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

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

Rui Gong [1], Dengxin Dai [1,4], Yuhua Chen [1], Wen Li [3], Luc Van Gool [1,2]
[1] Computer Vision Lab, ETH Zurich, [2] VISICS, KU Leuven, [3] UESTC, [4] MPI for Informatics
{gongr, dai, yuhua.chen, vangool}@vision.ee.ethz.ch, liwenbnu@gmail.com

Abstract

One challenge of object recognition is to generalize to new domains, to more classes and/or to new modalities. This necessitates methods to combine and reuse existing datasets that may belong to different domains, have partial annotations, and/or have different data modalities. This paper formulates this as a multi-source domain adaptation and label unification 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 unmatching label spaces 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. Extensive experiments on three different tasks - image classification, 2D semantic image segmentation, and joint 2D-3D semantic segmentation - show that our method outperforms all competing methods significantly.

1. Introduction

The development of object recognition is carried by two pillars: 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 new tasks, the creation of new datasets can hardly keep up due to its huge cost. In the literature, a diverse set of learning paradigms, such as self-learning [13], semi-supervised learning [17] and transfer learning [6], have been developed to come to the 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 time goes, research goals will become more and more ambitious, so object recognition models for more classes, new domains, and/or more data modalities are necessary.

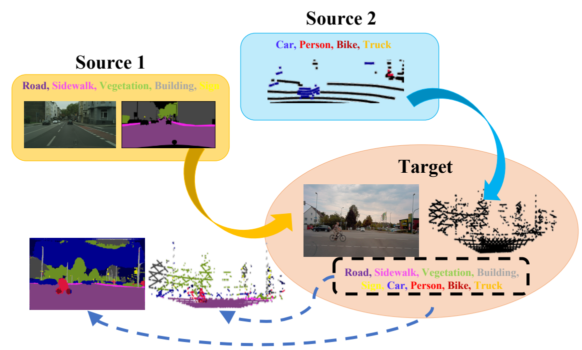


Figure 1: mDALU learns a complete-class and complete-modality object recognition model for a new, unlabeled target domain, by using multiple datasets with partial-class annotation and partial data modality as source domains.

To address this, we propose a multi-source domain adaptation and label unification (mDALU) problem. In this setting, there are multiple source domains and an unlabeled target domain. In each source domain, only samples (images, pixels, or LiDAR points) 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, and can have inconsistent taxonomies, *e.g.*, truck is labeled as "truck" in one source domain but labeled as "vehicle" together with other types of vehicles in another. Further, the data modalities in different source domains can also be different, *e.g.*, one contains images and the other contains LiDAR point clouds. The goal is to obtain an object recognition model for all classes in the target domain. Fig. 1 shows an exemplar setting of mDALU. A comparison to other domain adaptation settings, in Table 1, shows that mDALU is very flexible.

This goal is challenging. Firstly, there is the notorious issue of negative transfer. While negative transfer is

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?	Can Handle Inconsistent Taxonomy?
Unsupervised Domain Adaptation [10]	No	No	–	Same Size	No	–
Partial Domain Adaptation [3]	No	No	–	Reduced	No	–
Multi-Source Domain Adaptation [26, 43]	Yes	No	No	Same Size	No	No
Category-Shift Multi-Source Domain Adaptation [39]	Yes	No	Yes	Increased	No	No
Multi-Modal Domain Adaptation [18]	Yes	Yes	No	Same Size	No	No
Multi-Source Open-Set Domain Adaptation [27, 25]	Yes	No	No	Same Size + 1*	Yes	No
Multi-Source Domain Adaptation and Label Unification (mDALU)	**Yes**	**Yes**	**Yes**	**Increased**	**Yes**	**Yes**

Table 1: Comparison between our mDALU and other domain adaptation settings (see Sec. 2 for details). It is clear that mDALU offers a very flexible and general setting. * "1" means an additional "unknown" class in the target domain.

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 avoid making confident predictions for unlabeled samples in the source domains, and to enable robust 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 fuse the supervision of all partial datasets to transfer the supervision 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 samples in the source domains and all samples in the target domain. Standard supervised learning is then performed in the unified label space for the final model.

