mDALU: Multi-Source Domain Adaptation and Label Unification with Partial Datasets

Rui Gong [1], Dengxin Dai [1], Yuhua Chen [1], Wen Li [3], Luc Van Gool [1,2]
[1] Computer Vision Lab, ETH Zurich, [2] VISICS, KU Leuven, [3] UESTC
{gongr, dai, yuhua.chen, vangool}@vision.ee.ethz.ch, liwenbnu@gmail.com

Abstract

Object recognition advances very rapidly these days. One challenge is to generalize existing methods to new domains, to more classes and/or to new data modalities. In order to avoid annotating one dataset for each of these new cases, one needs to combine and reuse existing datasets that may belong to different domains, have partial annotations, and/or have different data modalities. This paper treats this task as a multi-source domain adaptation and label unification (mDALU) problem and proposes a novel method for it. Our method consists of a partially-supervised adaptation stage and a fully-supervised adaptation stage. In the former, partial knowledge is transferred from multiple source domains to the target domain and fused therein. Negative transfer between unmatched label space is mitigated via three new modules: domain attention, uncertainty maximization and attention-guided adversarial alignment. In the latter, knowledge is transferred in the unified label space after a label completion process with pseudo-labels. We verify the method on three different tasks, image classification, 2D semantic image segmentation, and joint 2D-3D semantic segmentation. Extensive experiments show that our method outperforms all competing methods significantly.

1. Introduction

The development of object recognition is powered by two pillars now: large-scale data annotation and deep neural networks. With new applications coming out every day, researchers need to constantly develop new methods and create new datasets. While we are able to develop novel neural networks for all the new tasks, creating a new dataset for every new task can hardly work due to its huge cost. In the literature, a diverse set of learning paradigms, such as self-learning [14], semi-supervised learning [46] and transfer learning [6], have been developed to rescue. We enrich this repository by developing a method to combine multiple existing datasets that have been annotated in different domains, for smaller-scale tasks (fewer classes), and/or with fewer data modalities. The importance of the method can be justified by the fact that as the field grows, research goals will become more and more ambitious, so object recognition methods for more classes, in new domains, and/or with more data modalities is becoming a necessity.

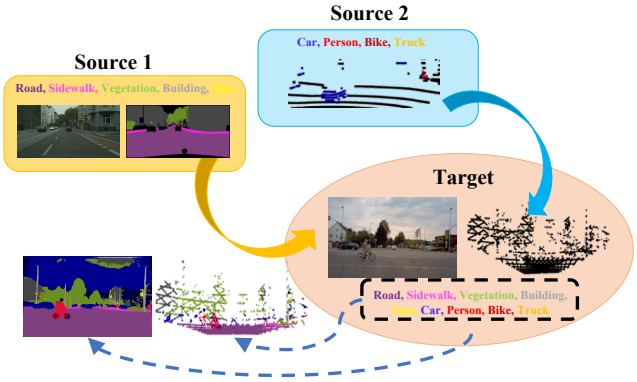


Figure 1: **mDALU: multi-source domain adaptation and label unification with partial datasets.** With partially labeled source domain and unlabeled target domain, mDALU allows knowledge from multiple source domains, partial label space and different modalities to be transferred to the target domain in a complete label space.

In this work, we tackle the problem as a multi-source domain adaptation and label unification (mDALU) problem. In this setting, there are multiple source domains and one target domain. In each source domain, only samples (images or pixels) belonging to a subset of classes are labeled; the rest are unlabeled. The subsets of classes having labels can be different over different source domains. Further, the data modalities in different source domains can also be different, *e.g.*, one contains images and the other contains LiDAR point clouds (cf. Fig. 1). We compare mDALU to other domain adaptation settings in Table 1 and find that mDALU offers a flexible setting. The goal is to obtain an object recognition method for all classes on target domain.

Domain Adaptation Setting	Can Handle Multiple Source Domains?	Can Handle Multiple Data Modalities?	Can Handle Different Label Spaces of Source Domains?	Change of Label Space Size from Source to Target Domain	Can Handle Partial Annotations?
Unsupervised Domain Adaptation [10]	No	No	–	Same Size	No
Partial Domain Adaptation [3]	No	No	–	Reduced	No
Multi-Source Domain Adaptation [30, 50]	Yes	No	No	Same Size	No
Category-Shift Multi-Source Domain Adaptation [47]	Yes	No	Yes	Increased	No
Cross-Modal Unsupervised Domain Adaptation [19]	Yes	Yes	No	Same Size	No
Multi-Source Domain Adaptation and Label Unification (mDALU)	**Yes**	**Yes**	**Yes**	**Increased**	**Yes**

Table 1: Comparison between our mDALU and other domain adaptation settings (see the details in Sec. 2). It is clear that mDALU offers a very flexible and general setting.

There are some challenges towards this goal. The first one is the notorious issue of negative transfer. While negative transfer is an issue also for standard transfer and multi-task learning, it is especially severe in our mDALU task due to the influence of unlabeled classes. To address this, we propose three novel modules, termed domain attention, uncertainty maximization and attention-guided adversarial alignment, to suppress making confident predictions for unlabeled classes and to enable robust feature distribution alignment between the source domains and the target domain. The method with the aforementioned modules and attention-guided prediction fusion is able to generate good results in the unified label space and on the target domain. In order to further improve the results, we need to solve another challenge, which is how to fuse the supervision from partial label space, to transfer in the unified label space. To this aim, we propose a pseudo-label based supervision fusion module. In particular, we generate pseudo-labels for the unlabeled classes in all source domains, and pseudo-labels for all classes in the target domain. A standard supervised learning is then performed to train a model in the unified label space.

To showcase the effectiveness and the applicability of our method, we evaluate it on three different tasks: image classification, 2D semantic image segmentation, and joint 2D-3D semantic segmentation. In these settings, synthetic data, real data, images and LiDAR point clouds are all involved. Also, non-overlapping, partially-overlapping and fully-overlapping label spaces across source domains are all covered. Extensive experiments show that our method is able to successfully distill semantic knowledge from partial datasets and perform semantic object recognition in the target domain for a *complete* set of class labels. Our method also outperforms all competing methods significantly in all considered settings.

2. Related Work

Domain Adaptation. The work belongs to multi-source domain adaptation. Transfer learning and domain adaptation have been extensively studied in the past years to mitigate the domain gap. Several effective strategies have been developed such as minimizing maximum mean discrepancy [42], moment matching [30], adversarial domain confusion [10], entropy regularization [44], and curriculum domain adaptation [9]. While great progress has been achieved, most algorithms focus on the single-source adaptation setting. This limits the methods from being used when data is collected from multiple source domains.

That is why multi-source domain adaptation (msDA) methods are proposed [8]. Some of the aforementioned learning strategies have been successfully applied in the multi-source setting. For instance, adversarial alignment for multiple source domain adaptation is studied in [49] and a k-way domain discriminator for digit classification and object recognition has been proposed in the Deep Cocktail Network (DCTN) [47]. Moment alignment across multiple domains has proven effective in [30]. Hoffman *et al*. have derived new normalized solutions for the cross-entropy loss and and other similar losses for the multi-source adaptation setting [16]. These methods are proposed mainly for image classification. msDA has also been applied to semantic segmentation. For example, cycle-consistency image translation and adversarial domain aggregation are employed to transfer knowledge from multiple sources in [50]. The method assumes the same label space for all domains.

Multi-modal domain adaptation [20, 27] explores the multi-modal nature of data. In addition to adversarial alignment and training with pseudo labels, they use the correspondence of modalities for self-supervised alignment.

In the view of making domain adaptation more practically useful, our method is related to open set domain adaptation [29, 40, 32]. They work with one source domain and address the challenge that some classes in the target domain are unseen in the source domain. We employ a closed solution space, *i.e.,* , the label space of the target domain is the union of that of all datasets in multiple source domains.

Learning from multiple datasets. Several successful methods have been proposed to learn a universal model from multiple existing datasets, for image classification [33, 34], object detection [45], or semantic segmentation [22]. The main goal is to learn a single universal network that can represent different domains with minimal number of domain-specific parameters. The domain-specific parameters are embodied as residual adapters [33, 34] or domain-specific decoders [22]. Those methods do not consider domain shifts and label space unification. The closest work to

ours is the object detection method by Zhao *et al*. [51]. It also performs label space unification from multiple datasets with partial annotations. Their method is based on self-training with pseudo-labels. Our method contains novel modules such as domain attention, uncertainty maximization, attention-guided adversarial alignment, and attention-guided label fusion. We also consider multiple tasks and multiple data modalities.

Recently, Lambert *et al*. [24] presented a composite dataset that unifies semantic segmentation datasets from different domains by reconciling the taxonomies, merging and splitting classes manually. Unal *et al*. [43] learns 3D semantic segmentation and 3D object detection together with two partial datasets. Our method also shares some similarity to incremental learning methods [35, 39]. Instead of learning in the standard batch setting, they incrementally learn about new classes/tasks when new data becomes available. Our goal is different. We fuse the supervision and unify the label spaces of multiple existing datasets to perform object recognition of more classes in a new domain. Also, they only consider image classification task.

Partial domain alignment. Recently, there is a stream of research on partial domain alignment [2, 48, 3, 21]. Their focus is to transfer knowledge from existing large-scale domains (e.g. 1K classes) to unknown small-scale domains (e.g. 20 classes) for customized applications. Our method, however, addresses the *opposite* problem – it completes the label space and fuses the supervision of multiple existing datasets, which have been created in multiple source domains, to achieve a bigger goal of recognizing *all* classes in the target domain. As the field advances, research goals will become more and more ambitious. Therefore, it is imperative to have methods that are able to combine (partial) datasets in order to recognize objects of more classes. Furthermore, a diverse set of tasks including image classification, 2D image semantic segmentation, and 2D-3D image-LiDAR joint semantic segmentation are considered in this work instead of image classification only.

