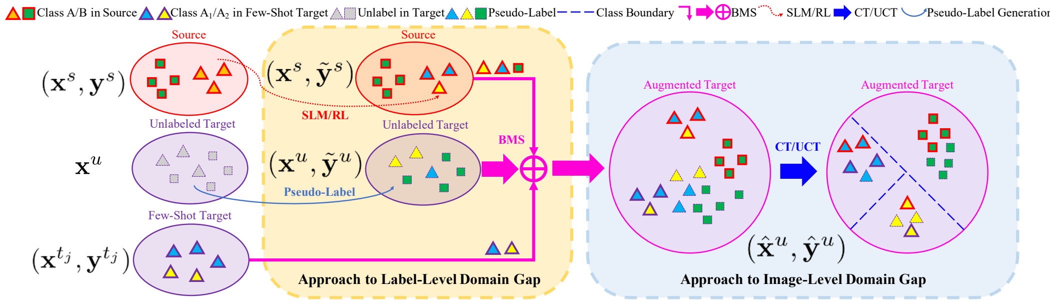

Figure 2: Framework overview. Class A is an inconsistent taxonomy class (e.g., "person") in the source domain, related to class A_1 (e.g., "pedestrian") and A_2 (e.g., "rider") in the target domain. Class B is a consistent taxonomy class. On the label level, SLM/RL module maps the inconsistent taxonomy class A in the source domain to the related classes A_1, A_2 in the target domain. BMS module unifies label space, by randomly selecting samples from the source and the few-shot/partially labeled target domain and mixing them in the unlabeled target domain. On the image level, CT/UCT module adopts the pseudo-label to distinguish positive and negative pixel samples, and then conducts the pixel-wise contrastive learning, to learn more domain-invariant and class-discriminative features.

target class, it will generate incorrect supervision, degrading the performance of the model. However, if we instead take the inconsistent source class as unlabeled, the source domain information is not fully exploited. ii) The complete target domain label taxonomy is partially inherited from the source domain for the consistent taxonomy, and partially provided by the few-shot/partially labeled target domain. The problem of how to *unify the consistent and inconsistent taxonomy classes* for the target domain is non-trivial. The naive way is to train the model on the source domain for the consistent taxonomy classes, and on the few-shot/partially labeled target domain for the inconsistent taxonomy classes separately, in the supervised way. However, the few-shot labeled target domain samples are far fewer than the labeled source domain samples, causing the model training to be easily dominated by the consistent taxonomy classes, therefore the inconsistent taxonomy classes are possibly ignored. Meanwhile, most of the pixels in the few-shot/partially labeled target domain samples are unlabeled except for the pixels of class $\mathbf{c}_j^t$, and the arbitrarily incorrect prediction on these unlabeled parts can bring the negative effect since most of these parts belong to the consistent taxonomy classes or other inconsistent taxonomy classes. On the **image level**, the image domain distribution difference between the source and target domain, $P_S \neq P_T$, still exists in TADA.

3.2 Our Approach to the TADA Problem

Motivation. Motivated by the technical challenge i) of the label level in Sec. 3.1, the stochastic label mapping (SLM) module and pseudo-label based relabeling (RL) module are proposed to solve the problem of the one-to-many mappings from the source domain to the target domain classes. Motivated by the technical challenge ii) of the label level in Sec. 3.1, the bilateral mixed sampling (BMS) module is proposed to unify the consistent and inconsistent taxonomy classes for the target domain. Motivated by the technical challenge of the image level in Sec. 3.1, the contrastive learning (CT/UCT) module is proposed to train the domain-invariant but class-discriminative features.

Training Strategy. The whole framework adopts pseudo-label based self-training strategy. Following the self-training structure of Tranheden et al. (2021); Olsson et al. (2021), there are two components of our framework, namely a semantic segmentation network $\mathcal{F}_\theta$ and a mean-teacher network $\mathcal{F}_{\theta'}$. The semantic segmentation network $\mathcal{F}_\theta$ is used to output the predicted semantic map. The pseudo-labels $\tilde{\mathbf{y}}^u = \mathcal{F}_{\theta'}(\mathbf{x}^u)$ are generated by the mean-teacher network $\mathcal{F}_{\theta'}$ by feeding the unlabeled target sample $\mathbf{x}^u$. The parameters θ' are the exponential moving average of the parameters θ during the optimization process, which is proven to bring more stable prediction (Tarvainen & Valpola, 2017) during training.

Framework Overview. The framework overview is shown in Fig. 2. The SLM and RL modules (Sec. 3.2.1) are used to map inconsistent taxonomy class labels $\mathbf{y}^s$ in the source domain to target-domain class labels $\tilde{\mathbf{y}}^s$. Then in order to unify the label spaces, the source domain sample $(\mathbf{x}^s, \tilde{\mathbf{y}}^s)$ and the few-shot/partially labeled target domain sample $(\mathbf{x}^{t_j}, \mathbf{y}^{t_j})$ is cut and mixed with the unlabeled target domain sample and corresponding pseudo-label $(\mathbf{x}^u, \tilde{\mathbf{y}}^u)$, to synthesize the sample $(\hat{\mathbf{x}}^u, \hat{\mathbf{y}}^u)$ through the BMS module (Sec. 3.2.1). In this way, the synthesized sample $(\hat{\mathbf{x}}^u, \hat{\mathbf{y}}^u)$ is a cross-domain mixed sample and covers the consistent taxonomy class from $(\mathbf{x}^s, \tilde{\mathbf{y}}^s)$ and inconsistent taxonomy class from $(\mathbf{x}^{t_j}, \mathbf{y}^{t_j})$. The CT/UCT module (Sec. 3.2.2) is further utilized on the $(\hat{\mathbf{x}}^u, \hat{\mathbf{y}}^u)$ to train the domain-invariant and class-discriminative features using pixel-wise contrastive learning. All the modules are thus employed together in a single framework. Next, we detail individual components.

3.2.1 Approach to the Label Level Domain Gap

In order to solve the problem of *one-to-many class mappings*, the SLM and RL modules are proposed. In the initial training stage, the model is unable to distinguish the different inconsistent taxonomy classes reliably. Thus, taking the coarse-to-fine taxonomy as example, we propose the SLM module, and it stochastically assigns the source "coarse class" to different corresponding target "finer classes" to guide the model to predict the uniform distribution over the "finer classes" on the source domain samples. In this way, in the early training stage, the prediction of the model on the "finer classes" will be mainly guided by the few-shot labeled target samples. As the training goes on, with the help of the few-shot labeled target samples, the teacher network gradually has the capacity to distinguish the "finer classes". In the second stage, we then replace the SLM module with the RL module. It relabels the "coarse-class" pixel in the source domain with the "finer class" predicted by the teacher network.