To showcase the effectiveness of our method, we evaluate it on three different tasks: image classification, 2D semantic image segmentation, and joint 2D-3D semantic segmentation. Synthetic and real data, and images and LiDAR point clouds are involved. Also, non-overlapping, partially-overlapping and fully-overlapping label spaces, and consistent and inconsistent taxonomies across source domains are covered. Experiments show that our method outperforms all competing methods significantly.

2. Related Work

Multi-Source Domain Adaptation. Transfer learning and domain adaptation have been extensively studied in the past years. Several effective strategies have been developed such as minimizing maximum mean discrepancy [36, 23], moment matching [40], adversarial domain confusion [10, 35], entropy regularization [37], 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 methods are proposed [8, 42, 26, 15, 43]. Yet, these methods all assume the same label space for all domains. Xu *et al.* [39] explores the problem of the category shift among different source domains, and adopts the k-way domain discriminator to reduce the effect of category shift. But the method is mainly proposed for the image classification task, and cannot deal with the problem of partial annotation, inconsistent taxonomies and modal differences among different source domains.

Open-Set/Partial Domain Adaptation. Recent research explores the category "openness" between the source domain and the target domain, which is divided into open-set domain adaptation and partial domain adaptation. Open-set domain adaptation [25, 33, 27] assumes that the target domain includes new classes that are unseen in the source domain, and aims to classify the unseen class samples as "unknown" class in the target domain. Partial domain adaptation [2, 41, 3, 19] aims to transfer knowledge from existing large-scale domains (e.g. 1K classes) to unknown small-scale domains (e.g. 20 classes) for customized applications. Different than both open-set and partial domain adaptation, our label space of the target domain is the union of label spaces of all source domains.

Learning from multiple datasets. Several successful methods [28, 29, 38, 20] have been proposed to learn a single universal network, that can represent different domains with a minimum of domain-specific parameters. But those methods do not consider domain adaptation and label space unification. Recently, Lambert *et al.* [21] presented a composite dataset that unifies different semantic segmentation datasets by reconciling the taxonomies, merging and splitting classes manually. But they do not address the problem of domain adaptation, partial annotation and cross-modal data, and they rely on the manual re-annotation for unification. The object detection method by Zhao *et al.* [44] performs label space unification from multiple datasets with partial annotations, but it does not consider other problems that are considered by our method such as domain discrepancies, inconsistent taxonomies and mismatched data modalities across the datasets.

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, resp. All the source domains can contain both partially labeled and unlabeled samples. The unlabeled samples can belong to the labeled classes of other domains. The label spaces $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K$ can be non-, partially-, or fully-overlapping with each other. Moreover, both consistent and inconsistent taxonomies among $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K$ are allowed. Then the union of the label spaces $\mathcal{C}_i, i = 1, ..., K$ forms 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 unlabeled 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 the partially-supervised adaptation stage, the partial supervision is transferred to the target domain from different source domains, respectively. Then in the fully-supervised adaptation stage, the supervision, in complete label space, is fused and self-completed on the unlabeled samples, and jointly transferred in the source domains and target domain. In order to realize adaptation under partial supervision, we propose three modules: DAT, UM and A^3 for the first stage. Then in the second stage, we use PSF and further learning. Below we provide details of all these components. From Sec. 3.2.1 to Sec. 3.2.5, we first introduce our method for mDALU under consistent taxonomies. In this part, we first describe a basic version of our method composed of DAT and inference via attention-guided fusion, which will be followed by UM and A^3 to enhance the adaptation ability. Finally, we present PSF. Then in Sec. 3.2.6, we extend our proposed method towards mDALU under inconsistent taxonomies.

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 cover partial label spaces $\mathcal{C}_i$, 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, coined multi-branch (MBR) fusion, this generates 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 that dominate the averages. We thus propose the domain attention (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 indices 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: $M_i = \{E_i, B_i^M\}$. M_i is trained under an MSE loss $\mathcal{L}_{\text{att}}$, together with G_i in a multi-task setting.