3. Approach

3.1. Problem Statement

For the problem of mDALU, we are given K source domains $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_K$. The K source domains contain the samples from K different distributions $P_{S_1}, P_{S_2}, ..., P_{S_K}$, which are labeled with $C_1, C_2, ..., C_K$ classes respectively and noted as $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K$ label space. Then the union of the label space $\mathcal{C}_i, i = 1, ..., K$ composes the unified and complete label space $\mathcal{C}_\cup = \mathcal{C}_1 \cup \mathcal{C}_2 \cdots \mathcal{C}_K$, including $C_\cup$ classes. Besides, the target domain $\mathcal{T}$ is given, containing samples from the distribution P_T. Denoting the source samples $\mathbf{x}^{s_i} \in \mathcal{S}_i, i = 1, ..., K$ and the target samples $\mathbf{x}^t \in \mathcal{T}$, we have $\mathbf{x}^{s_i} \sim P_{S_i}, \mathbf{x}^t \sim P_T$, $P_{S_1} \neq P_{S_2} \neq ... \neq P_{S_K} \neq P_T$. The mDALU problem aims at training the model on the K source domains $\mathcal{S}_i, i = 1, ..., K$, labeled with C_i classes in each, and the unlabeled target domain $\mathcal{T}$, to improve the performance of the model on the target domain $\mathcal{T}$ in the unified label space $\mathcal{C}_\cup$. We use $\mathbf{y}^{s_i}$ to indicate the ground-truth label map of $\mathbf{x}^{s_i}$. Note that we present most of our approach with the notation of 2D semantic image segmentation. The translation to image classification and 3D point cloud segmentation is straightforward – by replacing pixels with images and by replacing pixels with 3D LiDAR points.

3.2. Our Approach to mDALU problem

As shown in Fig. 2, there are two stages in our approach, the partially-supervised adaptation stage and the fully supervised adaptation stage. In order to realize adaptation under partial supervision, we propose three modules: domain attention module (DAT), uncertainty maximization module (UM) and attention-guided adversarial alignment module (A^3) for the first stage. Then in the second stage, we use pseudo-labels for supervision fusion (PSF) and further learning. Below we provide details of all these components. We first describe a basic version of our method for partially-supervised learning with DAT, which will be followed by UM and A^3 to enhance the adaptation ability. Finally, we present PSF.

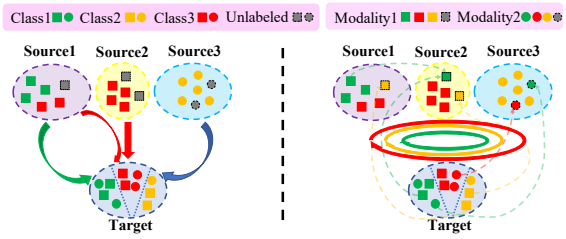
3.2.1 Partially-Supervised Learning

Different segmentation networks $G_i, i = 1, ..., K$ are adopted for different source domains $\mathcal{S}_i$. While their annotations are done in partial label space $\mathcal{C}_i$, we train each network G_i in the unified label space $\mathcal{C}_\cup$ – some classes have no training data – with a standard cross-entropy loss $\mathcal{L}_{\text{psu}}$. The network G_i is composed of a feature extractor E_i and a label predictor B_i, *i.e.,* $G_i = \{E_i, B_i\}$. While we can average the results of these models directly in the target domain for predictions in the unified label space, named as multi-branch (MBR) fusion, this will generate poor results. This is because the predictions of each model G_i for its unlabeled classes in $\mathcal{C}_\cup \setminus \mathcal{C}_i$ can be arbitrary numbers and dominate the averaged results. We thus propose the DAT module, which learns the attention map for G_i to signal on which area its prediction is reliable for more effective fusion.

The attention map $\mathbf{a}^{s_i}$ in domain $\mathcal{S}_i$ is defined as:

$$\mathbf{a}^{s_i}(h,w) \begin{cases} = 1, \text{if } \mathbf{y}^{s_i}(h,w) \in \mathcal{C}_i \\ = 0, \text{if } \mathbf{y}^{s_i}(h,w) = \texttt{void}, \end{cases} \tag{1}$$

where (h, w) are pixel indexes and `void` means no label. We train an attention network M_i for each source domain $\mathcal{S}_i$. The attention maps are predicted as $\tilde{\mathbf{a}}^{s_i} = M_i(\mathbf{x}^{s_i})$ and $\tilde{\mathbf{a}}_i^t = M_i(\mathbf{x}^t)$. The attention network M_i is composed of the feature extractor E_i and a new label predictor B_i^M:



(a) Partially-Supervised Adaptation (b) Fully-Supervised Adaptation

Figure 2: Illustration of our approach to mDALU problem. Our approach includes two stages, (a) partially supervised adaptation stage and (b) fully-supervised adaptation stage. More detailed network structure is put into the supplementary due to the space limit.

$M_i = \{E_i, B_i^M\}$. M_i is trained under an MSE loss $\mathcal{L}_{\text{att}}$, together with G_i under a multi-task setting.

3.2.2 Inference via Attention-Guided Fusion

We feed an image $\mathbf{x}$ into semantic segmentation networks G_k to generate the corresponding probability maps $\hat{\mathbf{p}}_k \in [0,1]^{H\times W\times C_\cup}$, and into different attention networks M_k to generate attention maps $\hat{\mathbf{a}}_k$. Then we fuse the predictions by averaging $\hat{\mathbf{p}}_k$ weighted by $\hat{\mathbf{a}}_k$:

$$\mathbf{f} = \frac{\sum_{i=1}^{K} \hat{\mathbf{a}}_k \otimes \hat{\mathbf{p}}_k}{\sum_{j=1}^{C_\cup}(\sum_{i=1}^{K} \hat{\mathbf{a}}_k \otimes \hat{\mathbf{p}}_k)^{(j)}}, \tag{2}$$

where $(\sum_{i=1}^{K} \hat{\mathbf{a}}_k \otimes \hat{\mathbf{p}_k})^{(j)}$ represents the probability of the j^{th} class. The final predicted class labels can then be obtained via a normal `argmax` operation.

3.2.3 Uncertainty Maximization (UM)

While we have the attention-guided fusion, the wrong prediction to the unlabeled classes can still have negative effects for our across-domain prediction. In order to further reduce the effects of these unlabeled samples (pixels, images or LiDAR points), $\mathbf{x}_u^{s_i}$, we propose a module specifically to maximize uncertainties of the predictions on unlabeled samples. In particular, $G_i(\mathbf{x}_u^{s_i})$ is expected to equally spread the probability mass to all classes, *i.e.,* obeying the uniform categorical distribution $\mathcal{U}(C_\cup)$. The probability density function $q(j)$ of $\mathcal{U}(C_\cup)$ is formulated as $q(j) = \frac{1}{C_\cup}$, where $j = 1, 2, ..., C_\cup$ is to represent different classes. The probability distribution of the network prediction on unlabeled samples $G_i(\mathbf{x}_u^{s_i})$ is denoted as $p(j) = G_i(\mathbf{x}_u^{s_i})^{(j)}$, where $G_i(\mathbf{x}_u^{s_i})^{(j)}$ represent the probability of the j^{th} class. In order to maximize the uncertainty of the prediction on the unlabeled samples, the distribution distance between $p(j)$ and $q(j)$ is expected to be minimized. Following the distribution distance measurement metric in [5], we adopt the Pearson χ^2-divergence for measuring the distribution distance, which is formulated as,

$$D_{\chi^2}(p||q) = \int_j ((\frac{p(j)}{q(j)})^2 - 1)q(j), \tag{3}$$

$$D_{\chi^2}(p||q) = C_\cup \sum_{j=1}^{C_\cup} p(j)^2 - 1. \tag{4}$$

On the basis of the Eq. (4), we propose the square loss $\mathcal{L}_{um}$ for minimizing the Pearson χ^2-divergence, *i.e.,* maximizing the uncertainty of the prediction on the unlabeled samples. The $\mathcal{L}_{um}$ can be written as,

$$\mathcal{L}_{um} = \sum_{j=1}^{C_\cup}(G_i(\mathbf{x}_u^{s_i})^{(j)})^2. \tag{5}$$

This loss is used together with other losses.

3.2.4 Attention-Guided Adversarial Alignment (A^3)

It has been proven in the literature that adversarial feature alignment is effective for domain adaptation. We extend the idea to our mDALU task. For adversarial alignment, one discriminator D_i is exploited for each source domain, to align the distribution between the source domain $\mathcal{S}_i$ and the target domain $\mathcal{T}$. In the general unsupervised domain adaptation, the discriminator training loss $\mathcal{L}_d$ and the adversarial loss $\mathcal{L}_{adv}$ [41] for the source domain $\mathcal{S}_i$ and the target domain $\mathcal{T}$ is defined as,

$$\mathcal{L}_{adv}^{(i)}(\mathbf{x}^t) = -\log(D_i(G_i(\mathbf{x}^t))) \tag{6}$$

$$\mathcal{L}_{d}^{(i)}(\mathbf{x}_i^{s_i}, \mathbf{x}^t) = -\log(D_i(G_i(\mathbf{x}^{s_i}))) - \log(1 - D_i(G_i(\mathbf{x}^t))). \tag{7}$$

However, in our mDALU problem, there is no ground truth label guidance available for the unlabeled samples. A direct alignment between the source domain and the target domain will cause the negative transfer effect, *i.e.,* transfer incorrect knowledge of the unlabeled parts in source domain to the target domain.

Here, we again use our attention map $\mathbf{a}^{s_i}$ for alleviate this problem by proposing an attention-guided adversarial loss:

$$\mathcal{L}_{a^3}^{(i)}(\mathbf{x}^t) = -\log(D_i(G_i(\mathbf{x}^t) \otimes M_i(\mathbf{x}^t))), \tag{8}$$

$$\mathcal{L}_{d}^{(i)}(\mathbf{x}_i^{s_i}, \mathbf{x}^t) = -\log(D_i(G_i(\mathbf{x}^{s_i}) \otimes M_i(\mathbf{x}^{s_i}))) - \log(1 - D_i(G_i(\mathbf{x}^t) \otimes M_i(\mathbf{x}^t))), \tag{9}$$

where $\otimes$ represents the element-wise multiplication.

