TACS: Taxonomy Adaptive Cross-Domain Semantic Segmentation

Rui Gong[1], Martin Danelljan[1], Dengxin Dai[3], Danda Pani Paudel[1], Ajad Chhatkuli[1], Fisher Yu[1], and Luc Van Gool[1,2]

[1] Computer Vision Lab, ETH Zurich, Switzerland
{gongr,martin.danelljan,paudel,ajad.chhatkuli,vangool}@vision.ee.ethz.ch
i@yf.io
[2] VISICS, ESAT/PSI, KU Leuven, Belgium
[3] VAS, MPI for Informatics, Germany ddai@mpi-inf.mpg.de

Abstract. Traditional domain adaptive semantic segmentation 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 cross-domain semantic segmentation (TACS) 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 TACS settings: open taxonomy, coarse-to-fine taxonomy, and implicitly-overlapping taxonomy. Our approach outperforms the previous state-of-the-art by a large margin, while being capable of adapting to target taxonomies. Our implementation is publicly available at https://github.com/ETHRuiGong/TADA.

Keywords: Domain Adaptation, Semantic Segmentation, Inconsistent Taxonomy

1 Introduction

Traditional unsupervised domain adaptation (UDA) approaches for semantic segmentation [7, 15, 20, 34, 35, 37] typically focus on the *image level* domain gap, which can involve 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 (Fig. 1 (a-c)), which is often not the

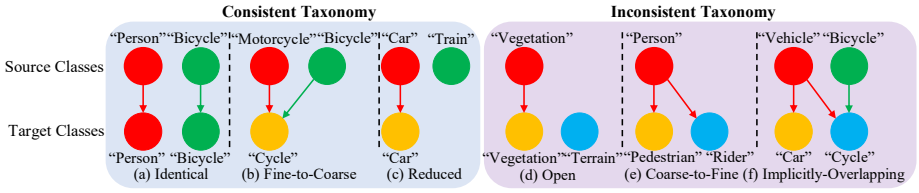

Fig. 1: Consistent *vs.* inconsistent taxonomy. In (a)-(f), the upper row shows the source domain classes, and the lower row the target domain classes. Circles represent classes while an arrow represents a mapping from a source domain class to a target domain class. (a)-(c) and (d)-(f) are examples of consistent and inconsistent taxonomies, resp. Different from other domain adaptation problems, *e.g.,* universal/partial/open-set domain adaptation [2,27,43], that only touch the consistent taxonomy or special case of open taxonomy, our TACS provides a more general problem, including the consistent taxonomy and different inconsistent taxonomies types. More detailed comparisons with other domain adaptation problems are put in Sec. 2 and Sec. S2 in the supplementary.

case. In many applications, the label spaces of the source and target domains are inconsistent, due to different scenarios or requirements, inconsistent annotation practices, or the strive towards an increasingly fine-grained taxonomy [8,19,25].

The aforementioned considerations motivate us to consider the *label level* domain gap problem. Even though recent open/universal/class-incremental domain adaptation works [18,27,43] touched upon the label level domain gap, they 1) only took image classification as test-bed, and 2) only focused on unseen classes in the target domain. However, the label level domain gap in practical scenarios is more complicated than only involving unseen classes. We therefore formulate and explore the label level domain gap problem in a more general and complete setting. We identify three typical types 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 split into several sub-classes in the target domain, *e.g.,* "pedestrian" and "rider' (Fig. 1(e)). iii) *Implicitly-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 a 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 cross-domain semantic segmentation* (TACS). In traditional UDA for semantic segmentation, 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, TACS allows for inconsistent taxonomies between a labeled source domain and a few-shot/partially labeled target domain, where the inconsistent classes of the target domain are exemplified by a few labeled samples. Thus TACS approaches domain adaptation on both the

image and label side, under the few-shot/partially labeled setting. Such task setting is realistic, but involves practical challenges. On the one hand, TACS allows methods to make full use of the labeled source domain without annotation costs in the target domain for the consistent classes. On the other hand, for the inconsistent classes the taxonomy adaptation should only require very limited supervision in the target domain, *i.e.,* only few samples should be labeled there.