3.2.2 Inference via Attention-Guided Fusion

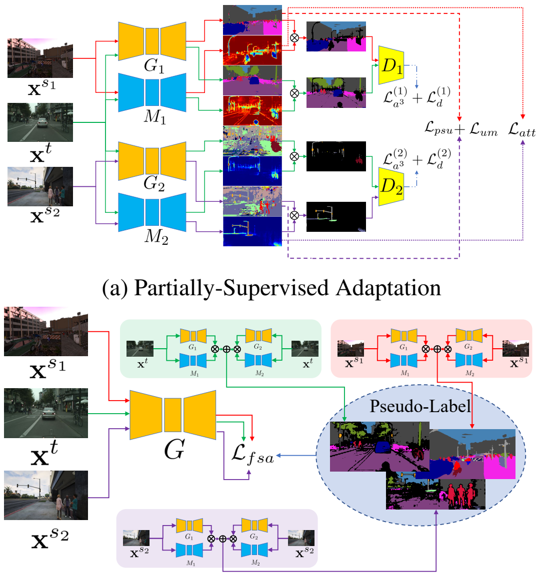
We feed an image $\mathbf{x}$ into semantic segmentation networks G_i to generate the corresponding probability maps $\hat{\mathbf{p}}_i \in [0, 1]^{H \times W \times C_\cup}$, and into different attention networks M_i to generate attention maps $\hat{\mathbf{a}}_i$. Then we fuse the predictions by averaging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

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

where $(\sum_{i=1}^K \hat{\mathbf{a}}_i \otimes \hat{\mathbf{p}}_i)^{(j)}$ yields the probability of the j^{th} class. The predicted class is then obtained via argmax.

3.2.3 Uncertainty Maximization (UM)

Due to the lack of ground truth class supervision, while we have the attention-guided fusion, the wrong prediction of unlabeled samples in the source domains can still have negative effects for our cross-domain prediction fusion. In order to further reduce the negative effects of unlabeled samples $\mathbf{x}_u^{s_i}$ in source domains, we propose a module specifically to maximize uncertainties of the predictions on unlabeled samples in those domains. 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 $G_i(\mathbf{x}_u^{s_i})^{(j)}$ represents the probability of the j^{th} class. In order to maximize the



(a) Partially-Supervised Adaptation

(b) Fully-Supervised Adaptation

Figure 2: Illustration of our approach to mDALU. There are 2 stages, (a) partially supervised adaptation and (b) fully-supervised adaptation.

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 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 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. $\mathcal{L}_{um}$ can be written as

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

Through the UM module, we encourage the model to make uniform categorical probability predictions, $\frac{1}{C_\cup}$, for unlabeled samples over the unlabeled classes, to best preserve the uncertainty to let the ground truth supervision of those classes from other source domains make the decision in the further attention-guided fusion and PSF process.

3.2.4 Attention-Guided Adversarial Alignment (A^3)

It has been proven in the literature that adversarial alignment is effective for domain adaptation. We extend the idea to mDALU. For adversarial alignment, one discriminator D_i is used for each source domain, to align the distribution between the source domain $\mathcal{S}_i$ and the target domain $\mathcal{T}$. In general unsupervised domain adaptation, the discriminator training loss $\mathcal{L}_d$ and the adversarial loss $\mathcal{L}_{adv}$ [34] for the source domain $\mathcal{S}_i$ and the target domain $\mathcal{T}$ is defined as

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

$$\begin{aligned}\mathcal{L}_d^{(i)}(\mathbf{x}_i^{s_i}, \mathbf{x}^t) = &-\log(D_i(G_i(\mathbf{x}^{s_i}))) \\ &-\log(1 - D_i(G_i(\mathbf{x}^t))).\end{aligned} \tag{7}$$

Yet, in our mDALU problem, there is no ground truth label guidance available for the unlabeled classes. A direct alignment between the source domain and the target domain will cause negative transfer, *i.e.,* the transfer of incorrect knowledge from the unlabeled parts in the source domains to the target domain. Here, we again use our attention map $\mathbf{a}^{s_i}$ to alleviate this problem by proposing an attention-guided adversarial loss:

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

$$\begin{aligned}\mathcal{L}_d^{(i)}(\mathbf{x}_i^{s_i}, \mathbf{x}^t) = &-\log(D_i(G_i(\mathbf{x}^{s_i}) \otimes M_i(\mathbf{x}^{s_i}))) \\ &-\log(1 - D_i(G_i(\mathbf{x}^t) \otimes M_i(\mathbf{x}^t))),\end{aligned} \tag{9}$$

where $\otimes$ represents 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 out the attention-guided adversarial loss against 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 from different source domains to the target domain. In the second fully-supervised adaptation stage, we aim at learning and transferring knowledge in the complete and unified label space $\mathcal{C}_\cup$ between all domains jointly. In order to realize that, we complete the label spaces 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 spaces $\mathcal{C}_i$ 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 in the source domain $\mathcal{S}_i$, we feed each of the source image samples $\mathbf{x}^{s_i}$ into every semantic model $G_k, k = 1, ..., K$, to generate 'partial' semantic probability maps $\hat{\mathbf{p}}_k^{s_i} \in [0, 1]^{H \times W \times C_\cup}$ 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 the source domain $\mathcal{S}_i$ 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 labels are available, we obtain pseudo labels 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.

3.2.6 Inconsistent Taxonomies

The above method is able to deal with the mDALU problem under consistent taxonomies, *i.e.,* the different classes in all source domains are exclusive with each other. Yet, there might be inconsistent taxonomies between different source domains, causing a performance drop for the inconsistent taxonomies classes. Here, we introduce the extension of our above method, to handle the inconsistent taxonomies problem. Denoting the classes in the label spaces $\mathcal{C}_i$ as $\mathbf{c}_i^o$, we have $\mathcal{C}_i = \{\mathbf{c}_i^o, o = 1, 2, ..., C_i\}$. Then the inconsistent taxonomies among different source domains can be defined as, $\exists \mathbf{c}_p^q \in \mathcal{C}_p, \mathbf{c}_m^n \in \mathcal{C}_m, p, m = 1, ..., K, p \neq m, q = 1, ..., C_p, n = 1, ..., C_m$, we have $\mathbf{c}_p^q \neq \mathbf{c}_m^n$, and $\mathbf{c}_p^q \cap \mathbf{c}_m^n \neq \emptyset$. The inconsistent taxonomies classes between different source domains $\mathcal{S}_p$ and $\mathcal{S}_m$ are denoted as $\mathbf{c}_p^q \in \mathcal{C}_p$ and $\mathbf{c}_m^n \in \mathcal{C}_m$. For example, the truck is labeled as "truck" class $\mathbf{c}_p^q$ in one dataset $\mathcal{S}_p$, while it is labeled as "vehicle" class $\mathbf{c}_m^n$ together with other vehicles in another dataset $\mathcal{S}_m$. Another typical example is motorcycles being labeled as "cycle" class $\mathbf{c}_p^q$ together with other cycles in one dataset $\mathcal{S}_p$, but being labeled as "vehicle" class $\mathbf{c}_m^n$ together with other vehicles in another dataset $\mathcal{S}_m$. In the unified label space of the target domain, the conflict part $\mathbf{c}_p^q \cap \mathbf{c}_m^n$ is assigned to either $\mathbf{c}_p^q$ or $\mathbf{c}_m^n$ exclusively. Without loss of generality and for reasons of clarity, it is assumed that the $\mathbf{c}_p^q \cap \mathbf{c}_m^n$ is assigned to $\mathbf{c}_p^q$. Then in order to solve the conflict of $\mathbf{c}_p^q$ and $\mathbf{c}_m^n$, in the attention-guided fusion, we introduce the additional class-wise weight map $\mathbf{w}_i \in \mathbb{R}^{H\times W\times C_\cup}$, and Eq. (2) is extended to Eq. (16),

$$\mathbf{w}_i(h,w,j) = \begin{cases} = v, \text{if } \arg\max \hat{\mathbf{p}}_i(h,w) = q', \text{and } i = p, \\ \ \text{and } \arg\max \hat{\mathbf{p}}_m(h,w) = n', \text{and } j = q' \\ = 1, \text{otherwise} \end{cases} \tag{15}$$