Then the overall loss for our method at the first stage is:

$$\mathcal{L}_{all} = \mathcal{L}_{psu} + \mathcal{L}_{att} + \mathcal{L}_{um} + \lambda \sum_{i=1}^{K} \mathcal{L}_{a^3}^{(i)}, \tag{10}$$

where λ is the hyper-parameter to balance between the attention-guided adversarial loss and other losses. The whole optimization objective for our first partially-supervised domain adaptation stage can be formulated as:

$$\min_{G_i} \max_{D_i} \mathcal{L}_{all}. \tag{11}$$

3.2.5 Pseudo-Label Based Supervision Fusion (PSF)

In the first 'partially-supervised' adaptation stage, knowledge in different label spaces $\mathcal{C}_i$ is transferred between different domains. In the second 'fully-supervised' learning stage, we aims at learning and transferring knowledge in the complete and unified label space $\mathcal{C}_\cup$ between all domains. In order to realize that, we complete the label space for all the related domains $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_K, \mathcal{T}$ with pseudo-labels, *i.e.,* fuse the supervision from different label space $\mathcal{C}_i$ to get the complete and unified supervision $\mathcal{C}_\cup$. Here we present our pseudo-label based supervision fusion (PSF) method.

In order to complete the label space on the source domain $\mathcal{S}_i$, we feed each of the source image sample $\mathbf{x}^{s_i}$ into every semantic model $G_k, k = 1, ..., K$, to generate 'partial' semantic probability map $\hat{\mathbf{p}}_k^{s_i} \in [0,1]^{H \times W \times C_\cup}$ and to every attention network $M_k, k = 1, ..., K$ for the attention map $\hat{\mathbf{a}}_k^{s_i} \in [0,1]^{H \times W}$. The fused prediction $\mathbf{f}^{s_i}$ is obtained via Eq. 2. We denote the predicted label map as $\bar{\mathbf{y}}^{s_i}$, generated by using an `argmax` operation over $\mathbf{f}^{s_i}$. The 'pseudo-label' map $\hat{\mathbf{y}}^{s_i}$ for source domain s_i is defined as:

$$\hat{\mathbf{y}}^{s_i}(h,w) = \begin{cases} \mathbf{y}^{s_i}(h,w), \text{if } \mathbf{y}^{s_i}(h,w) \neq \texttt{void} \\ \bar{\mathbf{y}}^{s_i}(h,w) \text{if } \mathbf{y}^{s_i}(h,w) = \texttt{void} \text{ and } \mathbf{f}^{s_i(h,w,\bar{\mathbf{y}}^{s_i}(h,w))} > \delta \\ \texttt{void}, \text{ otherwise} \end{cases} \tag{12}$$

where δ is a threshold determining whether to select the predicted pseudo-label.

On the target domain $\mathcal{T}$, since no ground truth label is available, we obtain the pseudo label directly from the predicted label map $\bar{\mathbf{y}}^t$ (obtained from $\mathbf{f}^t$ via an `argmax`):

$$\hat{\mathbf{y}}^t(h,w) = \bar{\mathbf{y}}^t(h,w) \text{ if } \mathbf{f}^{t(h,w,\bar{\mathbf{y}}^t(h,w))} > \delta. \tag{13}$$

By using the generated fused pseudo-label $\hat{\mathbf{y}}^{s_i}, \hat{\mathbf{y}}^t, i = 1, ..., K$, we complete the label space from $\mathcal{C}_i$ to $\mathcal{C}_\cup$ for the source domain $\mathcal{S}_i$, and from $\emptyset$ to $\mathcal{C}_\cup$ for the target domain $\mathcal{T}$. We then train the network G for all the related domains $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_K, \mathcal{T}$ with all the datasets in the unified label space. In total, the loss $\mathcal{L}_{fsa}$ for our second 'fully-supervised' adaptation stage is:

$$\mathcal{L}_{fsa} = \sum_{i=1}^{K} \mathcal{L}_{ce}^{s_i} + \mathcal{L}_{ce}^{t}, \tag{14}$$

where $\mathcal{L}_{ce}$ is the standard cross-entropy loss.

4. Experiments

We evaluate the effectiveness of our method for mDALU problem under different settings. We build benchmarks for image classification, 2D semantic image segmentation, and 2D-3D cross-modal semantic segmentation. We first introduce the benchmarks, and then compare our method to other state-of-the-art (SOTA) methods on them.

4.1. Image Classification: Setup

In the classification benchmark, in order to evaluate the classification model adaptation and label unification performance under diverse appearances and backgrounds images, we adopt the digits classification images, from different datasets, as source domains and target domain. There are three datasets involved in the classification setting, the MNIST [25], the Synthetic Digits [10], and the SVHN [28], which are named as "MT", "SYN" and "SVHN", respectively. Each time, one of them is taken as the target domain, while the other two as source domains. There are 10 classes, from '0' to '9', measured on the target domain. In our main setting, we adopt the most difficult setup to evaluate different methods, where the label space of different source domains are non-overlapping. Only half classes are labeled in each of the source domains. We also compare our methods with other SOTA methods in the partially-overlapping situation. We also studied how performance changes as the number of overlapping classes increases. For a fair comparison, we adopt the same network architecture used in [30] for all the methods. The classification performance is evaluated for all the ten classes in the target domain.

4.2. Image Classification: Results

Comparison with SOTA. Table 2 shows the comparison results between our method and other SOTA methods which include 1) unsupervised domain adaptation method, DANN [10], 2) category-shift unsupervised domain adaptation method, DCTN [47], 3) multi-source unsupervised domain adaptation method, M^{3}SDA [30], and 4) label unification method, AENT [51]. It can be seen that without the pseudo label (PL) generation part, other domain adaptation based methods, DANN, DCTN, and M^{3}SDA show the negative transfer effect, or have similar performance to the baseline trained with source data only. This is due to that each of the source domain can only provide the guidance for a partial label space, and the adaptation in the partial label space guides the prediction on the target domain to the biased label space when training with data from different source domains. This makes the prediction on the target domain contradictory, and the model hard to be adapted to the complete label space. On the contrary, the label-unification based method, AENT, obtained a performance gain of 4.25%, from 60.65% to 64.90%, compared with the

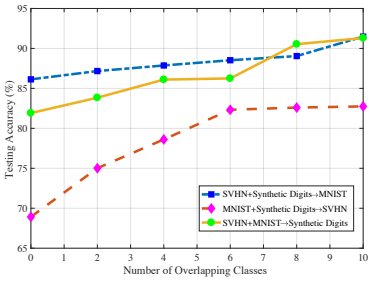


Figure 3: Accuracy in the target domain as a function of the number of over-lapping classes between the source domains.

Method	MT	SYN	SVHN	Avg
Source	76.76 ± 0.63	61.77 ± 1.05	43.42±1.89	60.65±1.19
DANN[10]	77.30 ± 2.57	60.31 ± 0.99	41.65±2.34	59.75±1.97
DANN *	71.29 ± 0.48	55.94 ± 0.51	35.60 ± 1.63	54.28 ± 0.87
DCTN [47]	68.10±0.2	62.72±0.30	48.11±0.57	59.64±0.36
DCTN *	72.01 ± 1.22	63.33 ± 0.20	49.34 ± 1.28	61.59 ± 0.90
M^{3}SDA [30]	76.56±0.71	61.25±2.33	43.13±3.55	60.31±2.20
M^{3}SDA *	72.50 ± 2.64	55.92 ± 1.04	36.24 ± 1.70	54.89 ± 1.79
AENT[51]	73.24±1.76	68.66±1.32	52.80 ± 0.92	64.90 ± 1.33
Ours w/o PSF	**81.23±0.92**	**78.97±0.45**	**65.20±0.58**	**75.13±0.65**
DCTN w/ PL [47]	73.40±0.85	65.63±0.43	52.12±0.07	63.72 ± 0.45
AENT[51] w/ PL	78.56±1.23	70.25 ± 0.39	59.24 ± 1.01	69.35 ± 0.88
Ours	**86.18±0.45**	**81.91±0.33**	**68.92±0.81**	**79.00 ± 0.53**

Table 2: Quantitative comparison between our method and other SOTA methods, under image classification setting. "M", "SYN", and "SVHN" represent the target domain. "PL" represents to add the pseudo-label training module, which is specifically adjusted according to their own papers design. * represents to remove the unlabeled samples in the training data. We realize AENT on classification by utilizing the ambiguity cross entropy loss proposed in [51]. The best results are denoted in bold.

source-only baseline. This is because it uses an ambiguity cross entropy loss, to avoid the prediction of the source domain data being restricted in a partial label space.

In our first partially-supervised adaptation stage, by exploiting our domain attention (DAT), uncertainty maximization square loss (UM), and the attention-guided adversarial alignment (A^3) module, the performance is further improved to 75.13%, which proves the effectiveness of our DAT, UM and A^3 module for preventing the negative transfer effect. After the second 'fully-supervised' adaptation stage, by adding the PSF module, our model highly outperforms the DCTN [47] and AENT [51], both with pseudo-label training, by 15.28% and 9.65%, respectively. It proves the effectiveness of our whole method for domain adaptation, label space completion and supervision fusion. In Table 3, the performance comparison between different ablations and our full model are shown. It proves that each part of the full model, the DAT, UM, A^3, and PSF all contribute to our final performance.

Partially Overlapping. In Fig. 3, it is shown that the testing accuracy on the target domain increases, as more and more common classes on the source domain are available. In Table 4, we compare the model performance of our method with other SOTA methods when the source domains are partially overlapping, with 4 common classes. It is shown that our method still highly outperforms the adaptation-based methods, DANN, DCTN, M^{3}SDA, and the label unification based method, AENT, 82.40% v.s. 69.37%, 70.58%, 69.56%, 76.73%. It further verifies the effectiveness of our model in partially overlapping situation.