We put forward the first approach for TACS, addressing both the image and label domain gaps. As to the latter, we aim to remedy the gap 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 the model's predictions, to further enforce taxonomy adaptation during the training process. To tackle the 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 TACS demonstrates strong results in different inconsistent taxonomy settings (*i.e.,* open, coarse-to-fine, and implicitly-overlapping). Moreover, our suggested mixed-sampling and contrastive-learning scheme outperforms current state-of-the-art methods by a large margin in the traditional UDA setting.

To summarize, our contributions are three-fold:

- A new problem – *taxonomy adaptive cross-domain semantic segmentation* (TACS) – of addressing both image and label domain gaps is proposed. It opens up a new avenue for more flexible cross-domain semantic segmentation.
- A generic solution for UDA and TACS is proposed, for which the unified mixed-sampling, pseudo-labeling and uncertainty-rectified contrastive learning scheme is presented to solve both image and label level domain gaps.
- Extensive experiments are conducted under the traditional UDA and the new TACS settings, showing the effectiveness of our approach.

2 Related Work

Domain adaptation: The traditional unsupervised domain adaptation (UDA) [9, 16, 21, 35, 47, 48] 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. **Open-set/universal domain adaptation** [27, 31, 43] aims at recognizing the new unseen classes in the target domain together as the "unknown" class. **Class-incremental/zero-shot domain adaptation** [1, 18] are proposed to recognize the new unseen classes explicitly and separately in 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 the open taxonomy setting in TACS. 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 TACS also allows for the unseen classes to exist in the source domain, where they are unlabelled. Besides, the above works do not consider the coarse-to-fine and implicitly-overlapping taxonomy problems, which are covered by the more general TACS formulation. Recent **few-shot/semi-supervised domain adaptation** works [24,33,46] aim at improving the domain adaptation performance by introducing few-shot fully labeled target domain samples. However, they still assume a consistent taxonomy between the source and target domain. Moreover, all the aforementioned non-UDA works, except for [1] and [46], only touch upon the image classification task. Instead, our TACS aims at semantic segmentation, which is more challenging and raises particular interest due to its great importance in autonomous driving [23,34,35,37]. More detailed comparisons between our TACS and different domain adaptation problems are put in the supplementary.

Contrastive learning: Recently, contrastive learning [4–6,12,13,36] was 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 [38], image translation [28], object detection [41] and domain adaptation [17]. In [17], 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 image classification task, it utilizes the contrastive learning between the whole feature vectors of the different image samples, which is not directly 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 Problem Statement

In our taxonomy adaptive cross-domain semantic segmentation (TACS) 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. [4] 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 $\mathbf{c}_i^s \cap \mathbf{c}_j^t = \varnothing, \forall \mathbf{c}_i^s \in \mathcal{C}_s$. 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 *implicitly-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 $\mathbf{c}_j^t \not\subseteq \mathbf{c}_i^s$,$\mathbf{c}_i^s \cap \mathbf{c}_j^t \neq \varnothing$, and $(\mathbf{c}_j^t \setminus (\mathbf{c}_i^s \cap \mathbf{c}_j^t)) \notin \{\varnothing, \mathbf{c}_q^s, q = 1, ..., C_S\}$.

Few-shot/Partially Labeled. In TACS, 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 TACS 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 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

[4] 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.

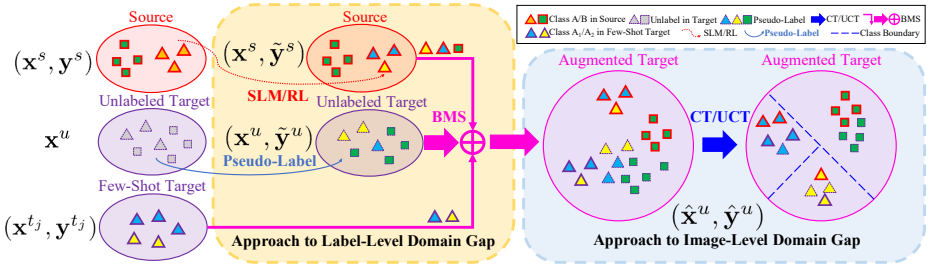

Fig. 2: Framework overview. Class A is an inconsistent taxonomy class (*e.g.,* "person") in the source domain, related to class A_1 (*e.g.,* "pedestrian") and A_2 (*e.g.,* "rider") in the target domain. Class B is a consistent taxonomy class. On the label level, SLM/RL module maps the inconsistent taxonomy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

image level, the image domain distribution difference between the source and target domain, $P_S \neq P_T$, still exists in TACS.