Stochastic Label Mapping (SLM). We propose the SLM module, which maps the source domain classes of inconsistent taxonomy, *e.g.,* "person" in Fig. 1 (e), to the corresponding target domain classes stochastically, *e.g.,* "pedestrian" and "rider" in Fig. 1 (e), in the initial training stage and *in each training iteration*. Under the inconsistent taxonomy setting, there might be the one-to-many class mapping from the source domain classes to the target domain label space. Without loss of generality and for the convenience of clarification, we take the example that the corresponding classes in $\mathcal{C}_t$ of $\mathbf{c}_i^s$ include q classes $\mathbf{c}_p^t, \mathbf{c}_{p+1}^t, ..., \mathbf{c}_{p+q-1}^t$. Then the SLM module can be described as,

$$\tilde{\mathbf{y}}^{s(m,n)} = \begin{cases} \mathcal{I}^t(\mathrm{rand}(\mathbf{c}_p^t, \mathbf{c}_{p+1}^t, ..., \mathbf{c}_{p+q-1}^t)), & \text{if } \mathbf{y}^{s(m,n)} = \mathcal{I}^s(\mathbf{c}^{s_i}), \\ \mathcal{I}^{st}(\mathbf{y}^{s(m,n)}), \text{otherwise}, \end{cases} \tag{1}$$

where the (m, n) is the (row, column) index. The $\mathrm{rand}(\cdot)$ represents the uniformly discrete sampling function. The function $\mathcal{I}^s(\cdot)$ maps the source domain class to the corresponding class index in source domain, *i.e.,* $\mathcal{I}^s : \mathbf{c}_i^s \to [1, C_S]$. The function $\mathcal{I}^t(\cdot)$ maps the target domain class to the corresponding class index in the target domain, *i.e.,* $\mathcal{I}^t : \mathbf{c}_j^t \to [1, C_T]$. The function $\mathcal{I}^{st}$ maps the consistent taxonomy source domain class index to the corresponding target domain class index, *i.e.,* $\mathcal{I}^{st} : [1, C_S] \to [1, C_T]$. With the new labels obtained in Eq. (1), we employ the standard cross-entropy loss, $\mathcal{L}_{slm} = CE(\mathcal{F}_\theta(\mathbf{x}_s), \tilde{\mathbf{y}}^s)$ to learn the model.

Pseudo-Label based Relabeling (RL). As the training goes on, the model learns to distinguish the different inconsistent taxonomy classes to some extent. Instead of adopting SLM strategy at the latter part of the training, we introduce an alternative strategy. To exploit the capabilities learned by the model, we perform pseudo-label based relabeling (RL), which relabels the pixels of inconsistent taxonomy classes in the source domain with the classes predicted by the model. Without generality loss and for writing convenience, we take the same example that $\mathbf{c}_i^s$ is related to $\mathbf{c}_p^t, \mathbf{c}_{p+1}^t, ..., \mathbf{c}_{p+q-1}^t$ as in SLM module. We generate predictions $\mathbf{f}^s = \mathcal{F}_{\theta'}(\mathbf{x}^s)$ by feeding the source domain sample $\mathbf{x}^s$ into the mean-teacher network $\mathcal{F}_{\theta'}$. Then the prediction $\mathbf{f}^s$ is used to relabel the source domain sample $\mathbf{x}^s$ for the inconsistent taxonomy classes $\mathbf{c}_i^s$, to generate the complete label $\tilde{\mathbf{y}}^s$ as,

$$\tilde{\mathbf{y}}^{s(m,n)} = \begin{cases} \arg\max_c \mathbf{f}^{s(m,n)}, & \text{if } \max_c \mathbf{f}^{s(m,n)} > \delta, \text{ and } \mathbf{y}^{s(m,n)} = \mathcal{I}^s(\mathbf{c}_{s_i}) \\ & \text{and } \arg\max_c \mathbf{f}^{s(m,n)} \in \mathcal{I}^t(\{\mathbf{c}_p^t, \mathbf{c}_{p+1}^t, ..., \mathbf{c}_{p+q-1}^t\}), \\ \mathcal{I}^{st}(\mathbf{y}^{s(m,n)}), \text{otherwise}, \end{cases} \tag{2}$$

where the δ represents the threshold to decide whether the predicted label is used. The pseudo-label based relabeling module loss is written as $\mathcal{L}_{rl} = CE(\tilde{\mathbf{y}}^s, \mathcal{F}_\theta(\mathbf{x}^s))$. The SLM module and the RL module are used in the sequential manner during the training process, *i.e.,* initially SLM and then RL.

Bilateral Mixed Sampling (BMS). In order to *unify the consistent and inconsistent taxonomy classes* for the target domain, we propose the bilateral mixed sampling (BMS) module, which cuts and mixes the source domain and few-shot/partially labeled target domain samples on the unlabeled target domain. Recently, the mixed sampling based data augmentation approach (Zhang et al., 2018; Ghiasi et al., 2020; Yun et al., 2019) is proven to be able to generate the synthetic data to combine the samples and corresponding labels, thus provides such a potential to unify the label space. In Tranheden et al. (2021), the cross-domain mixed sampling (DACS) is shown helpful to UDA of consistent taxonomy.

Similar to DACS for UDA, we adopt the class-mixed sampling strategy for TADA. Different from DACS, which only focus on the labeled source domain and the unlabeled target domain, our BMS module conducts the class-mixed sampling in the bilateral way: 1) from labeled source domain samples $\mathbf{x}^s$ to unlabeled target domain samples $\mathbf{x}^u$; 2) from few-shot/partially labeled target domain samples $\mathbf{x}^{t_j}$ to unlabeled target domain samples $\mathbf{x}^u$. The bilateral mixed sampling mask $\mathbf{m}^s$ of $\mathbf{x}^s$ is,

$$\mathbf{m}^{s(m,n)} = \begin{cases} 1, \text{if } \tilde{\mathbf{y}}^{s(m,n)} = \mathcal{I}^t(\mathbf{c}_r) \\ 0, \text{otherwise}, \end{cases} \tag{3}$$

where the sampling class $\mathbf{c}_r$ is randomly selected from the available classes in $\tilde{\mathbf{y}}^s$. Following Tranheden et al. (2021), half of all the available classes in $\tilde{\mathbf{y}}^s$ is randomly selected as the sampling class in each training iteration. Similar to $\mathbf{m}^s$, the bilateral mixed sampling mask $\mathbf{m}^{t_j}$ of $\mathbf{x}^{t_j}$ is defined. Then the augmented target domain sample and the corresponding pseudo-label $\hat{\mathbf{x}}^u$, $\hat{\mathbf{y}}^u$ are,

$$\hat{\mathbf{x}}^u = \mathbf{m}^s \odot \mathbf{x}^s + (1 - \mathbf{m}^s) \odot (\mathbf{m}^{t_j} \odot \mathbf{x}^{t_j} + (1 - \mathbf{m}^{t_j}) \odot \mathbf{x}^u), \tag{4}$$

$$\hat{\mathbf{y}}^u = \mathbf{m}^s \odot \tilde{\mathbf{y}}^s + (1 - \mathbf{m}^s) \odot (\mathbf{m}^{t_j} \odot \mathbf{y}^{t_j} + (1 - \mathbf{m}^{t_j}) \odot \tilde{\mathbf{y}}^u). \tag{5}$$

where the $\odot$ denotes element-wise multiplication. On this basis, the pseudo-label based self-training loss of our BMS module is formulated as, $\mathcal{L}_{bms} = CE(\hat{\mathbf{x}}^u, \hat{\mathbf{y}}^u)$.