MBR	UM	A^3	PSF	MT	SYN	SVHN	Avg
				76.76 ± 0.63	61.77 ± 1.05	43.42±1.89	60.65±1.19
✓				72.21±1.89	62.41±0.58	50.24±1.23	61.62±1.23
✓	✓			84.74±0.54	76.12±0.85	58.39±0.57	73.08± 0.65
✓	✓	✓*		81.38±0.79	78.20±1.3	65.12±0.64	74.90 ± 0.91
✓	✓	✓		81.23±0.92	78.97±0.45	65.20±0.58	75.13 ± 0.65
✓	✓	✓	✓	**86.18±0.45**	**81.91±0.33**	**68.92±0.81**	**79.00 ± 0.53**

Table 3: Ablation study under the image classification setting. MBR: multi-branch network, *i.e.,* adopts different networks G_i for different source domains. * represents there is no adversarial part in A^3 module, *i.e.,* only DAT module.

Method	MT	SYN	SVHN	Avg
Source	82.10±1.50	73.37± 0.67	57.50±1.93	70.99 ± 1.37
DANN[10]	80.13±1.60	72.97±0.49	55.00±0.73	69.37 ± 0.94
DCTN[47]	78.56±0.47	72.33 ± 0.04	60.86±0.21	70.58 ± 0.24
M^{3}SDA[30]	81.52 ± 1.55	72.91 ± 0.68	54.26±0.66	69.56 ± 0.96
AENT[51]	79.12 ± 1.07	81.99 ± 0.87	69.07 ± 1.93	76.73 ± 1.29
Ours w/o PSF	**85.39 ± 1.32**	**85.33± 1.21**	**76.48±1.31**	**82.40 ± 1.28**

Table 4: Quantitative comparison between our method and other SOTA methods under image classification setting, when source domains are partially overlapping and have 4 common classes. "MT", "SYN", and "SVHN" represent the target domain. The best results are denoted in bold.

4.3. 2D Semantic Image Segmentation: Setup

In this single modal semantic segmentation setting, in order to evaluate the semantic segmentation model adaptation and label unification performance under the same data mode, we adopt the synthetic-to-real image semantic segmentation setup. The synthetic image datasets GTA5 [36] and the SYNTHIA [38] are taken as the source domains, while the real image dataset Cityscapes [7] is used as the target domain. Information of 19 classes needs to be transferred to the Cityscapes dataset. In our main setting, the label space of the SYNTHIA and GTA5 is non-overlapping. In SYNTHIA dataset, the label of 7 classes are available, including road, sidewalk, building, vegetation, sky, person and car. In GTA5 dataset, the label of 12 classes are available, containing wall, fence, pole, light, sign, terrain, rider, truck, bus, train, motorcycle and bicycle. Further more, we also explore the performance of our model when the images of the two source domains are fully labeled. In order to further evaluate the performance of all methods when combined with the pixel-level domain adaptation methods

Method	NT	T
Source	17.7	24.0
AdaptSegNet[41]	7.7	30.8
MinEnt[44]	27.1	30.1
Advent[44]	11.8	30.3
Ours w/o PSF	36.3	38.1
Ours (ADV)	40.1	41.5
Ours (PSF)	37.3	42.4
Ours (ADV+PSF)	**40.6**	**42.8**

(a)

MBR	UM	A^3	PSF	ADV	NT	T
					17.7	24.0
✓					20.9	21.4
✓	✓				27.6	36.8
✓	✓	✓*			29.1	37.0
✓	✓	✓			36.3	38.1
✓	✓			✓	35.4	40.9
✓	✓		✓		31.4	41.5
✓	✓	✓		✓	40.1	41.5
✓	✓	✓	✓		37.3	42.4
✓	✓	✓	✓	✓	**40.6**	**42.8**

(b)

Table 5: (a) Quantitative comparison of semantic segmentation under the setting, SYNTHIA+GTA5→ Cityscapes. "NT" means source images are not translated with CycleGAN. "T" means source images are translated by CycleGAN. The mIoU results are reported over 19 classes. The best results are denoted in bold. (b) Ablation study of our method when applied to single modal semantic segmentation. * represents there is no adversarial part in A^3 module, *i.e.,* , only DAT module. "ADV" represents the output space alignment between the source domain and target domain as done in [41], after generating the complete pseudo-label on the source domain. "ADV+PSF" means to combine the "ADV" and "PSF" by completing the label space and generating pseudo-label on both of the source and target domain, then adversarial alignment in the output space is also adopted during the second stage training.

[52, 17], we conduct the experiments under two settings; 1) source domain images are not translated with CycleGAN [52], named as "NT"; 2) source domain images are translated with CycleGAN, named as "T". Meanwhile, in order to verify the model performance combined with output-level adaptation method [41], we conduct additional experiments which include "ADV" in the fully-supervised adaptation stage. "ADV" generates the complete source domain label as done in PSF, and then trains the semantic segmentation model through the adversarial adaptation between pseudo-complete source domain and unlabeled target domain in the output-level space. For a fair comparison, all the methods adopt the DeepLabv2-ResNet101 [4, 15] semantic segmentation network.

4.4. 2D Semantic Image Segmentation: Results

Comparison with SOTA. In Table 5a, we show the quantitative comparison for semantic segmentation between our method and other SOTA methods. It is shown that our method without adding PSF highly outperforms the adaptation-based method AdaptSegNet[41], the self-supervision-based method MinEnt[44], and the method combining the adaptation and the self-supervision Advent [44]. Our method archives 36.3% and 38.1% in "NT" and "T" setting, respectively. Similar to the image classification results, without using the translated source images, the adaptation-based methods show the negative transfer effect and the performance is lower than the source-only baseline. By using the translated source images in "T", different source domain images are all Citysapes-like images. The different source domains can be seen as a larger unified source domain, which can provide the guidance of the complete label space to some extent by combination. So all the adaptation-based or self-supervision based methods perform much better in the "T" situation, compared with the non-adapted baseline. However, even in the "T" situation, our method can still provide the advantage by further completing label space, through our partial-supervised adaptation stage. It proves the effectiveness of our method for preventing the negative transfer and completing the label space. By further adding the second "fully-supervised" adaptation stage, the model achieves a new state-of-the-art performance in both the "T" and the "NT" settings. In Table 5b, we report the results of different ablations of our method. It confirms that different parts of our method are all useful for the final performance, and the output space alignment "ADV" is helpful to our method as well. In Fig. 4, we show the qualitative results of the 2D semantic segmentation on the target domain, under the "NT" setting.

Method	mIoU
Source	39.1
AdaptSegNet[41]	40.8
Minentropy[44]	42.2
Advent[44]	42.9
Ours w/o PSF	**43.1**

Table 6: Segmentation results for SYNTHIA + GTA5 → Cityscapes in the fully-labeled setting. The images in source domains are translated with CycleGAN.

Fully labeled. In Table 6, we show the quantitative comparison between our method and other SOTA methods under the fully labeled setting, *i.e.,* , the source domain images are labeled with all the considered classes, 16 classes in SYNTHIA and 19 classes in GTA5. Table 6 shows that our model still outperforms other unsupervised domain adaptive semantic segmentation methods under this setting as shown by the numbers: 43.1% v.s. 40.8%, 42.2%, and 42.9%. It shows that our method can also be used for the standard multi-source domain adaptation problem.

4.5. Cross-Modal Semantic Segmentation: Setup

In the cross-modal semantic segmentation setting, we aim to evaluate the performance of semantic segmentation model when different data modalities are used, such as 2D images and 3D LiDAR point clouds. For this setting, the 2D RGB images from the Cityscapes dtaset [7], and the 3D point clouds from the Nuscenes dataset [1] are treated as two different source domains. while the paired but unlabeled 2D RGB images and 3D point clouds from the A2D2 dataset [12] are used as the target domain. There are in total 10 classes that need to be transferred to the target domain. In the Cityscapes dataset, the label for 6 classes are

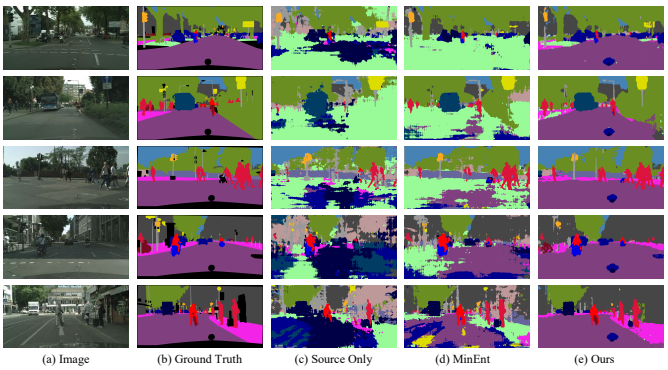


Figure 4: Qualitative comparison of semantic segmentation, SYNTHIA + GTA5→ Cityscapes, under the "NT" setting.

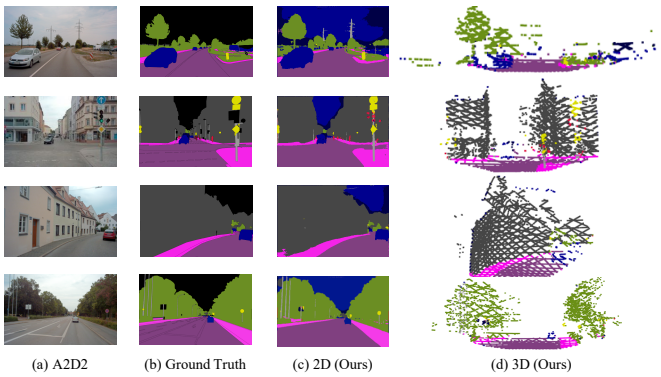


Figure 5: Qualitative results of the cross modal semantic segmentation on the target domain, A2D2.