3.2 Our Approach to the TACS Problem

Motivation. Motivated by the technical challenge i) of the label level in Sec. 3.1, the stochastic label mapping (SLM) 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 and augment the few-shot supervision 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 the pseudo-label based self-training strategy. Following the self-training structure of [26], there are two components of our framework, namely a student network $\mathcal{F}_\theta$ and a mean-teacher network $\mathcal{F}_{\theta'}$, which are both semantic segmentation networks. The student network $\mathcal{F}_\theta$ is used to backpropagate the gradients and update θ according to the training loss. 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 training [32, 34]. During inference, the mean-teacher network $\mathcal{F}_{\theta'}$ is used to output the final segmentation map.

Framework Overview. The framework overview is shown in Fig. 2. The SLM and RL modules (Sec. 3.3) 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. 3.3). 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. 3.4) 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.

3.3 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, $\tilde{\mathbf{y}}^{s(m,n)} = \text{rand}(\mathbf{c}_p^t, \mathbf{c}_{p+1}^t, ..., \mathbf{c}_{p+q-1}^t)$, where the (m, n) is the (row, column) index. The $\text{rand}(\cdot)$ represents the uniformly discrete sampling function. With the obtained new labels $\tilde{\mathbf{y}}^s$, 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 the 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 loss of generality and for the 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, $\tilde{\mathbf{y}}^{s(m_i^s,n_i^s)} = \arg\max_c \mathbf{f}^{s(m_i^s,n_i^s)}$, if $\max_c \mathbf{f}^{s(m_i^s,n_i^s)} > \delta$, and $\arg\max_c \mathbf{f}^{s(m_i^s,n_i^s)} \in \{\mathbf{c}_p^t, ..., \mathbf{c}_{p+q-1}^t\}$. (m_i^s, n_i^s) is the index of the pixel corresponding to $\mathbf{c}_i^s$. 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* and *augment the few-shot supervision* 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 [11,44,45] 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 [34], 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 TACS. 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,

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

where the sampling class $\mathbf{c}_r$ is randomly selected from the available classes in $\tilde{\mathbf{y}}^s$. Following [34], 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,

$$\hat{\mathbf{x}}^u = \mathbf{m}^s \odot \mathbf{x}^s + (1 - \mathbf{m}^s) \odot (\mathbf{m}^{t_j} \odot \mathbf{x}^{t_j} + (1 - \mathbf{m}^{t_j}) \odot \mathbf{x}^u), \tag{2}$$

$$\hat{\mathbf{y}}^u = \mathbf{m}^s \odot \tilde{\mathbf{y}}^s + (1 - \mathbf{m}^s) \odot (\mathbf{m}^{t_j} \odot \mathbf{y}^{t_j} + (1 - \mathbf{m}^{t_j}) \odot \tilde{\mathbf{y}}^u). \tag{3}$$

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

3.4 Approach to the Image Level Domain Gap

Besides dealing with the label-level domain gap, we also need to tackle 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 [9, 10, 35]. 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 [38], the pixel-wise contrastive learning is proven to be helpful for semantic segmentation. However, it relies on ground truth label, which is unavailable for our unlabeled 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 $\mathcal{V}$ 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.,* $\mathcal{E}_\theta : \mathcal{S}, \mathcal{T} \rightarrow \mathcal{V}$. 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,

$$\mathcal{L}_{ct} = -\sum_h \sum_w \log \sum_{v^+ \in \mathcal{P}_v} \text{Contrast}(v, v^+), \tag{4}$$

$$\text{Contrast}(v, v^+) = \frac{\exp(v \cdot v^+ / \tau)}{\exp(v \cdot v^+ / \tau) + \sum_{v^- \in \mathcal{N}_v} \exp(v \cdot v^- / \tau)}, \tag{5}$$