$$\mathbf{f} = \frac{\sum_{i=1}^{K} \hat{\mathbf{a}}_i \otimes \hat{\mathbf{p}}_i \otimes \mathbf{w}_i}{\sum_{j=1}^{C_\cup} (\sum_{i=1}^{K} \hat{\mathbf{a}}_i \otimes \hat{\mathbf{p}}_i \otimes \mathbf{w}_i)^{(j)}}, \tag{16}$$

where $v > 1$ in Eq. (15) is a hyper-parameter, set to 5.0. v is used to increase the weight of class $\mathbf{c}_p^q$ of the corresponding prediction $\hat{\mathbf{p}}_p$ in Eq. (16), to convert $\mathbf{c}_p^q \cap \mathbf{c}_m^n$ to $\mathbf{c}_p^q$ in the prediction fusion. q', n' are the class indices of $\mathbf{c}_p^q$ and $\mathbf{c}_m^n$ in the unified label space $\mathcal{C}_\cup$. Correspondingly, under inconsistent taxonomies, besides the unlabeled samples in the source domains being completed with the predicted pseudo-label as in Eq. (12), the conflict part $\mathbf{c}_p^q \cap \mathbf{c}_m^n$, which is labeled as $\mathbf{c}_m^n$ originally in $\mathcal{S}_m$, is relabeled with the predicted pseudo-label $\bar{\mathbf{y}}^{s_i}(h,w)$, *i.e.,*

$$\hat{\mathbf{y}}^{s_m}(h,w) = q', \text{if } \mathbf{f}^{s_m}(h,w,q) > \delta \\ \text{and } \bar{\mathbf{y}}^{s_m}(h,w) = q' \text{ and } \mathbf{y}^{s_m}(h,w) = n'. \tag{17}$$

4. Experiments

We evaluate the effectiveness of our method mDALU under different settings. We build benchmarks for image classification, 2D semantic image segmentation, and 2D-3D cross-modal semantic segmentation.

4.1. Image Classification

Setup. In the classification benchmark, we adopt the digits classification images from three different datasets, MNIST [22], Synthetic Digits [10], and SVHN [24], coined "MT", "SYN" and "SVHN", resp. Each time, one of them is taken as the target domain, the other two as source domains. There are 10 classes, from '0' to '9', in the target domain. In our main setting, we adopt the most difficult setup to evaluate different methods, where the label spaces of different source domains are non-overlapping. Only half the classes are labeled in each of the source domains. The partially-overlapping situation is also explored. For fair comparison, we adopt the same network architecture used in [26] for all methods. The classification performance is evaluated on all 10 classes in the target domain.

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 [39]	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 [26]	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[44]	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 [39]	73.40±0.85	65.63±0.43	52.12±0.07	63.72 ± 0.45
AENT[44] 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 of image classification. "MT", "SYN", and "SVHN" represent the target domain. "PL" represents to add the pseudo-label training module, which is specifically adjusted according to their own paper's design. * represents to remove the unlabeled samples in the training data. We implement AENT for classification by utilizing the ambiguity cross entropy loss proposed in [44].

Comparison with SOTA. Table 2 compares our method with other SOTA methods which include 1) unsupervised domain adaptation method DANN [10], 2) category-shift unsupervised domain adaptation method DCTN [39], 3) multi-source unsupervised domain adaptation method M^{3}SDA [26], and 4) label unification method AENT [44]. 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 perform similarly to the baseline trained with source data only. This is because each source domain can only provide 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 renders the prediction on the target domain contradictory and the model hard to adapt to the complete label space. In contrast, the label-unification based method AENT obtained a performance gain of 4.25%, from 60.65% to 64.90%, compared with the 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, 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 strongly outperforms DCTN [39] and AENT [44], both with pseudo-label training, by 15.28% and 9.65%, resp. This proves the effectiveness of our entire method for domain adaptation, label space completion and supervision fusion. The ablation results in Table 3 show that each part of our model contributes to its performance.

Partially Overlapping. In Fig. 3, it is shown that the testing accuracy on the target domain increases, as more and more common classes in the source domains 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 strongly 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 the partial overlap situation.