Cityscapes + Nuscenes → A2D2	2D	3D	Fuse
Source	37.5	2.0	42.5
xMUDA[19]	16.3	1.7	9.1
ES + MinEnt[44]	22.3	1.5	20.8
ES + KL[19]	21.7	1.5	19.7
xMUDA + AKL	27.5	2.3	21.1
xMUDA + AKL + COMP	32.1	2.9	37.7
Ours w/o PSF	38.1	2.4	49.9
Ours	**54.9**	**37.1**	**55.7**

Table 7: Cross modal semantic segmentation performance under the setting, Nuscenes+Cityscapes→ A2D2. "Fuse" represents the average fusion of the prediction probability from 2D model and 3D model, then the final class prediction is chosen as the maximum of the fused probability. "ES" represents 2D and 3D average fusion ensemble. "KL" means KL-divergence alignment. "AKL" means adaptive KL-divergence alignment. "COMP" means complementary condition constraint for the point. The results are reported on mIoU over 10 classes on A2D2. The best results are denoted in bold.

given, covering road, sidewalk, building, pole, sign and nature. In the Nuscenes dataset, the label for 4 classes are given, including person, car, truck and bike. The 2D RGB images and 3D point clouds in the target domain are registered via a projection matrix between the 2D pixel and 3D points. By following [19], we adopt the U-Net-ResNet34 [37, 15] as the 2D semantic segmentation network, and the SparseConvNet [13] for 3D semantic segmentation. Due to the challenge of aligning features for the 3D point clouds, the A^3 module is not included in this cross-modal semantic segmentation setting.

4.6. Cross-Modal Semantic Segmentation: Results

Comparison with SOTA. In Table 7, we show the quantitative comparison between our method and the SOTA method for cross-modal unsupervised domain adaptation, xMUDA [19]. Similar to the image classification and the single modal semantic segmentation results, the adaptation-based method, xMUDA, shows obvious negative transfer effect, resulting in performance drop for the 2D model, 3D model and the fused one. Furthermore, we designed reasonable baseline methods for comparison: 1) ES + MinEnt: the prediction from 2D and 3D network is averaged in the target domain through the 2D and 3D point correspondence during training, and the fused prediction probability is optimized using the self-supervised minimum entropy loss [44]. 2) ES + KL: the KL-divergence [19] is utilized to align between the 2D/3D prediction and the fused prediction on the corresponding point in the target domain, respectively. 3) xMUDA + AKL: the KL-divergence alignment between 2D and 3D on the target domain is weighted adaptively, to reduce the wrong guidance from the unlabeled parts. 4) xMUDA + AKL + COMP: following baseline 3), another constraint, that the weight related to 2D and 3D needs to be complementary, is added. It is shown that our method prevents the negative transfer without the PSF component, outperforming the non-adapted baseline. Then by adding the PSF module, the 2D and 3D single-model performance is highly improved, achieving 54.9% and 37.1%, respectively. In Fig. 5, we show the qualitative results of the cross-modal semantic segmentation on the target domain. The good performance of our method in this setting proves the effectiveness of our method for the mDALU problem when learning with partial modalities. This opens the avenue to combine datasets collected with different sensors and offers the possibility of cheaply evaluating new combinations of sensors without annotating their data.

5. Conclusion

In this paper, we propose the multi-source domain adaptation and label unification with partial datasets problem, named as mDALU. Then we propose a novel multi-stage approach, including the partially and fully adaptation stage,to the mDALU problem. We further develop the

benchmarks on the image classification, the single modal semantic segmentation and the cross modal semantic segmentation, and demonstrate the effectiveness of our approach to mDALU problem through extensive experiments.

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 7, 13

[2] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Partial transfer learning with selective adversarial networks. In *CVPR*, 2018. 3

[3] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *ECCV*. 2018. 2, 3

[4] 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. 7

[5] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *CVPR*, 2019. 4

[6] 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. 1

[7] 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. 6, 7, 11, 13

[8] Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *JMLR*, 9(57):1757–1774, 2008. 2

[9] Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *IJCV*, 128(5):1182–1204, 2020. 2

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2, 5, 6, 15

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 11, 13

[12] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 7, 13

[13] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. *CVPR*, 2018. 8

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7, 8

[16] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *NeurIPS*, pages 8246–8256. 2018. 2

[17] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *ICML*, 2018. 7

[18] Jonathan J. Hull. A database for handwritten text recognition research. *TPAMI*, 16(5):550–554, 1994. 11, 13

[19] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *CVPR*, 2020. 2, 8, 11, 13

[20] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Perez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *CVPR*, 2020. 2

[21] Hu JIan, Hongya Tuo, Chao Wang, Lingfeng Qiao, Haowen Zhong, Yan Junchi, Zhongliang Jing, and Henry Leung. Discriminative partial domain adversarial network. In *ECCV*. 2020. 3

[22] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and C.V. Jawahar. Universal semi-supervised semantic segmentation. In *ICCV*, 2019. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 11

[24] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *CVPR*, 2020. 3

[25] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 5, 11

[26] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008. 14

[27] Jonathan Munro and Dima Damen. Multi-modal Domain Adaptation for Fine-grained Action Recognition. In *CVPR*, 2020. 2

[28] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS workshops*, 2011. 5, 11

[29] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *ICCV*, 2017. 2

[30] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 2, 5, 6, 11, 15

[31] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. In *ICLR*, 2020. 11

[32] Sayan Rakshit, Dipesh Tamboli, Pragati Shuddhodhan Meshram, Biplab Banerjee, Gemma Roig, and Subhasis Chaudhuri. Multi-source open-set deep adversarial domain adaptation. In *ECCV*, 2020. 2

[33] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *NeurIPS*, 2017. 2

[34] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *CVPR*, 2018. 2

[35] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: incremental classifier and representation learning. In *CVPR*, 2017. 3

[36] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 6, 11

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 8

[38] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 6, 11

[39] Amir Rosenfeld and John K Tsotsos. Incremental learning through deep adaptation. *TPAMI*, 42(03):651–663, 2020. 3

[40] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, 2018. 2

[41] 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. 4, 7, 15

[42] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[43] Ozan Unal, Luc Van Gool, and Dengxin Dai. Improving point cloud semantic segmentation by learning 3d object detection. In *WACV*, 2021. 3

[44] 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. 2, 7, 8, 15

[45] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. Towards universal object detection by domain attention. In *CVPR*, 2019. 2

[46] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 1

[47] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *CVPR*, 2018. 2, 5, 6, 15

[48] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *CVPR*, 2018. 3

[49] Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In *NeurIPS*. 2018. 2

[50] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *NeurIPS*, 2019. 2

[51] Xiangyun Zhao, Samuel Schulter, Gaurav Sharma, Yi-Hsuan Tsai, Manmohan Chandraker, and Ying Wu. Object detection with a unified label space from multiple datasets. In *ECCV*, 2020. 3, 5, 6, 15

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *ICCV*, 2017. 7, 14

Supplementary

In this supplementary, we provide additional information for,

S1 detailed framework structure and implementation of our approach,

S2 more detailed information about the datasets involved in experiments,

S3 experimental results when having more than two source domains,

S4 more experimental results and additional visualization results for semantic segmentation.

S1. Framework Structure and Implementation

In Sec. 3 and Fig. 2 of the main paper, we introduce our approach to mDALU problem, and here we provide more detailed structure and implementation of our approach. The overview of our approach is shown in Fig. S1. In the image classification experiment, the hyperparameter λ in Eq. (10) of the main paper is set as 1.0, and δ in Eq. (12) and Eq. (13) of the main paper is set as 0.5. The images are resized to 32×32. We use the the Adam optimizer [23] with $\beta_1 = 0.9, \beta_2 = 0.999$ and the weight decay as 5×10^{-4}. The learning rate is set as 2×10^{-4}. We adopt the same network architecture as that of the digits classification experiments in [30]. In the 2D semantic image segmentation experiments, the hyperparameter λ in Eq. (10) of the main paper is set as 0.001, and δ in Eq. (12) and Eq. (13) of the main paper is set as 0.2, 0.5 and 0.4 for SYNTHIA, GTA5 and Cityscapes dataset, respectively. The images are resized to 1024×512. We use the SGD optimizer for training the semantic segmentation network, whose momentum is 0.9, weight decay is 5×10^{-4} and learning rate is 2.5×10^{-4} with polynomial decay of power 0.9. Meanwhile, the Adam optimizer is used for training the discriminator network, whose momentum is $\beta_1 = 0.9, \beta_2 = 0.99$, weight decay is 5×10^{-4} and learning rate is 1×10^{-4} with polynomial decay of power 0.9. In the cross-modal semantic segmentation experiments, we follow the exactly same data augmentation and preprocess procedure as that of [19]. The hyperparameter δ in Eq. (12) and Eq. (13) of the main paper is set as 0.2. We use the Adam optimizer for training the 2D and 3D semantic segmentation network, with $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate is set as 1×10^{-3}.

S2. Datasets Overview of mDALU Benchmark

In Sec. 4 of the main paper, we introduce the benchmark setup of the mDALU problem. Here we provide more details about the datasets involved in the benchmark.

