國家實驗研究院
National Applied Research Laboratories

（編號：NCHC-112-01-F-004）

經費來源： ■01 公務□02 非公務

機密(E)： □是 ■否

出國類別： □ A 考察/訪問 □ C 進修/研究 ■ F 工作會議/研討會
□ G 推廣佈展 □ H 學術會議

分項計畫名稱：技術研發與環境開發

2023 IIAE International Conference on Industrial Application Engineering

參加 IIAE ICIAE 國際研討會

論文發表

出國報告書

服務單位： 國家實驗研究院高速網路與計算中心

出國人姓名職稱： 盧沛怡 助理研究員

出國地點： 日本那霸市

出國日期： 112 年 3 月 26 日至 112 年 3 月 31 日

報告日期： 112 年 3 月 30 日

NARLabs
國家實驗研究院
National Applied Research Laboratories

摘　要

此次會議 International Conference on Industrial Application Engineering 由日本工業應用工程師學會(IIAE)所主辦，IIAE 成立於 2012 年，它是一個非盈利的學術組織，旨在促進各行業研究和新實踐的發展。 IIAE 的研究領域包括工業原型的開發、技術創新的商業化以及技術在工業管理和行銷中的應用。ICIAE 主要為一個針對電機、資工與機械工程相關研究發表論文的平台，希望能成為研究學術機構和工業組織之間溝通的橋樑。而 keynotes 部分也邀請了包含日本、中國與英國等地的知名學者來演講，主題非常多樣化，包含了結合當地首里城重建相關的科技演講，以及 AI 人工智慧與深度學習相關的偵測問題，另外，也有與軟體開發與跨學科關係的探討。而本人的論文發表為「Federated Multi-source Domain Adaptive Object Detection with Probabilistic Teacher」，發表日期為 3/30 日，內容主要是介紹透過聯邦式學習架構，在多攝影機的架構下，將多攝影機視為來源域，可以在保護各個攝影機隱私內容的同時，共同訓練一個物件偵測模型，可以在目標域有良好的正確率。發表論文當天，也從各國學者的問題中得到了一些的啟發。另外，在同一發表場次中，也有一些攝影機影像相關論文發表，對於未來的系統創新及設計上，有極大的幫助。

NARLabs
國家實驗研究院
National Applied Research Laboratories

活動日程表

日期	行程規劃
3/26	出發並至會場報到
3/27	參加 technical tool 與學者交流
3/28	參加 ICIAE 會議
3/29	參加 ICIAE 會議
3/30	參加 ICIAE 會議
3/31	早上前往機場搭機返台

NARLabs
國家實驗研究院
National Applied Research Laboratories

目　　次

NARLabs
國家實驗研究院
National Applied Research Laboratories

1.目的

於 IIAE ICIAE(International Conference on Industrial Application Engineering)國際會議中發表論文：Federated Multi-source Domain Adaptive Object Detection with Probabilistic Teacher。

2. 國際研究會議紀要

此 IIAE ICIAE 2023 國際會議是由 IIAE 所舉辦的，含蓋許多不同的領域，包括：Electrical technology, Sensing technology, Information technology, Network technology, Image processing…等。包含不同類型的 keynote，其中一位講者 Prof. Juan Jose Castro 是琉球大學的教授，講了一個比較不學術的題目” Reconstruction of the Shuri Castle: Resilient Structure Design and Use of Local Timber Material” 主要介紹對沖繩首里城進行重建的計劃。由於首里城曾多次被大火摧毀，最近一次是在 2019 年，因此在 2020 年啟動了這個項目。這個計劃考慮了增加結構的耐震性和使用當地木材的觀點，以創建一個更好的首里城。文章中還介紹了在設計結構時如何使用當地木材，並經過實驗驗證其材料強度足夠用於構造元件。另一位講者題目為 "Longitudinal Tear Detection of Conveyor Belt Based on Improved MFCC and DenseNet Network”, 主要是透過深度學習相關技術來達到機械材料中的 longitudinal tear detection，算是多領域結合的應用。很好的詮釋了本協會主要想推動的目標：希望能成為研究學術機構和工業組織之間溝通的橋樑。