3.2.2 Approach to the Image Level Domain Gap

Besides dealing with the label-level domain gap, we also need to handle the *image-level domain gap*. We propose a pseudo-label based contrastive learning (CT) module, and further the pseudo-label based uncertainty-rectified contrastive learning (UCT) module. They are easy to be plugged into our self-training pipeline and trained jointly with the BMS, SLM and RL modules.

Contrastive Learning (CT) for Domain Adaptation. The typical strategy of image-level adaptation is to train the domain-invariant but class-discriminative features in the cross-domain embedding space (Ganin & Lempitsky, 2015; Tsai et al., 2018; Ganin et al., 2016). The pixels of the same class from different or same domains need to have similar features in the feature embedding space, while the pixels of different classes needs be distinguishable in the feature embedding space. This kind of distinction between features can naturally be formulated as a contrastive learning problem, where positive pairs stem from pixels of the same class, irrespective of their domain. In Wang et al. (2021), the pixel-wise contrastive learning is proven to be helpful for semantic segmentation. However, it relies on ground truth supervision of the pixel, which is unavailable for our unlabeled target samples.

In order to exploit contrastive learning to train domain-invariant and class-discriminative features under cross-domain setting, we propose the pseudo-label based contrastive learning for domain adaptation. We employ pseudo-labels as guidance for distinguishing the positive and negative samples. The contrastive learning is conducted on the augmented target domain image sample $\hat{\mathbf{x}}^u$, and corresponding pseudo-label $\hat{\mathbf{y}}^u$ in the BMS module. Our main semantic segmentation network $\mathcal{F}_\theta$ can be decomposed into the encoder $\mathcal{E}_\theta$ and the decoder $\mathcal{M}_\theta$. The decoder is used to map the embedding space $\mathcal{B}$ to the label domain $\mathcal{Y}$. The encoder $\mathcal{E}_\theta$ maps the source image domain $\mathcal{S}$ and the target image domain $\mathcal{T}$ to the embedding space $\mathcal{V}$, *i.e.,* $\mathcal{E}_\theta : \mathcal{S}, \mathcal{T} \to \mathcal{V}$. The feature embedding corresponding to the sample $\hat{\mathbf{x}}^u$ is denoted as $\hat{\mathbf{v}}^u$, *i.e.,* $\hat{\mathbf{v}}^u = \mathcal{E}_\theta(\hat{\mathbf{x}}^u)$. Then the pseudo-label based contrastive learning module loss $\mathcal{L}_{ct}$ can be described as,

$$\mathcal{L}_{ct} = -\sum_h \sum_w \log \sum_{v^+ \in \mathcal{P}_v} \frac{\exp(v \cdot v^+/\tau)}{\exp(v \cdot v^+/\tau) + \sum_{v^- \in \mathcal{N}_v} \exp(v \cdot v^-/\tau)} \tag{6}$$

where $v = \hat{\mathbf{v}}^{u(h,w)}$ is the feature vector of $\hat{\mathbf{v}}^u$ at the position (h, w). The positive samples in $\mathcal{P}_v$ are the feature vectors whose corresponding pixels in $\hat{\mathbf{y}}^u$ have the same class label as that of the corresponding pixel of v. The negative samples in $\mathcal{N}_v$ are the feature vectors whose corresponding pixels in $\hat{\mathbf{y}}^u$ have the different class label from that of the corresponding pixel of v. Eq. (6) tries to learn similar features for the pixels of the same class label, and learn discriminative features for the pixels of different class label, no matter the pixels are in the same domain or not.

Uncertainty-Rectified Contrastive Learning (UCT) for Domain Adaptation. There unavoidably exist incorrect predictions in the pseudo-label $\hat{\mathbf{y}}^u$ of the unlabeled part in CT module, resulting in incorrect guidance to the contrastive module for the selection of the positive and negative samples. In order to alleviate the incorrect guidance, we propose the uncertainty-rectified contrastive learning

(UCT) module based on the CT module. In our UCT module, we use the prediction uncertainty of the pseudo-label $\hat{\mathbf{y}}^u$ to rectify the contrastive learning, so that the uncertain prediction of $\hat{\mathbf{y}}^u$ has less effect on the contrastive learning. The uncertainty estimation map of $\hat{\mathbf{y}}^u$ is denoted as $\hat{\mathbf{u}}^u$, and the uncertainty measurement function is denoted as $\mathcal{U}(\cdot)$, *i.e.,* $\hat{\mathbf{u}}^u = \mathcal{U}(\hat{\mathbf{y}}^u)$. We adopt the maximum prediction probability of $\hat{\mathbf{x}}^u$ as the uncertainty estimation function $\mathcal{U}(\cdot)$, formulated as,

$$\hat{\mathbf{u}}^u = \max_c \mathcal{F}_{\theta'}(\hat{\mathbf{x}}^u). \tag{7}$$

Then, based on Eq. (6), the uncertainty-rectified contrastive learning loss $\mathcal{L}_{uct}$ is formulated as,

$$\mathcal{L}_{uct} = -\sum_h \sum_w \hat{\mathbf{u}}^u(v)\hat{\mathbf{u}}^u(v^+) \log \sum_{v^+ \in \mathcal{P}_v} \frac{\exp(v \cdot v^+/\tau)}{\exp(v \cdot v^+/\tau) + \sum_{v^- \in \mathcal{N}_v} \exp(v \cdot v^-/\tau)} \tag{8}$$

where $\hat{\mathbf{u}}^u(v)$, $\hat{\mathbf{u}}^u(v^+)$ are the uncertainty estimation value of the pixel corresponding to v, v^+, resp.

3.3 Joint Training

With the above proposed BMS, SLM, RL and UCT modules, the total loss function is derived as,

$$\mathcal{L}_{total} = \mathcal{L}_{bms} + \lambda_1 \mathcal{L}_{slm} + \lambda_2 \mathcal{L}_{rl} + \lambda_3 \mathcal{L}_{uct} \tag{9}$$

where λ_1 and λ_2 are used to train the SLM and RL module in a sequential manner. When iteration $t < T$, $\lambda_1 = 1, \lambda_2 = 0$. When iteration $t \geq T$, $\lambda_1 = 0, \lambda_2 = 1$. T is the number of iterations to start training the RL module. λ_3 is the hyper-parameter to balance the UCT module loss and other loss, which is set as 0.01 in our work. Our model is trained end-to-end with the total loss in Eq. (9).

4 Experiments

We evaluate the effectiveness of our framework under different scenarios, including the consistent and inconsistent taxonomy settings. For the consistent taxonomy, we follow the traditional UDA setting. For the inconsistent taxonomy, we build different benchmarks for TADA, including an open taxonomy setting, a coarse-to-fine taxonomy setting, and a partially-overlapping taxonomy setting. The DeepLabv2-ResNet101 (Chen et al., 2017; He et al., 2016) is adopted as the segmentation network. Experimental details for baselines and our framework training are put in the supplementary.

4.1 Experimental Setup

UDA: Consistent Taxonomy. We adopt the UDA setting for the consistent taxonomy. The target domain is completely unlabeled. SYNTHIA (Ros et al., 2016) is used as the source domain, while Cityscapes (Cordts et al., 2016) is treated as the target domain. The source domain and target domains share the same label space, where there are 16 classes in total: *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *sky*, *person*, *rider*, *car*, *bus*, *motorcycle* and *bike*.