S2.1. Image Classification

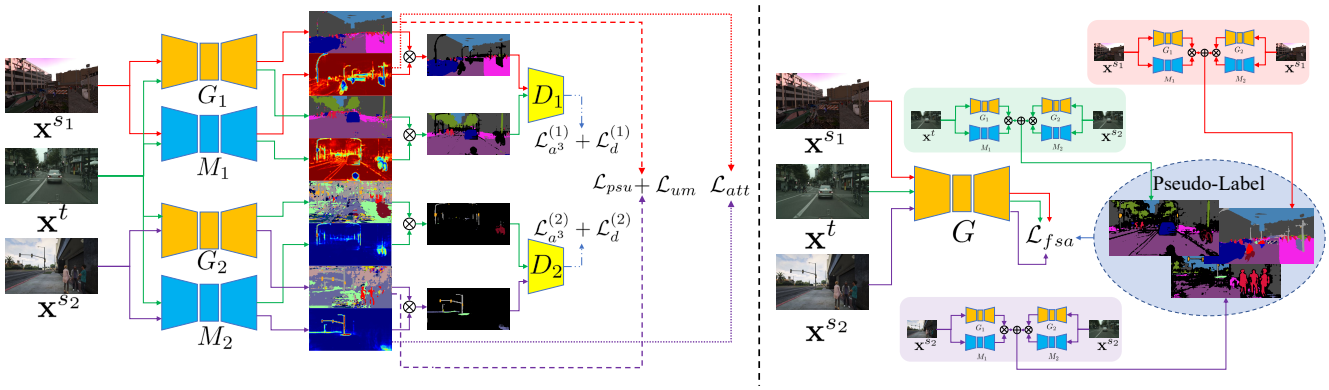
In the image classification benchmark of the main paper, we adopt three digits dataset, including MNIST [25], Synthetic Digits [11], and SVHN [28] dataset. The MNIST is the hand-written numbers image dataset, the SVHN is street view house numbers image dataset and the Synthetic Digits is synthetic numbers image dataset. In the image classification benchmark of the main paper, we adopt these three different style digits images, to introduce larger domain gap between different source domains to effectively evaluate the validity of different methods for mDALU problem. In Sec. S3, we introduce two more datasets, MNIST-M [11] and USPS [18] ,to evaluate the effectiveness of our approach when dealing with more than two source domains. The MNIST-M is synthetic numbers image dataset, and the USPS is the hand-written numbers image dataset. We follow the setup of splitting the dataset in [30, 31]. In each of MNIST, MNIST-M, SVHN and Synthetic Digits, 25000 images for training are sampled from the training subset, and 9000 images for testing are sampled from testing subset. And for the USPS dataset, due to there are only 9298 images in total are available, the whole training set covering 7438 images are used for training, while the whole testing set with 1860 images are adopted for testing. The MNIST, MNIST-M, SVHN, Synthetic Digits, USPS are abbreviated as MT, MM, SVHN, SYN, and UP, respectively. The detailed label space of different source domains and target domain under different experiments setup is listed in Table S1 and Table S2. The example images of different datasets are shown in Fig. S2.

S2.2. 2D Semantic Image Segmentation

In the 2D semantic image segmentation benchmark of the main paper, we adopt the synthetic image datasets, GTA5 [36] and SYNTHIA [38] and the real image dataset, Cityscapes [7]. We introduce the label space of different datasets in the main paper. Here we provide more additional information about the datasets.

Cityscapes. Cityscapes is a dataset composed of the street scene images collected from different European cities. We use the training set of Cityscapes covering 2993 images, without label information, as the target domain during training stage. And we adopt the validation set of Cityscapes, which are composed of 500 images and densely labeled with 19 classes, to evaluate the semantic segmentation performance of the model on the target domain.

GTA5. GTA5 is a synthetic urban scene image dataset, whose images are rendered from the game engine. The scene of the images is based on the city of Los Angeles. In our 2D semantic image segmentation benchmark, we use 24966 densely labeled images in the GTA5 dataset as one of our source domains, whose annotation is compatible with that of Cityscapes.



(a) Partially-Supervised Adaptation (b) Fully-Supervised Adaptation

Figure S1: Overview of our approach to mDALU problem. Our approach is composed of two stages: (a) partially-supervised adaptation stage, and (b) fully-supervised adaptation stage. In the partially-supervised adaptation stage, there are three modules involved, the domain attention (DAT) module, the uncertainty maximization (UM) module, and the attention-guided adversarial alignment (A^3) module. Besides the supervised semantic segmentation loss $\mathcal{L}_{psu}$ on the source domain, the DAT module is trained in the supervised way with $\mathcal{L}_{att}$, the UM module is trained in the supervised way with $\mathcal{L}_{um}$ and the A^3 module is trained in the adversarial way with $\mathcal{L}_{a^3} + \mathcal{L}_d$. In the fully-supervised adaptation stage, in order to complete the label space, the pseudo-label, for all the samples $\mathbf{x}^{s_1}$, $\mathbf{x}^{s_2}$, $\mathbf{x}^t$ from all related domains, is generated by fusing the probability map weighted by attention map from different branches, G_1, M_1 and G_2, M_2. Then the semantic segmentation network G is trained in the complete and unified label space with the generated pseudo-label and supervised loss $\mathcal{L}_{fsa}$. In the implementation, G_1, G_2, M_1, M_2 share the same encoder and adopt different label predictors.

Experiment	Label Space									
Non-Overlapping(Table 2 in main paper)	Domain	Source1	Source2	Target	Source1	Source2	Target	Source1	Source2	Target
	Dataset	SVHN	SYN	MT	MT	SVHN	SYN	MNIST	SYN	SVHN
	Class	0∼4	5∼9	0∼9	0∼4	5∼9	0∼9	0∼4	5∼9	0∼9
Partially-Overlapping(Table 4 in main paper)	Domain	Source1	Source2	Target	Source1	Source2	Target	Source1	Source2	Target
	Dataset	SVHN	SYN	MT	MT	SVHN	SYN	MNIST	SYN	SVHN
	Class	0∼6	3∼9	0∼9	0∼6	3∼9	0∼9	0∼6	3∼9	0∼9

Table S1: The label space of different source domains and target domain in the mDALU image classification benchmark of the main paper.

More Source Domains Experiments (Table S4 in supplementary)					
Domain	Source1	Source2	Source3	Source4	Target
Dataset	SVHN	SYN	MM	UP	MT
Class	0∼2	2∼4	4∼6	7∼9	0∼9
Dataset	MT	SYN	MM	UP	SVHN
Class	0∼2	2∼4	4∼6	7∼9	0∼9
Dataset	MT	SVHN	MM	UP	SYN
Class	0∼2	2∼4	4∼6	7∼9	0∼9
Dataset	MT	SVHN	SYN	UP	MM
Class	0∼2	2∼4	4∼6	7∼9	0∼9
Dataset	MT	SVHN	SYN	MM	UP
Class	0∼2	2∼4	4∼6	7∼9	0∼9

Table S2: The label space of different source domains and target domain in the mDALU image classification benchmark of the more source domains experiments in the supplementary.

(a) MT (b) SYN (c) SVHN (d) MM (d) UP

Figure S2: Example images of different datasets in mDALU image classification benchmark.

SYNTHIA. SYNTHIA is a synthetic dataset, containing photo-realistic images rendered from a virtual city. We use the SYNTHIA-RAND-Cityscapes subset, which contains 9400 densely labeled images and the 16 class annotation of which is compatible with that of Cityscapes. In our 2D semantic image segmentation benchmark, the labeled SYNTHIA dataset serves as one of our source domains.

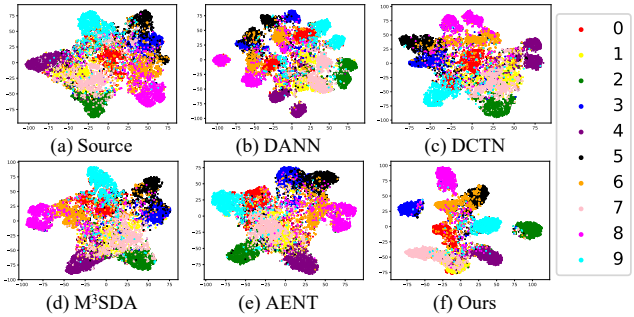
S2.3. Cross-Modal Semantic Segmentation

In the cross-modal semantic segmentation benchmark of the main paper, three datasets are involved, Cityscapes [7], Nuscenes [1] and A2D2 [12]. We introduce the label space of different datasets in the main paper. Here we provide more information on the datasets and the mapping between our label space and the annotated class label in different datasets.

Cityscapes. Cityscapes [7] is 2D urban scene image dataset, and has been introduced in the Sec. S2.2. In the cross-modal semantic segmentation benchmark, we adopt the training set of the Cityscapes, covering 2993 images, as the 2D source domain. Unlike the Sec. S2.2 does not use the label information for the training image, we use the ground truth label of the training images, but the label space of Cityscapes in our experiments only covers 6 classes, road, sidewalk, building, pole, sign and nature. The mapping from the original Cityscapes annotated classes and our label space is listed in Table S3.

Nuscenes. Nuscenes [1] is an autonomous driving dataset covering 1000 driving scenes, which are collected from the Boston and Singapore. Each scene, of 20-second length, is sampled and annotated at 2HZ, resulting in 40K well-annotated keyframes for 3D bounding boxes of the objects. In our cross-modal semantic segmentation benchmark, we adopt the training set of the Nuscenes, including 28130 keyframes 3D LiDAR points, as the 3D source domain. Then as done in [19], we generate the 3D point-wise semantic labels from the 3D bounding boxes, by assigning the object label to the points inside the bounding box and taking the points outside the bounding box as unlabeled points. The label space of the 3D source domain includes 4 classes, person, car, truck and bike. The mapping between the object label annotation in Nuscenes and our label space is reported in Table S3.

A2D2. A2D2 [12] is an autonomous driving dataset, including simultaneously recorded paired 2D images and 3D



(a) Source (b) DANN (c) DCTN

(d) M^{3}SDA (e) AENT (f) Ours

Figure S3: t-SNE Visualization of the feature embedding on the mDALU image classification benchmark, MT, SVHN, MM, UP $\rightarrow$ SYN. We adopt the same t-SNE parameters for all visualization.

LiDAR points. The A2D2 covers 20 scenes, which are corresponding to 28637 frames for training. And the scene 20180807_145028 is used for validation. The 2D images are densely labeled with 38 semantic classes. Following [19], the 3D point-wise semantic labels are generated by the reprojection to the 2D images. In our cross-modal semantic segmentation benchmark, the A2D2 serves as the target domain. We use the training set of A2D2 without label information during training, including the paired 2D images and 3D LiDAR points. And we use the validation set 20180807_145028 with ground truth label for evaluating the performance. The label space of the target domain for evaluation includes 10 classes, road, sidewalk, building, pole, sign, nature, person, car, truck and bike. The mapping between the label space and the annotated 38 semantic classses in A2D2 is shown in Table S3.

S3. Experiments with More Source Domains

In this section, we evaluate the effectiveness of our approach when dealing with more than two source domains. Based on the classification benchmark of the main paper, we here introduce two more source domains, MNIST-M [11] and USPS [18], which are abbreviated as "MM" and "UP" respectively. Then as done in main paper, each time, one of "MT", "SYN", "SVHN", "'MM' and "UP" is taken as the target domain, while the other four are used as the source domains. The label space of different source domains in the experiments is listed in Table S2.