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. (5) tries to learn similar features for the pixels of the same class, and learn discriminative features for the different class pixels, no matter whether 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 $\mathcal{U}(\cdot)$, formulated as,

$$\hat{\mathbf{u}}^u = \max_c \mathcal{F}_{\theta'}(\hat{\mathbf{x}}^u). \tag{6}$$

Then, based on Eq. (5), the uncertainty-rectified CT loss $\mathcal{L}_{uct}$ is formulated as,

$$\mathcal{L}_{uct} = -\sum_h \sum_w \hat{\mathbf{u}}^u(v)\hat{\mathbf{u}}^u(v^+)\,\text{Contrast}(v, v^+), \tag{7}$$

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

3.5 Joint Training

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

$$\mathcal{L}_{total} = \mathcal{L}_{bms} + \lambda_1\mathcal{L}_{slm} + \lambda_2\mathcal{L}_{rl} + \lambda_3\mathcal{L}_{uct} \tag{8}$$

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 loss in Eq. (8).

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 TACS, including the open, coarse-to-fine and implicitly-overlapping taxonomy setting. The DeepLabv2-ResNet101 [3,14] is adopted as the segmentation network. The baselines in Table 2-4 adopt the SOTA few-shot cross-domain semantic segmentation training strategy, *i.e.,* fine-tuning [46] and pseudo-label [26], to exploit the supervision from the few-shot labeled target domain. More experimental details 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 [30] is used as the source domain, while Cityscapes [8] is treated as the target domain. The

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

Method	Road	SW	Build	Wall*	Fence*	Pole*	TL	TS	Veg	Sky	Person	Rider	Car	Bus	MC	Bike	mIoU*	mIoU
ADVENT [37]	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 [42]	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 [23]	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 [34]	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 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 BMS module.

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 [37]	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 [42]	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 [23]	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 [34]	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 [39]	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

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*.

TACS: Open Taxonomy. The SYNTHIA dataset [30] is used as the source domain, and the Cityscapes dataset [8] 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.

TACS: Coarse-to-Fine Taxonomy. The GTA5 dataset [29] is utilized as the source domain, and the Cityscapes dataset [8] 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.

TACS: Implicitly-Overlapping Taxonomy. The Synscapes dataset [40] is treated as the source domain, while the Cityscapes dataset [8] is seen as the target domain. The label space of the source domain contains the *road*, *sidewalk*,

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	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 [37]	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 [42]	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 [23]	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 [34]	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 [39]	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

building, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *terrain*, *sky*, *person*, *rider* and *vehicle*. The *vehicle* class in source 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. 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 [37], the image translation based FDA [42], the self-training based IAST [23], and the data augmentation based DACS [34]. 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 other SOTA methods performance by a large margin, under the open, coarse-to-fine and implicitly-overlapping taxonomy settings. It validates the proposed framework for dealing with both of the image- and label-level domain gap. In Fig. 5, we show qualitative semantic segmentation results on the target domain.

Ablation Study. 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 implicitly-overlapping taxonomy settings. In different settings, the improvement brought by different modules are different. It is mainly because different settings in TACS touch diverse and broad aspects of inconsistent taxonomy. For example, the open taxonomy setting includes the new classes which are unseen or unlabeled in the source domain. The RL module is especially helpful to those unlabeled classes, *e.g.,* "wall" class. The

Table 4: Implicitly-Overlapping Taxonomy: Synscapes→Cityscapes. There are 3 classes (in gray) in the Cityscapes corresponding to the implicitly-overlapping taxonomy. "M": BMS. "*": with SLM.