TADA: Open Taxonomy. The SYNTHIA dataset (Ros et al., 2016) is used as the source domain, and the Cityscapes dataset (Cordts et al., 2016) is adopted as the target domain. In the SYNTHIA dataset, the main 13 classes are labeled: *road*, *sidewalk*, *building*, *traffic light*, *traffic sign*, *vegetation*, *sky*, *person*, *rider*, *car*, *bus*, *motorcycle* and *bike*. In the Cityscapes dataset, the 6 classes *wall*, *fence*, *pole*, *terrain*, *truck* and *train* are few-shot labeled, with 30 image samples per class.

TADA: Coarse-to-Fine Taxonomy. The GTA5 dataset (Richter et al., 2016) is utilized as the source domain, and the Cityscapes dataset (Cordts et al., 2016) as the target domain. The label space of source domain is composed of *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *sky*, *person*, *car*, *truck*, *bus*, *train*, *cycle*. The *vegetation* class of source domain is further divided into *vegetation* and *terrain* in the target domain, *person* in source domain is mapped to *person* and *rider* in the target domain, and *cycle* in the source domain is fine-grained labeled into *bicycle* and *motorcycle* in the target domain. In Cityscapes, each of the fine-grained 6 classes is 30-shot labeled.

TADA: Partially-Overlapping Taxonomy. The Synscapes dataset (Wrenninge & Unger, 2018) is treated as the source domain, while the Cityscapes dataset (Cordts et al., 2016) is seen as the target domain. The label space of the source domain contains the *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *terrain*, *sky*, *person*, *rider* and *vehicle*. The *vehicle* class in source

Method	Road	SW	Build	Wall*	Fence*	Pole*	TL	TS	Veg	Sky	Person	Rider	Car	Bus	MC	Bike	mIoU*	mIoU
ADVENT(Vu et al., 2019)	87.0	44.1	79.7	9.6	0.6	24.3	4.8	7.2	80.1	83.6	56.4	23.7	72.7	32.6	12.8	33.7	47.6	40.8
FDA(Yang & Soatto, 2020)	79.3	35.0	73.2	–	–	–	19.9	24.0	61.7	82.6	61.4	31.1	83.9	40.8	**38.4**	51.1	52.5	–
IAST(Mei et al., 2020)	81.9	41.5	**83.3**	17.7	4.6	32.3	30.9	28.8	83.4	85.0	65.5	30.8	**86.5**	38.2	33.1	**52.7**	57.0	49.8
DACS(Tranheden et al., 2021)	80.56	25.12	81.90	21.46	**2.85**	**37.20**	22.67	23.99	83.69	**90.77**	**67.61**	**38.33**	82.92	38.90	28.49	47.58	54.81	48.34
Ours (DACS+CT)	86.32	26.63	82.71	5.78	1.97	33.87	**34.60**	**40.00**	83.83	86.73	67.52	36.53	83.46	**55.23**	25.03	41.46	57.70	49.47
Ours (DACS+UCT)	**91.54**	**60.41**	82.52	**21.80**	1.48	31.66	31.59	27.95	**84.71**	88.95	66.68	35.78	81.04	42.79	28.49	45.88	**59.10**	**51.45**

Table 1: Consistent Taxonomy: SYNTHIA→Cityscapes. The mIoU are over 13 classes and 16 classes, resp. In the UDA setting, we adopt the class-mixed sampling strategy in DACS to augment the target domain. *The 3 classes are not included when calculating the mIoU over 13 classes.

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU	mIoU
Source	29.22	6.58	55.48	4.79	8.71	10.11	4.04	12.93	64.06	5.09	71.90	43.26	11.93	22.43	6.04	6.96	2.42	2.61	16.41	6.19	20.26
ADVENT(Vu et al., 2019)	75.72	24.62	74.94	0.00	0.17	18.98	11.30	16.01	76.87	**21.93**	78.91	48.24	14.20	54.97	2.54	18.38	17.58	12.22	20.90	10.20	30.97
FDA(Yang & Soatto, 2020)	28.87	13.22	67.10	4.63	14.52	18.94	10.99	14.75	51.56	12.48	78.85	56.78	25.81	70.10	14.24	20.85	21.27	19.22	41.14	14.35	30.81
IAST(Mei et al., 2020)	70.73	29.60	75.49	6.90	0.00	1.36	36.43	25.37	66.17	7.65	83.96	60.72	19.99	82.51	0.00	39.52	0.09	27.42	23.55	2.67	34.60
DACS(Tranheden et al., 2021)	66.48	1.42	6.55	10.26	9.47	4.39	0.47	2.09	33.38	3.75	36.45	46.75	18.23	20.90	1.91	2.78	7.18	1.30	5.08	6.16	14.68
Ours (M)	87.59	27.18	80.98	5.99	15.74	7.13	37.09	18.51	83.68	0.08	87.46	65.89	**37.45**	86.55	24.76	40.58	37.71	**37.57**	43.44	15.24	43.44
Ours (M+CT)	86.33	32.57	82.62	9.49	12.78	5.10	**37.49**	**39.32**	82.00	0.73	**88.03**	65.70	33.09	78.92	33.55	**62.53**	**41.90**	29.83	49.35	17.26	45.86
Ours (M+UCT)	90.84	57.64	80.77	5.79	**16.67**	8.40	32.82	33.21	83.68	1.68	86.89	63.54	26.57	86.87	33.43	48.65	35.57	31.51	49.29	16.92	45.99
Ours (M+UCT+RL)	**92.64**	**58.66**	**84.21**	**20.55**	15.04	**29.47**	35.26	32.41	**84.63**	4.45	87.91	**66.16**	34.07	**87.52**	**36.37**	57.63	31.21	34.17	**52.28**	**22.85**	**49.72**
n^t=2975	89.19	41.08	86.14	37.54	33.68	33.45	32.25	39.99	85.39	31.64	89.51	67.02	35.61	80.49	50.54	49.43	51.70	32.41	47.90	39.76	53.42
Oracle (Wang et al., 2020)	96.7	75.7	88.3	46.0	41.7	42.6	47.9	62.7	88.8	53.5	90.6	69.1	49.7	91.6	71.0	73.6	45.3	52.0	65.5	50.0	65.9

Table 2: Open Taxonomy: SYNTHIA→Cityscapes. There are 13 classes labeled in the SYNTHIA dataset, and 6 new classes few-shot labeled in Cityscapes. The gray columns are the 6 new classes and mean IoU of 6 new classes in Cityscapes. ”M” represents the BMS module.

domain can be seen as the union of the *car*, *truck*, *bus*, and *motorcycle* classes. In the target domain, each of 3 classes are few-shot labeled in 15 image samples, including the vehicle, public transport and cycle. The *vehicle* class in the target domain is the union of *car* and *truck*, the *public transport* is the union of *bus* and *train*, and *cycle* is the union of the *bicycle* and *motorcycle*.