Experimental results. In Table S4, we report the quantitative experimental results of the classification benchmark, after introducing two more source domains, MM and UP. It can be seen that our approach with the "partially-supervised adaptation" stage highly outperforms the source-only baseline, the adaptation-based methods DANN, DCTN, and M^{3}SDA, and the label-unification based method AENT. It achieves an average accuracy of 80.83% on the target do-

label space	A2D2	Cityscapes	Nuscenes
road	'rd normal street', 'zebra crossing', 'solid line', 'rd restricted area', 'slow drive area', 'drivable cobblestone', 'dashed line', 'painted driv. instr.'	'road'	–
sidewalk	'sidewalk', 'curbstone'	'sidewalk'	–
building	'buildings'	'building'	–
pole	'poles'	'pole'	–
sign	'traffic sign 1', 'traffic sign 2', 'traffic sign 3'	'traffic sign'	–
nature	'nature object'	'vegetation', 'terrain'	–
person	'pedestrian 1', 'pedestrian 2', 'pedestrian 3'	–	'pedestrian'
car	'car 1', 'car 2', 'car 3', 'car 4', 'ego car'	–	'car'
truck	'truck 1', 'truck 2', 'truck 3'	–	'truck'
bike	'bicycle 1', 'bicycle 2', 'bicycle 3', 'bicycle 4', 'small vehicles 1', 'small vehicles 2', 'small vehicles 3'	–	'motorcycle', 'bicycle'

Table S3: Class mapping between the label space and the annotated classes in different datasets.

main. Then by exploiting the "fully-supervised adaptation" stage, the performance is further improved to 82.88%. It proves the effectiveness and the robustness of our approach for addressing the mDALU problem when more than two source domains are given. In Fig. S3, the qualitative comparison of feature embedding, t-SNE visualization [26], between our approach and other methods is shown. It shows that our approach is able to learn more discriminative features than other methods. It further verifies the good performance of our approach to mDALU problem.

S4. More Experimental Results for Semantic Segmentation

Detailed experimental results for semantic segmentation. In Table 5a and Table 7 of the main paper, we show the quantitative comparison, through the mIoU, between our approach and other methods, on the 2D and cross-modal semantic segmentation benchmark. Correspondingly, we here provide more detailed experimental results in Table S5 and Table S6, covering the per-class IoU results.

Attention visualization for semantic segmentation. During the "partially-supervised adaptation" stage, we introduce the attention map in the domain attention (DAT) module, the attention-guided adversarial alignment (A^3) module and the inference via attention-guided fusion. In order to verify the effectiveness of our attention map prediction, we show the qualitative visualization of the attention map on the target domain images in Fig. S4. Corresponding to the Sec. 3.2.1 of the main paper, the attention map $\tilde{\mathbf{a}}_1^t$ and $\tilde{\mathbf{a}}_2^t$, are generated by feeding the target domain image $\mathbf{x}^t$ into the attention network M_1 and M_2. It is shown that our predicted attention map $\tilde{\mathbf{a}}_1^t$, corresponding to the source domain $\mathcal{S}_1$, has higher attention value, for the objects belonging to the partial label space $\mathcal{C}_1$, such as the road, sidewalk, building, vegetation, sky and car. And the predicted attention map $\tilde{\mathbf{a}}_2^t$, corresponding to the source domain $\mathcal{S}_2$, has higher attention value, for the objects belonging to the partial label space $\mathcal{C}_2$, such as the fence, pole, light, sign, bus, motorcycle and bicycle. It proves the validity of our attention map prediction.

Additional qualitative results for semantic segmentation. In Fig. 4 of the main paper, we show the qualitative comparison results between our approach and other methods on the 2D semantic image segmentation benchmark, and the source domain images are not translated with CycleGAN [52], *i.e.,* the "NT" setting. Here we provide additional qualitative comparison results between our approach and other methods on the 2D semantic image segmentation benchmark, and the source images are translated with CycleGAN [52], *i.e.,* the "T" setting. As shown in Fig. S5, it can be seen that our approach obviously outperforms other methods on the 2D semantic image segmentation benchmark. It further verifies the effectiveness of our approach to mDALU problem.

Method	MT	SYN	SVHN	MM	UP	Avg
Source	86.90 ± 0.40	63.80 ± 0.15	51.84 ± 2.13	52.09 ± 0.69	91.83 ± 0.78	69.29 ± 0.83
DANN[10]	86.38 ± 1.44	63.76 ± 0.88	51.58 ± 2.27	52.14 ± 0.61	89.98 ± 1.42	68.77 ± 1.32
DCTN [47]	63.87 ± 0.10	53.33 ± 1.15	43.57 ± 0.98	40.23 ± 0.48	59.78 ± 1.19	52.16 ± 0.78
M^{3}SDA [30]	87.26 ± 1.54	63.40 ± 0.32	48.96 ± 0.92	52.28 ± 1.60	90.20 ± 0.97	68.42 ± 1.07
AENT[51]	79.55 ± 2.40	63.22 ± 0.41	52.58 ± 2.27	48.65 ± 0.31	87.62 ± 1.36	66.32 ± 1.35
Ours w/o PSF	**94.90±0.23**	**78.37±0.58**	**72.18±0.44**	**63.01±0.74**	**95.70±0.44**	**80.83±0.49**
Ours	**96.60±0.07**	**80.68±0.30**	**73.82±0.35**	**66.62±0.62**	**96.70 ± 0.22**	**82.88±0.31**

Table S4: Quantitative comparison between our method and other SOTA methods, under mDALU image classification benchmark with 4 source domains. "MT", "SYN", "SVHN", "MM", and "UP" represent the target domain. We realize AENT on classification by utilizing the ambiguity cross entropy loss proposed in [51]. The best results are denoted in bold.

		GTA5+SYNTHIA→Cityscapes																			
Setting	Method	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrian	sky	person	rider	car	truck	bus	train	motorbike	bicycle	mIoU
NT	Source	3.0	10.1	42.1	7.3	6.6	10.6	18.2	**31.2**	61.3	3.9	73.2	27.5	16.2	9.9	1.4	1.6	0.0	8.6	3.8	17.7
	AdaptSegNet[41]	0.1	0.0	1.4	4.0	6.6	5.4	14.6	22.8	5.9	1.9	35.9	1.3	18.0	0.6	3.0	1.8	0.7	13.0	9.4	7.7
	MinEnt[44]	32.0	10.0	73.0	15.4	18.1	20.5	29.5	19.9	75.3	3.9	**79.6**	51.3	18.7	18.5	4.3	4.8	**9.2**	**20.3**	10.3	27.1
	Advent[44]	6.3	1.0	27.7	4.5	6.3	6.5	16.9	19.3	16.7	2.0	40.6	6.8	17.1	7.7	3.7	6.6	1.2	15.0	18.5	11.8
	Ours (w/o PSF)	**82.8**	30.8	78.9	17.5	15.8	28.0	34.8	18.9	79.1	10.5	78.4	52.0	18.2	71.4	16.8	34.3	2.0	11.0	8.0	36.3
	Ours (ADV)	82.1	**35.2**	78.1	**27.3**	18.8	**29.6**	33.0	21.1	78.3	**36.9**	75.3	**58.9**	**25.0**	69.6	19.3	33.8	0.0	15.6	**22.9**	40.1
	Ours (PSF)	77.8	31.9	**79.5**	17.9	18.1	29.0	**34.9**	20.9	80.2	9.0	**79.6**	55.6	20.9	**74.4**	16.9	25.5	0.0	17.0	18.9	37.3
	Ours (ADV+PSF)	81.7	34.1	**79.5**	26.7	**19.4**	29.0	32.0	23.2	**82.3**	31.4	79.5	57.5	22.3	66.6	**26.8**	**40.2**	0.0	19.4	20.4	**40.6**
T	Source	28.7	9.5	52.3	11.1	10.0	9.5	16.4	**30.6**	55.9	2.7	67.5	40.8	21.1	38.7	6.9	4.3	**6.4**	**22.1**	20.6	24.0
	AdaptSegNet[41]	78.3	34.5	75.7	16.2	15.6	11.5	19.0	10.8	78.0	16.5	76.3	42.6	8.4	59.6	10.9	8.8	0.5	14.2	8.7	30.8
	MinEnt[44]	58.5	20.6	70.5	12.0	17.9	18.3	19.9	27.1	74.3	8.0	79.1	46.5	20.5	37.7	9.1	20.4	2.8	18.9	10.6	30.1
	Advent[44]	78.0	34.3	75.9	14.5	5.8	9.8	17.2	10.2	76.4	15.0	76.9	40.6	3.1	61.3	19.3	14.5	0.0	9.9	12.5	30.3
	Ours(w/o PSF)	86.0	40.8	79.1	13.2	22.7	33.5	33.3	18.9	79.9	33.2	72.0	49.7	19.1	63.3	20.6	10.1	0.0	13.4	34.0	38.1
	Ours (ADV)	86.2	41.3	81.6	21.1	**23.3**	33.4	32.0	20.6	81.0	32.1	**79.8**	57.5	**26.4**	70.5	**24.8**	**31.4**	0.2	18.3	27.1	41.5
	Ours (PSF)	**87.8**	**42.9**	81.2	17.3	22.0	34.1	**36.9**	17.9	82.2	34.2	73.6	**58.9**	25.1	**76.5**	24.4	28.9	0.1	19.8	**41.9**	42.4
	Ours (PSF+ADV)	86.8	42.5	**82.5**	**23.0**	23.1	**34.4**	36.3	29.1	**82.9**	**34.3**	76.5	56.5	24.1	75.5	23.6	17.3	0.3	22.0	41.6	**42.8**

Table S5: Per-Class IoU on the mDALU 2D semantic image segmentation benchmark. "NT" means source domain images are not translated with CycleGAN, and "T" means source domain images are translated with CycleGAN. The mIoU results are reported over 19 classes. The best results are denoted in bold.