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 [37]	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 [42]	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 [23]	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 [34]	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

SLM module is significantly beneficial under the coarse-to-fine taxonomy setting since each fine class is corresponding to one coarse class unambiguously. The CT/UCT module contribution difference is mainly related to the image-level difference, *e.g.,* the style difference of SYNTHIA, GTA, Synscapes. 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. It is also observed that the combination of SLM and other baseline methods, *e.g.,* ADVENT, FDA, IAST and DACS, does not necessarily bring the performance improvement. It is because the model prediction, when using SLM, is guided by the few-shot labeled target samples, but the baseline methods cannot effectively extract and exploit few-shot supervision with the previous SOTA few-shot cross domain semantic segmentation strategy, *i.e.,* fine-tuning [46] and pseudo-label [26]. Instead, our proposed BMS can augment and utilize the few-shot supervision effectively, guiding the model prediction when using SLM.

Partially Labeled/Oracle. In Table 2, Table 3, and Table 4, under the open, coarse-to-fine and implicitly-overlapping taxonomy settings, 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 to the few-shot performance, the partially labeled performance is further improved 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 method serves as a strong baseline, but still provides the potential to develop stronger algorithms for the TACS problem.

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.

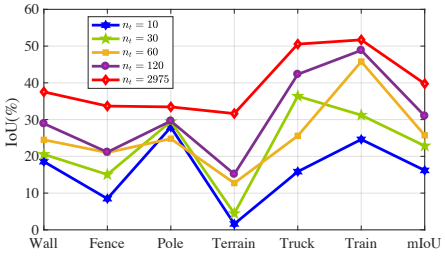

Fig. 3: Performance of inconsistent taxonomy classes under open taxonomy setting, varying n^t.

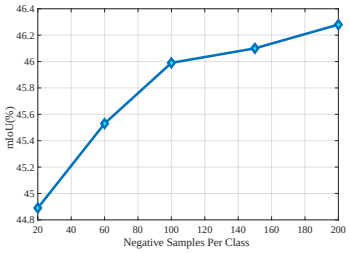

Fig. 4: Negative samples number study for contrastive learning, under M+UCT in Table 2.

Contrastive Learning. In Fig. 4, the performance when varying the number of negative samples in the contrastive learning is shown. It is observed that the performance increases as more samples are taken. Balancing the performance and memory, we adopt 100 samples per class. In Fig. 5, we compare the t-SNE visualization [22] of the feature embedding of the model trained with/without UCT, taking open taxonomy setting as example. It verifies the contrastive learning is helpful to train the cross-domain invariant and class-discriminative features.

5 Conclusion

We propose the new TACS problem, allowing inconsistent taxonomies between the source and the target domain in the cross-domain semantic segmentation. Three typical types of inconsistent taxonomies are identified. To resolve TACS, the mixed-sampling, pseudo-label and contrastive learning based techniques are developed. Extensive experiments prove the effectiveness of our approach.

Acknowledgements. This project was funded by the EU Horizon 2020 research and innovation program under grant agreement No. 820434. This project was also supported by the European Lighthouse on Secure and Safe AI (ELSA) Project, a Facebook Academic Gift on Robust Perception (INFO224), and the ETH Future Computing Laboratory (EFCL). Special thanks goes to Dr. Wenguan Wang.

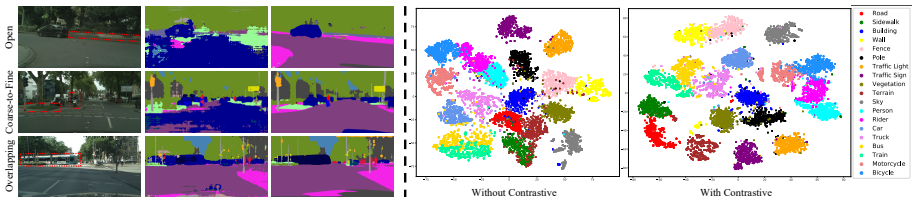

Fig. 5: **Left:** Qualitative results under different inconsistent taxonomy settings. Each group has the RGB image (left), the results without adaptation (middle) and adapted with our method (right). Refer to the red box region for the adaptation of the inconsistent taxonomy classes. **Right:** t-SNE visualization of the features with/without contrastive learning under the open taxonomy setting.