4.2 Experimental Results

Comparison with the SOTA and Ablation Study. In Table 1, it is shown that our proposed contrastive-learning based scheme outperforms the previous SOTA methods under the UDA setting, including the adversarial learning based ADVENT (Vu et al., 2019), the image translation based FDA (Yang & Soatto, 2020), the self-training based IAST (Mei et al., 2020), and the data augmentation based DACS (Tranheden et al., 2021). It proves the effectiveness of our contrastive learning for dealing with the domain gap on the image level. In Table 2, Table 3, and Table 4, it is shown that our proposed framework improves the other SOTA methods performance by a large margin, under the open, coarse-to-fine and partially-overlapping taxonomy settings. It validates the proposed framework for dealing with both of the image-level and label-level domain gap. The ablation study in Table 2, Table 3, and Table 4 proves that each module, BMS, SLM, RL, CT/UCT, all contributes to the final performance under open, coarse-to-fine, and partially-overlapping taxonomy settings. Besides, it is shown that the UCT module is able to reach higher performance than the CT module, verifying the help of our uncertainty rectification for contrastive learning. In Fig. 4, we show qualitative results.

Partially Labeled/Oracle. In Table 2, Table 3, and Table 4, under the open, coarse-to-fine, and partially-overlapping taxonomy setting, we report the partially labeled performance where inconsistent taxonomy classes are labeled in all the available target domain image samples, *i.e.*, $n^t = 2975$. Compared with the few-shot performance, the partially labeled performance is further improved

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU	mIoU
Source	54.12	16.20	70.08	13.07	19.37	22.56	28.59	20.59	75.87	13.49	74.36	47.91	5.35	36.15	16.08	9.71	1.61	8.77	21.34	28.79	29.22
Source*	63.38	20.95	67.65	15.07	18.60	23.03	27.74	18.00	76.03	14.11	75.19	38.36	10.25	49.01	26.32	9.23	2.68	9.93	27.26	29.32	31.20
ADVENT(Vu et al., 2019)	88.91	38.93	79.18	26.22	22.65	25.45	31.24	25.42	75.22	0.03	78.91	55.76	0.00	77.76	28.22	33.19	0.55	13.02	7.15	25.20	37.25
ADVENT*	86.72	34.02	79.22	22.32	23.60	26.92	31.36	24.89	59.86	3.39	75.47	41.83	7.73	69.62	32.71	20.39	0.49	12.06	39.25	27.35	36.41
FDA(Yang & Soatto, 2020)	90.83	45.07	81.62	28.37	31.04	32.56	34.00	29.80	83.09	6.31	72.61	60.67	10.13	82.71	29.06	51.51	0.11	15.69	45.61	36.92	43.73
FDA *	88.96	39.53	80.23	22.58	29.73	32.78	33.64	26.66	80.06	25.39	73.63	36.78	10.91	77.82	26.35	46.14	1.37	22.80	50.31	37.71	42.40
IAST(Mei et al., 2020)	83.20	37.84	82.63	36.00	21.59	32.34	43.48	44.69	84.92	36.51	88.77	59.71	28.04	84.34	32.64	38.66	2.52	31.27	35.57	46.00	47.62
IAST*	76.62	32.39	83.04	**37.52**	23.43	28.96	39.11	39.47	81.33	26.02	**89.10**	56.83	26.41	82.36	18.95	38.16	**23.03**	21.14	44.22	42.66	45.69
DACS(Tranheden et al., 2021)	82.93	29.50	69.67	31.58	24.87	18.17	20.71	17.43	69.69	8.54	64.06	32.17	9.78	76.99	36.40	44.26	0.00	8.64	30.39	26.54	35.57
DACS *	45.03	18.55	24.01	9.80	12.25	10.14	13.08	5.62	46.05	4.23	23.95	14.94	8.64	52.14	36.28	12.43	0.00	8.35	15.08	16.22	18.98
Ours(M)	93.60	60.14	85.64	34.57	25.27	33.67	34.67	41.84	83.03	2.67	86.96	60.15	2.34	87.25	52.06	47.66	0.00	17.81	42.53	34.76	46.94
Ours(M+SLM)	93.33	57.28	86.14	36.66	29.25	36.84	43.25	43.09	85.50	**39.17**	85.85	63.47	26.95	88.71	52.76	53.06	0.00	**41.46**	57.13	52.28	53.68
Ours(M+SLM+CT)	93.83	60.53	86.37	30.73	35.05	36.69	41.74	47.82	**85.70**	38.69	85.75	62.65	36.28	87.89	51.00	52.84	0.00	39.71	59.11	53.69	54.34
Ours(M+SLM+UCT)	**94.51**	**62.40**	87.15	29.95	35.96	37.96	44.17	52.17	84.56	34.33	84.80	65.79	37.41	**90.03**	**56.10**	52.57	0.00	40.46	59.82	53.73	55.27
Ours(M+SLM+UCT+RL)	93.97	59.71	**87.58**	29.81	**36.26**	**38.81**	**45.38**	**52.53**	85.26	35.18	87.28	**66.58**	**38.74**	89.74	55.23	**54.72**	0.00	40.72	**60.47**	**54.49**	**55.68**
n^t=2975	93.65	56.25	86.48	27.37	39.02	37.59	43.73	50.49	87.08	49.25	86.38	67.71	43.83	89.40	50.98	47.01	0.09	45.42	63.96	59.54	56.09
Oracle (Wang et al., 2020)	96.7	75.7	88.3	46.0	41.7	42.6	47.9	62.7	88.8	53.5	90.6	69.1	49.7	91.6	71.0	73.6	45.3	52.0	65.5	63.1	65.9

Table 3: Coarse-to-Fine Taxonomy: GTA5→Cityscapes. There are 3 classes in the GTA5 dataset fine-grained into 6 classes in the Cityscapes dataset. The gray columns are the 6 fine-grained classes in the Cityscapes and corresponding mean IoU of these classes. ”M”: BMS. ”*” with SLM module.