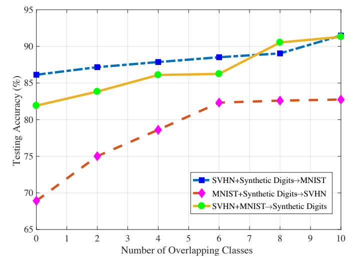


Figure 3: Accuracy in the target domain as a function of the number of overlapping classes between the source domains.

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. * indicates there is no adversarial part in the A^3 module, *i.e.,* only the DAT module. The best results are denoted in bold.

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[39]	78.56±0.47	72.33 ± 0.04	60.86±0.21	70.58 ± 0.24
M^{3}SDA[26]	81.52 ± 1.55	72.91 ± 0.68	54.26±0.66	69.56 ± 0.96
AENT[44]	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 of image classification, under the partial overlap setting with 4 common classes.

4.2. 2D Semantic Image Segmentation

Setup. In the single mode semantic segmentation setting, we adopt the synthetic-to-real image semantic segmentation setup. The synthetic image datasets GTA5 [30] and the SYNTHIA [32] 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 spaces of SYNTHIA and GTA5 are non-overlapping. In the SYNTHIA dataset, the label of 7 classes are available, incl. road, sidewalk, building, vegetation, sky, person and car. In GTA5, the labels of 12 classes are available, being wall, fence, pole, light, sign, terrain, rider, truck, bus, train, motorcycle and bicycle. Furthermore, we also explore the performance of our model when the images of the two source domains are fully labeled. Moreover, we ver-

Method	NT	T
Source	17.7	24.0
AdaptSegNet[34]	7.7	30.8
MinEnt[37]	27.1	30.1
Advent[37]	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)

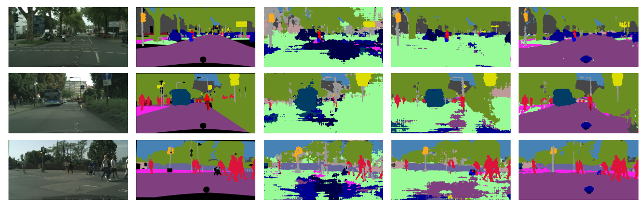
MBR	UM	A^3	PSF	ADV	NT	T
					17.7	24.0
✓					20.9	21.4
✓	✓				27.6	36.8
✓	✓	✓*			29.1	37.0
✓	✓	✓			36.3	38.1
✓	✓			✓	35.4	40.9
✓	✓		✓		31.4	41.5
✓	✓	✓		✓	40.1	41.5
✓	✓	✓	✓		37.3	42.4
✓	✓	✓	✓	✓	**40.6**	**42.8**

(b)

Table 5: (a) Quantitative comparison of single mode semantic segmentation, SYNTHIA+GTA5→ Cityscapes. The mIoU results are reported for 19 classes. (b) Ablation study for single mode segmentation. * indicates there is no adversarial part in the A^3 module, *i.e.,* only the DAT module. "ADV+PSF" means to combine "ADV" and "PSF" by completing the label space and generating pseudo-labels in the source and target domains, then adversarial alignment in the output space is adopted during the second stage training.

Method	Base	mIoU*	mIoU
Source	ResNet-101	42.8	39.1
AdaptSegNet[34]		45.2	40.8
Minentropy[37]		46.4	42.2
Advent[37]		46.7	42.9
Ours w/o PSF		**46.8**	**43.1**
Source[43]	VGG-16	37.3	–
MADAN[43]		41.4	–
Ours w/o PSF		**41.9**	**38.0**

Table 6: Single mode segmentation results, under fully-labeled setting and "T". mIoU* is the mean IoU of 16 classes in SYNTHIA, while mIoU is that of all 19 classes.