References

1. Bucher, M., Vu, T.H., Cord, M., Pérez, P.: Handling new target classes in semantic segmentation with domain adaptation. arXiv preprint arXiv:2004.01130 (2020) 3, 4
2. Cao, Z., Ma, L., Long, M., Wang, J.: Partial adversarial domain adaptation. In: ECCV (2018) 2
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI **40**(4), 834–848 (2017) 10
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020) 4
5. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. In: NeurIPS (2020) 4
6. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 4
7. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: CVPR (2018) 1
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016) 2, 10, 11
9. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML (2015) 3, 9
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(1), 2096–2030 (2016) 9
11. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. arXiv preprint arXiv:2012.07177 (2020) 8
12. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020) 4
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020) 4
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 10
15. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: ICML (2018) 1
16. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016) 3
17. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: CVPR (2019) 4
18. Kundu, J.N., Venkatesh, R.M., Venkat, N., Revanur, A., Babu, R.V.: Class-incremental domain adaptation. In: ECCV (2020) 2, 3
19. Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V.: MSeg: A composite dataset for multi-domain semantic segmentation. In: CVPR (2020) 2
20. Liu, Z., Miao, Z., Pan, X., Zhan, X., Lin, D., Yu, S.X., Gong, B.: Open compound domain adaptation. In: CVPR (2020) 1

21. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML (2015) 3
22. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. JMLR **9**(11) (2008) 14
23. Mei, K., Zhu, C., Zou, J., Zhang, S.: Instance adaptive self-training for unsupervised domain adaptation. In: ECCV (2020) 4, 11, 12, 13
24. Motiian, S., Jones, Q., Iranmanesh, S.M., Doretto, G.: Few-shot adversarial domain adaptation. In: NeurIPS (2017) 4
25. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV (2017) 2
26. Olsson, V., Tranheden, W., Pinto, J., Svensson, L.: Classmix: Segmentation-based data augmentation for semi-supervised learning. In: WACV (2021) 6, 10, 13
27. Panareda Busto, P., Gall, J.: Open set domain adaptation. In: ICCV (2017) 2, 3
28. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV (2020) 4
29. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV (2016) 11
30. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR (2016) 10, 11
31. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: ECCV (2018) 3
32. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017) 6
33. Teshima, T., Sato, I., Sugiyama, M.: Few-shot domain adaptation by causal mechanism transfer. In: ICML (2020) 4
34. Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: Dacs: Domain adaptation via cross-domain mixed sampling. In: WACV (2021) 1, 4, 6, 8, 11, 12, 13
35. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018) 1, 3, 4, 9
36. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., Van Gool, L.: Scan: Learning to classify images without labels. In: ECCV (2020) 4
37. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR (2019) 1, 4, 11, 12, 13
38. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L.: Exploring cross-image pixel contrast for semantic segmentation. arXiv preprint arXiv:2101.11939 (2021) 4, 9
39. Wang, Z., Wei, Y., Feris, R., Xiong, J., Hwu, W.M., Huang, T.S., Shi, H.: Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation. In: CVPR Workshops (2020) 11, 12
40. Wrenninge, M., Unger, J.: Synscapes: A photorealistic synthetic dataset for street scene parsing. arXiv preprint arXiv:1810.08705 (2018) 11
41. Xie, E., Ding, J., Wang, W., Zhan, X., Xu, H., Li, Z., Luo, P.: Detco: Unsupervised contrastive learning for object detection. arXiv preprint arXiv:2102.04803 (2021) 4
42. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: CVPR (2020) 11, 12, 13

43. You, K., Long, M., Cao, Z., Wang, J., Jordan, M.I.: Universal domain adaptation. In: CVPR (2019) 2, 3
44. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV (2019) 8
45. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018) 8
46. Zhang, J., Chen, Z., Huang, J., Lin, L., Zhang, D.: Few-shot structured domain adaptation for virtual-to-real scene parsing. In: ICCV Workshops (2019) 4, 10, 13
47. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: ICCV (2017) 3
48. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV (2018) 3