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Vehicle	PT	Cycle	mIoU	mIoU
Source	82.74	43.14	70.95	29.04	19.24	33.99	34.47	36.29	81.90	28.67	86.61	55.17	28.25	54.75	1.75	34.99	30.50	45.12
Source*	87.95	40.99	74.68	24.35	22.67	32.17	31.86	34.74	81.53	27.52	83.74	55.08	26.68	67.51	11.34	21.56	33.47	45.27
ADVENT(Vu et al., 2019)	92.84	54.32	82.54	31.40	25.90	37.67	38.92	40.55	85.46	35.95	87.69	58.12	29.75	73.19	2.42	3.23	26.28	48.75
ADVENT*	90.02	46.16	80.37	27.90	24.56	35.69	31.48	37.81	83.96	38.81	84.83	54.73	30.69	73.67	16.02	18.80	36.16	48.47
FDA(Yang & Soatto, 2020)	89.45	44.66	75.82	28.3	27.91	37.89	41.09	49.91	83.78	26.17	83.50	61.24	39.37	65.35	6.32	26.56	32.74	49.21
FDA *	86.86	43.56	75.32	28.01	27.68	38.50	39.50	50.31	83.80	21.69	83.93	63.45	**42.32**	80.99	10.96	42.64	44.86	51.22
IAST(Mei et al., 2020)	91.65	54.26	81.82	31.61	28.48	35.33	42.83	46.74	85.67	41.89	89.47	57.51	32.77	75.78	31.13	50.45	52.45	54.84
IAST *	93.00	55.31	83.55	32.80	30.49	38.21	46.04	53.09	86.46	41.91	88.57	60.58	29.17	83.18	39.01	36.76	52.98	56.13
DACS(Tranheden et al., 2021)	89.72	61.93	57.59	28.87	26.87	33.42	41.44	41.14	84.57	41.96	86.49	57.94	25.36	59.88	2.13	19.63	27.21	47.43
DACS *	82.27	41.83	13.43	17.67	18.84	23.23	23.93	23.54	56.89	18.20	68.49	44.60	13.75	22.09	2.39	16.75	13.74	30.49
Ours(M)	91.35	59.29	**86.81**	**34.60**	32.14	43.9	49.29	55.8	83.51	**42.28**	90.44	**67.98**	37.27	83.01	16.89	**43.92**	47.94	57.40
Ours(M+SLM)	93.66	65.25	81.31	28.81	26.43	44.96	51.70	55.84	87.59	38.47	88.80	67.93	35.10	87.71	35.55	36.29	53.18	57.84
Ours(M+SLM+CT)	**95.70**	**70.24**	85.42	29.16	25.78	42.10	49.77	54.14	**87.67**	42.11	90.10	66.59	36.67	87.55	34.97	40.43	54.32	58.65
Ours(M+SLM+UCT)	92.43	66.46	82.25	32.24	**32.47**	45.37	**52.29**	**57.15**	87.20	36.48	**91.85**	65.03	37.87	88.53	41.95	38.11	56.20	59.23
Ours(M+SLM+UCT+RL)	92.47	65.40	83.21	33.33	30.87	**45.94**	49.86	55.86	87.23	39.50	91.30	66.56	39.87	**88.75**	**42.59**	39.64	**56.99**	**59.52**
n^t=2975	94.62	63.90	85.13	28.52	31.03	46.46	53.44	50.16	86.98	41.21	91.00	67.61	35.04	89.98	74.72	52.85	72.52	62.04
Oracle	96.79	76.53	87.75	49.21	41.14	40.64	43.82	60.49	88.01	52.68	89.16	68.68	49.33	91.05	74.69	64.26	76.67	67.14

Table 4: Partially-Overlapping Taxonomy: Synscapes→Cityscapes. There are 3 classes (in gray) in the Cityscapes partially overlapping with the source domain classes. ”M”: BMS. ”*”: with SLM.

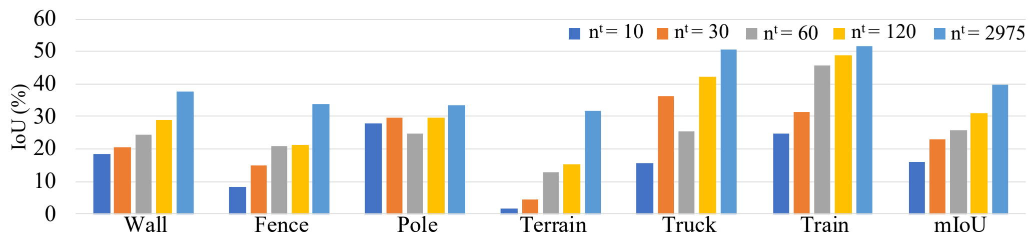

Figure 3: Performance of inconsistent taxonomy classes under open taxonomy setting, varying n^t.

due to more labeled samples on the target domain being available. But there is still gap to the fully supervised oracle performance on the target domain. It shows that our proposed framework serves as a strong baseline for TADA, but still provides the potential to develop stronger algorithms for TADA.

Effect of Few-shot Samples Number. In order to analyze the effect of the number of few-shot samples in the target domain for the inconsistent taxonomy adaptation performance, we take the open taxonomy setting as the example, and show the performance change with different number of few-shot samples in Fig. 3. It is shown that the inconsistent taxonomy class adaptation performance is improved, when more few-shot labeled samples are available, $n_t = 10, 30, 60, 120, 2975$.

Contrastive Learning. In Fig. 4, we compare the t-SNE visualization (Van der Maaten & Hinton, 2008) of the feature embedding of the model trained with/without UCT module, taking the open taxonomy setting as example. It verifies that the contrastive learning is helpful to train the cross-domain invariant and class-discriminative features.

5 Conclusion

We propose the new TADA problem, allowing inconsistent taxonomies between the source and target domain in domain adaptation. Pseudo-labeling and contrastive learning based techniques are developed, to reduce the domain gap on both of the image level and label level. Extensive experiments on both UDA and TADA settings prove the effectiveness of our approach.

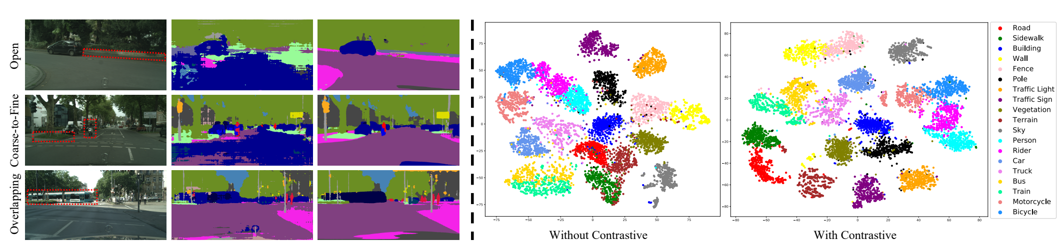

Figure 4: **Left:** Qualitative segmentation results under different inconsistent taxonomy settings. Each group has the RGB image (left), the results without adaptation (middle) and adapted with our pipeline (right). Refer to the red box region for the adaptation of the inconsistent taxonomy classes. **Right:** t-SNE visualization of the features with/without contrastive learning under the open taxonomy setting.

Ethics Statement. Our proposed approach provides the potential to adapt the model even under the inconsistent taxonomy, saving much cost and effort for labeling when new data and new requirements come. Thus, there is also a risk for reduced need of data labelling, leading fewer jobs in this domain and potential unemployment.

Reproducibility Statement. In Sec. 4.1 of the main paper, we introduce the experimental setup of consistent and inconsistent taxonomy experiments. In Sec. S1 of the supplementary, we list the framework implementation details such as the batch size, parameters, contrastive learning, baseline setup and compute resources. All of the aforementioned details make our paper reproducible. Besides, upon acceptance of our paper, we promise to make our codes publicly available, to further help reproduce our paper.

REFERENCES

Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Handling new target classes in semantic segmentation with domain adaptation. *arXiv preprint arXiv:2004.01130*, 2020.

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. *TPAMI*, 40(4):834–848, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020b.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018.

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.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *arXiv preprint arXiv:2012.07177*, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVRP*, 2019.

Jogendra Nath Kundu, Rahul Mysore Venkatesh, Naveen Venkat, Ambareesh Revanur, and R Venkatesh Babu. Class-incremental domain adaptation. In *ECCV*, 2020.

John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *CVPR*, 2020.

Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X. Yu, and Boqing Gong. Open compound domain adaptation. In *CVPR*, 2020.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, 2020.