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

Figure 4: Qualitative results of 2D semantic segmentation.

ify the effectiveness of our model when the taxonomies of different source domains are inconsistent. In those inconsistency experiments, for GTA5, the labels wall, fence, pole, light, sign, terrain, truck, bus, train, person (incl. person and rider), cycle (incl. bicycle and motorcycle) are available. In SYNTHIA, the labels road, sidewalk, building, vegetation, sky, person, rider, car, public facilities (incl. wall, fence, pole), motorcycle and bicycle are available. In order to further evaluate the performance of all methods when combined with the pixel-level domain adaptation methods [45, 16], we conduct experiments in two settings; 1) source domain images are not translated with CycleGAN [45], named as "NT"; 2) source domain images are translated with CycleGAN, named as "T". Also, in order to verify model performance combined with output-level adaptation method [34], we conduct additional experiments which include "ADV" in the fully-supervised adaptation stage. "ADV" generates the complete source domain label as in PSF, and then trains the semantic segmentation model via adversarial adaptation between pseudo-complete source domain and unlabeled target domain in the output-level space. For fair comparison, all the methods use the DeepLabv2-ResNet101 [4, 14] semantic segmentation network.

Comparison with SOTA. In Table 5a, we show a quantitative comparison for semantic segmentation between our method and other SOTA methods. It is shown that our method without adding PSF strongly outperforms the adaptation-based AdaptSegNet[34], the self-supervision-based MinEnt[37], and the method combining adaptation and self-supervision Advent [37]. Our method achieves 36.3% and 38.1% in the "NT" and "T" settings, resp. Similar to the image classification results, without using the translated source images, the adaptation-based methods suffer from negative transfer 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 guidance for the complete label space to some extent. So all adaptation-based or self-supervision based methods perform much better in the "T" situation, compared with the non-adapted baseline. Yet, even in the "T" situation, our method still provides an advantage by further completing the label space, through our partially supervised adaptation. This proves the effectiveness of our method in preventing negative transfer and in completing the label space. By further adding the second "fully-supervised" adaptation stage, the model achieves a new SOTA performance in both the "T" and the "NT" settings. An ablation study, see Table 5b, confirms all parts of our method add to its performance, and the output space alignment "ADV" is helpful as well. Fig. 4 shows qualitative results on Cityscapes.

Fully labeled. In the fully labeled setting, *i.e.,* the source domain images are labeled with all 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, 43.1% vs. 40.8%, 42.2%, and 42.9%. Our model also outperforms the SOTA method for multi-source domain adaptive semantic segmentation MADAN [43], 41.9% vs. 41.4%.

Inconsistent Taxonomies. Table 7 shows that our method is advantageous when taxonomies are inconsistent, 40.0% vs. 28.1%, 31.9%, 32.2%. In the partially supervised adaptation stage, as in Sec. 3.2.6, by adding higher weights to "person", "rider", "motorcycle" and "bicycle" for SYNTHIA and "wall", "fence" and "pole" for GTA5, our method can achieve a higher performance than inference without weighting, 37.2% vs. 35.3%. After the fully supervised adaptation stage, the performance can be further improved to 40.0%. The detailed performance for inconsis-

Method	wall	fence	pole	person	rider	motorcycle	bicycle	mIoU
Source	2.6	12.0	12.3	40.6	0.5	0.1	28.6	19.8
AdaptSegNet[34]	7.1	2.6	4.0	33.2	6.9	1.8	37.6	28.1
Minentropy[37]	6.7	**18.1**	23.0	28.8	6.6	1.0	42.3	31.9
Advent[37]	6.2	11.5	11.4	32.8	12.2	0.9	41.2	32.2
Ours w/o PSF	12.3	15.2	21.2	48.4	3.3	1.3	42.4	35.3
Ours w/o PSF *	**14.1**	15.3	**30.6**	48.1	17.9	13.0	42.1	37.2
Ours (PSF)	13.3	17.9	**30.6**	**53.7**	**18.2**	**19.8**	**43.2**	**40.0**

Table 7: Quantitative comparison of single mode segmentation, with inconsistent taxonomies, in the "T" setting. *During inference, an additional weights map is adopted in case of inconsistent taxonomies as in Sec. 3.2.6. The detailed performance on inconsistent taxonomies classes is also shown. The mIoU is reported for 19 classes.

tent taxonomies classes in Table 7 underlines the effectiveness of our method for the inconsistent taxonomies.