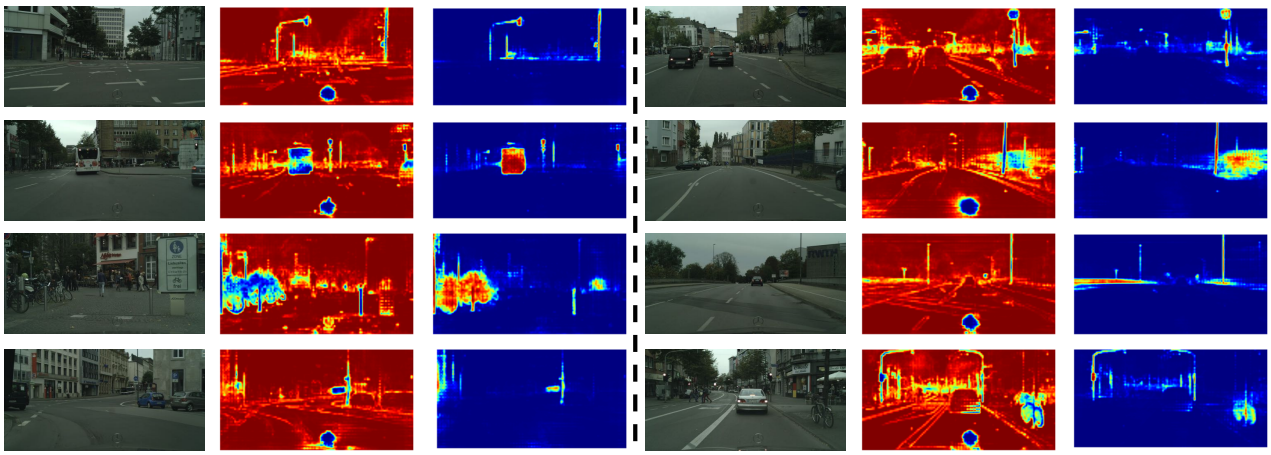
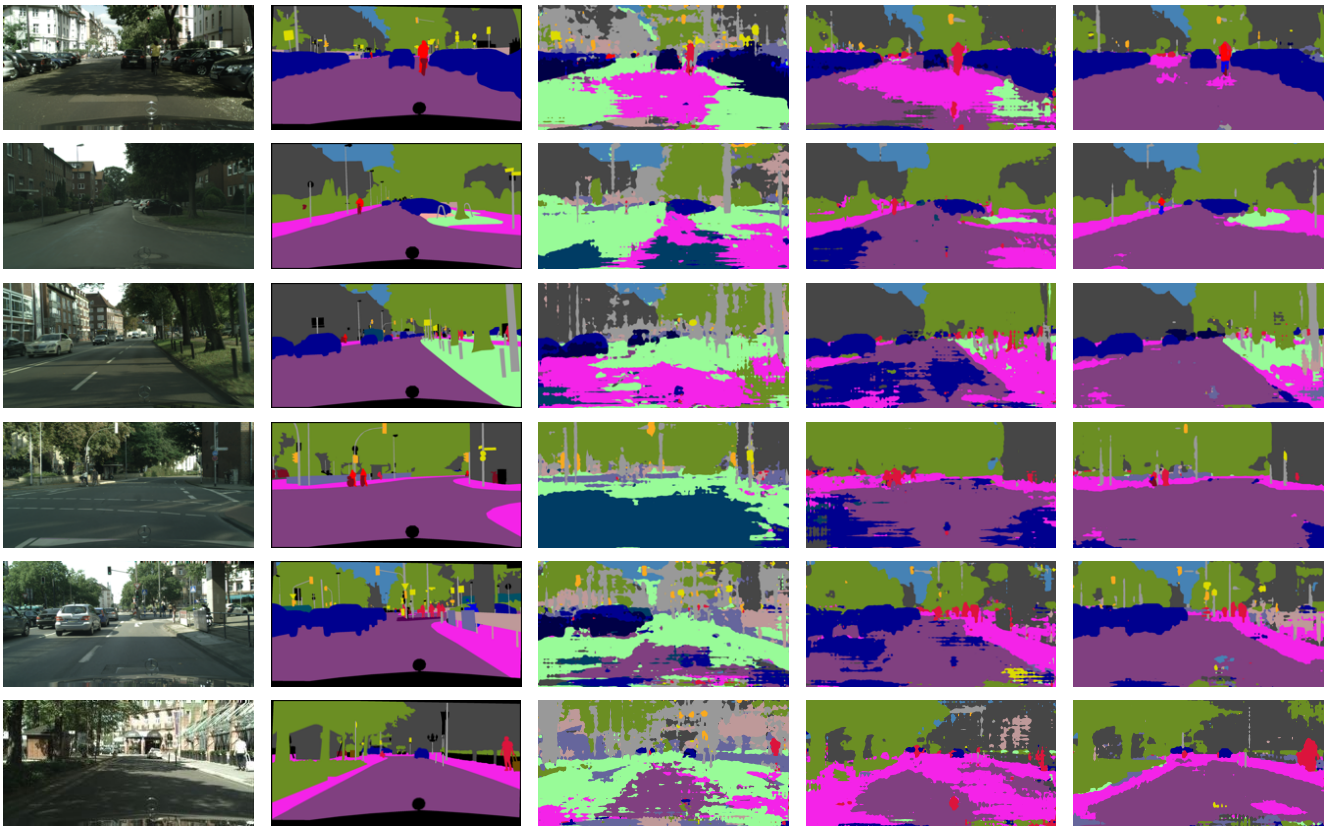


Figure S4: Visualization of the attention map $\tilde{\mathbf{a}}_1^t$ and $\tilde{\mathbf{a}}_2^t$ of the target domain images. (a) is the Cityscapes image $\mathbf{x}^t$. (b) is the attention map $\tilde{\mathbf{a}}_1^t$, generated by feeding the $\mathbf{x}^t$ into the attention network M_1. (c) is the attention map $\tilde{\mathbf{a}}_2^t$, generated by feeding the $\mathbf{x}^t$ into the attention network M_2. Red parts are the parts with higher attention value, while the blue parts with lower attention value.

Cityscapes+Nuscenes→A2D2												
Modality	Method	road	sidewalk	building	pole	sign	nature	person	car	truck	bike	mIoU
2D	Sources	83.1	48.7	85.0	**34.8**	36.1	87.0	0.0	0.0	0.0	0.0	37.5
	xMUDA	68.2	13.8	22.3	22.1	15.7	3.6	0.1	15.9	1.6	0.0	16.3
	ES + MinEnt	55.7	15.6	64.9	19.8	21.7	45.2	0.0	0.0	0.0	0.0	22.3
	ES + KL	14.4	20.0	74.2	15.3	36.6	46.2	0.0	9.3	1.5	0.0	21.7
	xMUDA + AKL	44.8	29.7	46.5	36.2	33.6	61.4	0.0	21.0	1.8	0.0	27.5
	xMUDA + AKL + COMP	70.3	38.1	76.4	25.0	30.5	80.8	0.0	0.0	0.0	0.0	32.1
	Ours (w/o PSF)	85.8	54.3	81.8	34.1	40.8	81.4	0.0	0.0	2.8	0.0	38.1
	Ours	**92.8**	**59.9**	**90.0**	30.4	**60.7**	**90.6**	**13.8**	**71.6**	**39.1**	**0.4**	**54.9**
3D	Source	0.0	0.0	0.0	0.0	0.0	0.0	2.1	16.1	1.4	0.0	2.0
	xMUDA	0.0	0.0	0.0	0.0	0.0	0.0	1.2	14.6	1.5	0.0	1.7
	ES + MinEnt	0.0	0.0	0.0	0.0	0.0	0.0	1.9	12.0	1.5	0.0	1.5
	ES + KL	0.0	0.0	0.0	0.0	0.0	0.0	1.9	9.8	1.8	1.2	1.5
	xMUDA + AKL	0.0	0.0	0.0	0.0	0.0	0.0	2.4	18.5	1.3	0.4	2.3
	xMUDA + AKL + COMP	6.1	1.8	0.0	0.0	0.0	0.0	2.1	17.9	1.3	0.0	2.9
	Ours (w/o PSF)	0.6	0.7	0.3	0.0	0.0	2.4	1.3	16.1	2.3	0.0	2.4
	Ours	**82.0**	**27.7**	**80.3**	**1.4**	**7.5**	**80.8**	**7.2**	**54.9**	**25.6**	**3.5**	**37.1**
Fuse	Source	85.5	51.8	83.8	**41.8**	40.2	83.8	6.3	23.0	8.8	0.0	42.5
	xMUDA	55.8	2.2	2.8	3.5	2.8	0.2	2.7	19.4	1.7	0.0	9.1
	ES + MinEnt	63.1	7.5	69.7	9.0	13.8	30.2	2.6	11.0	1.2	0.0	20.8
	ES + KL	10.6	21.2	65.0	18.2	26.8	34.7	5.4	12.0	2.4	0.3	19.7
	xMUDA + AKL	13.2	36.9	20.1	34.1	31.1	44.5	4.6	24.8	1.7	0.1	21.1
	xMUDA + AKL + COMP	74.1	43.5	74.4	35.2	35.5	71.0	4.1	34.7	5.0	0.0	37.7
	Ours(w/o PSF)	91.1	57.3	85.7	39.7	47.4	85.9	8.6	57.8	25.3	0.4	49.9
	Ours	**91.7**	**58.6**	**90.1**	34.5	**58.8**	**90.3**	**15.4**	**72.4**	**43.6**	**1.3**	**55.7**

Table S6: Per-Class IoU on the mDALU cross-modal semantic segmentation benchmark. The mIoU results are reported over 10 classes. The best results are denoted in bold.



(a) Image (b) Ground Truth (c) Source Only (d) AdaptSegNet (e) Ours

Figure S5: Qualitative comparison of semantic segmentation results, under the mDALU 2D semantic image segmentation benchmark, SYNTHIA + GTA5 $\rightarrow$ Cityscapes. The source images are translated with CycleGAN, *i.e.,* setting "T".