Saeid Motiian, Quinn Jones, Seyed Mehdi Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. In *NeurIPS*, 2017.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017.

Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *WACV*, 2021.

Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, 2017.

Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, 2020.

Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016.

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 *CVPR*, 2016.

Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, 2018.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

Takeshi Teshima, Issei Sato, and Masashi Sugiyama. Few-shot domain adaptation by causal mechanism transfer. In *ICML*, 2020.

Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021.

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.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019.

Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021.

Zhonghao Wang, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S Huang, and Honghui Shi. Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation. In *CVPR Workshops*, 2020.

Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. *arXiv preprint arXiv:2102.04803*, 2021.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020.

Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *CVPR*, 2019.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Junyi Zhang, Ziliang Chen, Junying Huang, Liang Lin, and Dongyu Zhang. Few-shot structured domain adaptation for virtual-to-real scene parsing. In *ICCV Workshops*, 2019.

Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018.

TADA: TAXONOMY ADAPTIVE DOMAIN ADAPTATION — SUPPLEMENTARY MATERIAL —

Anonymous authors
Paper under double-blind review

In this supplementary material, we provide the additional information for,

S1 detailed implementation of our proposed framework,

S2 detailed information of involved datasets in our experiments,

S3 additional quantitative and qualitative experimental results.

S1 FRAMEWORK IMPLEMENTATION

In the main paper, we propose the new taxonomy adaptive domain adaptation (TADA) problem, which allows inconsistent taxonomies between the source domain and the target domain in the domain adaptation. TADA approaches the domain adaptation on both of the image level and the label level. In order to address the TADA problem, a set of pseudo-labelling techniques and the contrastive learning scheme are developed to reduce both of the label-level and image-level domain gap (cf. Sec. 3 of the main paper). Our proposed complete approach demonstrates the strong performance under different TADA settings, open taxonomy, coarse-to-fine taxonomy and partially-overlapping taxonomy (cf. Table 2, Table 3 and Table 4 of the main paper). Moreover, our suggested mixed-sampling and contrastive learning based scheme outperforms the state-of-the-art (SOTA) methods by a large margin, under traditional unsupervised domain adaptation (UDA) setting (cf. Table 1 of the main paper). Here we present the implementation details of our proposed framework.

Batch Size. For the open taxonomy, coarse-to-fine taxonomy and partially-overlapping taxonomy experiments of TADA in Sec. 4 of the main paper, in each training batch, there are 2 source domain images, 2 unlabeled target domain images and 2 few-shot labeled target domain images mixed in the bilateral mixed sampling module. For the consistent taxonomy experiments of UDA in Sec. 4 of the main paper, we strictly follow the traditional UDA setting, and the target domain is completely unlabelled. Therefore, under UDA setting, in each training batch, there are 2 source domain images and 2 unlabelled target domain images mixed in the class mixed sampling way (Tranheden et al., 2021).

Parameters. The source domain images are resized to 1280×720, and the target domain images are resized to 1024×512. And the random crop with size 512×512 is then adopted. We adopt the SGD optimizer to train the semantic segmentation network, whose momentum is set as 0.9 and the weight decay is set to 5×10^{-4}. The learning rate is set as 2.5×10^{-4}, with polynomial decay of power 0.9. The iteration T in Sec. 3.3 for starting training the RL module is set as 130000. The total training iteration is set as 250000.

Contrastive Learning. Following Wang et al. (2021), for each mini-batch, we use 100 anchor pixel samples per category. The 100 pixel samples of the same category are taken as positive samples, while the other pixel samples of different categories are all taken as negative samples.

Baseline Setup. In the baseline methods setup of Table 2, Table 3 and Table 4 in the main paper, we add the additional supervised loss to train the model in the supervised way, with the few-shot/partially labeled samples in the target domain. For the baseline methods which adopt the pseudo-label based training strategy, such as FDA (Yang & Soatto, 2020), IAST (Mei et al., 2020), and DACS (Tranheden et al., 2021), the few-shot/partial label on the target domain samples is combined with the generated pseudo-label to attain the final pseudo-label. I.e., in the pseudo-label generation process on the few-shot/partially labeled samples, we adopt the ground-truth label for the labeled parts, while we adopt the generated pseudo-label for other unlabeled parts.

Compute Resources. The code is implemented with PyTorch (Paszke et al., 2019). Experiments are conducted on an NVIDIA GeForce RTX 2080 Ti GPU, with 11GB memory, where it takes 3 days for training the whole 250000 iterations. In the whole investigation process of our paper, the total compute used is around 390×3 GPU days.

S2 Datasets Information

As introduced in Sec. 4 of the main paper, there are 4 datasets in total involved in our experiments, including SYNTHIA (Ros et al., 2016), GTA5 (Richter et al., 2016), Synscapes (Wrenninge & Unger, 2018) and Cityscapes (Cordts et al., 2016). Here we provide more information about the datasets.

SYNTHIA. SYNTHIA is a syntheic image dataset, consisting of photo-realistic images rendered from a virtual city. We adopt the SYNTHIA-RAND-CITYSCAPES subset, including 9400 densely labeled synthetic images. SYNTHIA is licensed under a CC BY-NC-SA 3.0 license.

GTA5. GTA5 is a synthetic image dataset, containing 24966 urban scene images. The images in GTA5 dataset are rendered from game engine, and densely labeled with pixel-level semantic annotation. The scene of GTA5 dataset is based on the city of Los Angeles. We were unable to find the license for the GTA5 dataset. But the code for extracting the GTA5 dataset image from the game engine is released under the MIT license.

Synscapes. Synscapes is a photo-realistic synthetic dataset, created with physically based rendering techniques. Synscapes is built for street scene parsing, composed of 25000 densely pixel-level annotated images. Synscapes customizes the license, *i.e.,* Synscapes grants a non-exclusive, non-transferable, non-sublicensable, worldwide license to use the dataset for non-commercial purposes.

Cityscapes. Cityscapes is a real street scene image dataset, collected from different European cities. We adopt the training set of Cityscapes during the training stage, covering 2975 images. And we use the validation set of Cityscapes, including 500 images, to evaluate the performance of the semantic segmentation model. Cityscapes customizes the license, *i.e.,* Cityscapes is made freely available to academic and non-academic entities for non-commercial purposes such as academic research, teaching, scientific publications, or personal experimentation.

Whether the datasets cover personally identifiable information or offensive content? The SYNTHIA, GTA5 and Synscapes are all synthetic image datasets, and are rendered from the virtual city or game engine. The personally identifiable information or offensive content is not found in them. Cityscapes is a real street scene image dataset, but Cityscapes is for non-commercial use only. Even though Cityscapes covers the "person" class as one of the semantic annotation classes, the personally identificable information or offensive content is also not found in Cityscapes. Besides, Citysacpes creators state that, if any people find themselves or their personal belongings in the data, they will immediately remove the respective images from their servers after receiving the contact from the people.

S3 Additional Experimental Results

In Sec. 4 of the main paper, we report the experimental results under the traditional UDA setting and different TADA settings, *i.e.,* open taxonomy, coarse-to-fine taxonomy and partially-overlapping taxonomy. Here we provide additional quantitative and qualitative experimental results to further prove the effectiveness of our proposed approach.