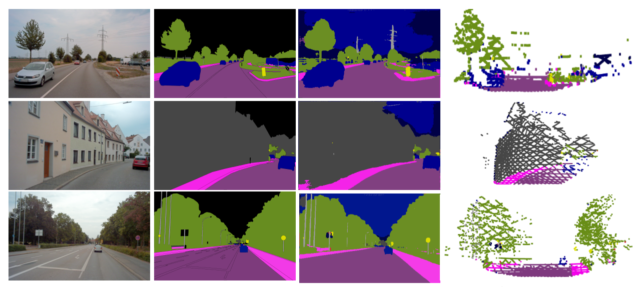
4.3. Cross-Modal Semantic Segmentation

Setup. In the cross-modal semantic segmentation setting, the 2D RGB images from Cityscapes [7], and the 3D LiDAR point clouds from Nuscenes [1] are treated as two different source domains, while the paired but unlabeled 2D RGB images and 3D point clouds from A2D2 [11] are used as the target domain. There are 10 classes in total that need to be transferred to the target domain. In Cityscapes, the label for 6 classes are given, covering road, sidewalk, building, pole, sign and nature. In Nuscenes the labels for 4 classes are given, incl. 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. Following [18], we adopt U-Net-ResNet34 [31, 14] as the 2D semantic segmentation network, and SparseConvNet [12] for 3D semantic segmentation. Due to the challenge of aligning features for the 3D point clouds, the A^3 module is not included in the cross-modal setting.

Comparison with the SOTA. As shown in Table 8, similar to the image classification and the single mode semantic segmentation results, the SOTA cross-modal unsupervised adaptation method xMUDA [18] shows an obvious negative transfer effect, resulting in a 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 networks are averaged in the target domain through the 2D and 3D point correspondence during training, and the fused prediction probability is optimized using the minimum entropy loss [37]. 2) ES + KL: the KL-divergence [18] is utilized to align between the 2D/3D prediction and the fused predictions for the corresponding points in the target domain, resp. 3) xMUDA + AKL: the KL-divergence alignment between 2D and 3D in the target domain is weighted adaptively, to reduce the wrong guidance from the unlabeled parts. 4)

Cityscapes + Nuscenes → A2D2	2D	3D	Fuse
Source	37.5	2.0	42.5
xMUDA[18]	16.3	1.7	9.1
ES + MinEnt[37]	22.3	1.5	20.8
ES + KL[18]	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 8: Quantitative comparison of cross modal segmentation, Nuscenes+Cityscapes→ A2D2. "Fuse" represents the average fusion of the prediction probability from 2D models and 3D models; the final class prediction is the maximum of the fused probability. "ES" means 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 mIoU is reported over 10 classes on A2D2.



(a) A2D2 (b) Ground Truth (c) 2D (Ours) (d) 3D (Ours)

Figure 5: Qualitative results of the cross-modal setting.

xMUDA + AKL + COMP: following baseline 3), another constraint, that the weights related to 2D and 3D need to be complementary, is added. It is shown that our method prevents 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 strongly improved, achieving 54.9% and 37.1%, resp. In Fig. 5, we show qualitative results in the target domain. The good performance proves the effectiveness of our method for the mDALU with partial modalities. This opens up 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 proposed the multi-source domain adaptation and label unification with partial datasets problem, called mDALU. Then we proposed a novel multi-stage approach for mDALU, including partially and fully supervised adaptation stages. Our approach is demonstrated through extensive experiments on different benchmarks.

Acknowledgements. This research has received funding from the EU Horizon 2020 research and innovation programme under grant agreement No. 820434. Dengxin Dai is the corresponding author.

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. 8

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

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

[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, 8

[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

[11] 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. 8

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

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

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

[15] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *NeurIPS*, 2018. 2

[16] 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

[17] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Köring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *CVPR*, 2021. 1

[18] 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, 8

[19] 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. 2

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

[21] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *CVPR*, 2020. 2

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

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 2

[24] 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

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

[26] 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

[27] 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

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

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

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

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

[32] 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

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

[34] 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, 8

[35] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2

[36] 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

[37] 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

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

[39] 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, 6

[40] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. In *ICLR*, 2017. 2

[41] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *CVPR*, 2018. 2

[42] 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

[43] 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, 7

[44] 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. 2, 6

[45] 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