S3.1 TADA: Coarse-to-Fine Taxonomy involving More Classes

In order to prove the effectiveness of our proposed approach when dealing with the inconsistent taxonomy involving more classes, we provide the experimental results under the coarse-to-fine taxonomy setting, with more fine-grained classes in the target domain.

Setup. We adopt the GTA5 dataset as the source domain, and the Cityscapes dataset as the target domain. The label space of source domain is composed of *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *terrain*, *vegetation*, *sky*, *moving objects*. The *moving objects* class in

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU	mIoU
Source	71.59	20.93	67.54	10.00	15.49	24.15	29.90	19.46	79.83	19.10	74.07	34.95	10.53	67.43	9.98	17.72	7.86	4.75	25.14	22.30	32.13
IAST(Mei et al., 2020)	81.87	35.74	79.58	37.35	25.77	32.26	**45.14**	39.14	85.34	34.09	85.14	57.58	27.32	81.64	28.01	45.54	26.03	21.58	44.28	41.50	48.08
Ours	**95.35**	**68.30**	**86.75**	**41.39**	**38.95**	**36.62**	43.96	**49.49**	**87.64**	**45.90**	**87.43**	**63.96**	**28.31**	**88.41**	**45.41**	**59.17**	**57.34**	**37.02**	**57.13**	**54.59**	**58.87**

Table S1: Coarse-to-Fine Taxonomy: GTA5→Cityscapes. The ”moving object” class in the GTA5 dataset is fine-grained into 8 classes in the Cityscapes dataset. The gray columns are the 8 fine-grained classes in the Cityscapes and corresponding mean IoU of these classes.

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU
ADVENT(Vu et al., 2019)	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
FDA(Yang & Soatto, 2020)	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
IAST(Mei et al., 2020)	**93.8**	**57.8**	85.1	**39.5**	26.7	26.2	43.1	34.7	84.9	32.9	88.0	62.6	29.0	87.3	39.2	49.6	**23.2**	34.7	39.6	51.5
DACS(Tranheden et al., 2021)†	89.90	39.66	87.87	30.71	**39.52**	38.52	**46.43**	52.79	**87.98**	43.96	**88.76**	**67.20**	**35.78**	84.45	45.73	50.19	0.00	27.25	33.96	52.14
DACS(Tranheden et al., 2021)*	93.25	50.20	87.21	36.75	34.80	38.83	39.80	48.68	87.06	44.06	88.76	65.19	34.38	89.25	**51.64**	52.71	0.00	28.59	48.42	53.66
Ours (DACS+UCT)	93.03	55.92	**87.91**	38.19	38.76	**40.44**	42.14	**54.50**	87.53	**46.67**	87.77	66.26	33.67	**90.18**	47.54	**54.15**	0.00	**41.24**	**53.34**	**55.75**

Table S2: Consistent Taxonomy: GTA5→Cityscapes. The mIoU is over 19 classes. In the UDA setting, we adopt the class-mixed sampling strategy in DACS to augment the target domain. The best results are denoted in bold. † is the performance reported in the DACS (Tranheden et al., 2021). * is the peak performance model publicly provided by the author of DACS (Tranheden et al., 2021).

the source domain is further divided into 8 classes, including *person*, *rider*, *car*, *truck*, *bus*, *train*, *motorcycle* and *bicycle* in the target domain.

Comparison with the SOTA. In Table S1, we show the quantitative comparison between our proposed method, the non-adapted baseline ”source” and other SOTA self-training based method IAST (Mei et al., 2020). Same as the ”source” baseline in the Table 2, Table 3 and Table 4 of the main paper, the non-adapted baseline ”source” in Table S1 is trained in the supervised way on the labeled source domain and the few-shot labeled target domain. It is shown that both of the adaptation-based methods, IAST and our proposed method, perform better than the non-adapted baseline method, 48.08%, 58.87% *v.s.* 32.13%. Moreover, our proposed method outperforms the IAST method by a large margin, 58.87% *v.s.* 48.08. It proves the effectiveness of our proposed method when dealing with the inconsistent taxonomy involving more classes.

S3.2 UDA: Consistent Taxonomy

In Table. 1 of the main paper, we show the comparison between our suggested mixed-sampling and contrastive learning based scheme and other SOTA methods under traditional UDA setting, SYNTHIA→Cityscapes. It is shown that our suggested mixed-sampling and contrastive learning based scheme outperforms other SOTA methods under traditional UDA setting. Here we provide additional quantitative experimental results under the traidtional UDA setting, GTA5→Cityscapes, to further prove the effectiveness of our suggested mixed-sampling and contrastive learning based scheme for traditional UDA problem.

Setup. We adopt the GTA5 dataset as the source domain, and the Cityscapes dataset as the target domain. The source domain and the target domain share the same label space, where there are 19 classes in total: *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *terrain*, *sky*, *person*, *rider*, *car*, *truck*, *bus*, *train*, *motorcycle* and *bicycle*. We strictly follow the traditional UDA setting, and the target domain is completely unlabelled.

Comparison with the SOTA. In Table S2, we report the quantitative experimental results of our suggested mixed-sampling and contrastive learning based scheme and other SOTA methods under the traditional UDA setting. It is shown that our suggested mixed-sampling and contrastive leaning based scheme outperforms current SOTA methods under the traditional UDA setting, 55.75% *v.s.* 53.66%. It further verifies the validity of our suggested mixed-sampling and contrastive learning based scheme for traditional UDA problem.

S3.3 Additional Qualitative Results

In Fig. 4 of the main paper, we show the qualitative semantic segmentation results, w/o adaptation and adapted with our proposed method, under the open taxonomy, coarse-to-fine taxonomy and partially-overlapping taxonomy setting. Here we further provide more qualitative segmentation results, w/o adaptation, adapted with other method, and adapted with our proposed method, under the aforementioned settings. In Fig. S1, under different inconsistent taxonomy settings, we show the qualitative semantic segmentation results on the target domain, w/o adaptation, adapted with IAST (Mei et al., 2020), and adapted with our proposed method. It is shown that our proposed method outperforms the non-adaptation baseline and other adaptation-based method IAST (Mei et al., 2020) qualitatively. It further proves the effectiveness of our proposed method for the TADA problem.

REFERENCES

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.

Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016.

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 *CVPR*, 2016.

Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019.

Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021.

Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020.

Figure S1: Qualitative semantic segmentation results on the target domain under different inconsistent taxonomy settings, open taxonomy, coarse-to-fine taxonomy and partially-overlapping taxonomy. (a) shows the RGB target domain image. (b) gives the ground truth semantic segmentation map. (c) is the semantic segmentation result without adaptation. (d) is the semantic segmentation result adapted by the IAST (Mei et al., 2020) method. (e) is the semantic segmentation result adapted by our proposed method. Refer to the red box region for the adaptation results of the inconsistent taxonomy classes. The target domain label space of open taxonomy and coarse-to-fine taxonomy setting both have 19 classes, whose corresponding color in the semantic segmentation map is listed in the top color grid. The target domain label space of the partially-overlapping taxonomy setting has 16 classes, whose corresponding color in the semantic segmentation map is listed in the low color grid.