3/30 日早上主要就是我的論文發表時間，這次發表的論文主

要是介紹透過聯邦式學習架構，在多攝影機的架構下，將多攝影機視為來源域，可以在保護各個攝影機隱私內容的同時，共同訓練一個物件偵測模型，可以在目標域有良好的正確率。我們的架構主要包含三部分(1) 透過 weak-strong augmentation 來增加輸入影像的多樣化 (2) 透過 Probabilistic teacher 來改善 pseudo-label 的品質 (3) 透過 FedAvg 來整合不同 clients 的模型。針對這三部分都有做深入的解釋。另外我們在結果部分也比較了各種方法，包含 single-source domain adaptation, multi-source domain adaptation 以及 privacy-preserving model aggregation algorithms，我們的方法在保護隱私之下，比其他聯邦式學習的演算法正確率都來的高。

Figure 1 論文發表當天現況

論文發表結束後，有位美國 California Polytechnic State University 的教授 Maria Pantoja 問到我們的 weak-strong augmentation 的功能以及差異為何，我告訴他我們的論文中有提

到所有 augmentation 的細節，之所以在 student model 用到 strong augmentation 的原因，是因為它包含了有 label 的 source image, 因此就算有很多種不同的 augmentation, 都可以有正確的標記，而在 teacher model 的部分，由於只把 target domain 當作輸入來源，並且產生的是 pseudo-label, 若是使用 strong augmentation, 會產生太多不正確的 pseudo-label，因此在這裡只使用 weak augmentation 給 teacher model。

另外幾個 session 中，有一些比較有趣的題目，是使用多個魚眼資料來判斷物件的距離，精準度可以到 2cm, 設定挺有趣，但實用價值有待商榷，會後我有問他是否每個魚眼攝影機都是用於他們調好的參數，其實是不行的，每個攝影機都要先校正一次，因此會花費蠻大的功夫在校正的，另外，魚眼攝影機的成本也比普通攝影機來的高，且影像較為扭曲，必須要進行校正，相較之下不如使用多個攝影機來建立環場影像，也可達到相同的目的，可節省成本，並且省去校正的工。

3.心得及建議

由於此會議為國際上知名的會議，不僅在 keynote 部分學習到不少新的技術，在其他論文發表場次上，也學習到許多新知。

論文發表部分，在會議前就假設了許多可能被提問的問題，發表論文前也思考過系統在不同層面是否會有漏洞，因此這次被提問的問題，都能夠順利回答，也讓提問者感到滿意。因此，相信未來不管在發表論文，或是發表成果，只要有充足的準備，都能從容應對。

另外，本次會議除了與提問學者交流，也主動的對有興趣的議題進行提問， 接下來我們內部也有一些魚眼交通資料的車流偵測，我覺得那幾篇九州大學老師的魚眼攝影機相關研究，都可能會對我們的研究有幫助，這次參加會議有機會認識那位老師，之後也許可以進行進一步的合作。

4.出國效益

1. 了解國際數位學習的潮流
2. 擴展國際學術人脈
3. 了解各種不同領域如何結合 AI 與深度學習

附錄—投稿論文

Federated Multi-source Domain Adaptive Object Detection with Probabilistic Teacher

1. Introduction

As surveillance devices become increasingly cheaper, cameras can now be found nearly everywhere. To obtain a general CNN model for a specific task in artificial intelligence, e.g., object detection, an intuitive idea is to use surveillance videos obtained from different cameras to collaboratively train a model which may achieve good performance on an unseen scene. However, video owners may be unwilling to share their data due to privacy concerns as the videos may reveal personal and private information. Meanwhile, it is difficult to obtain a decent detection model for images under different configurations, since videos captured by different cameras correspond to different scenes, styles and camera setups which cause large domain gaps. Currently, this challenge problem is treated as a multi-source domain adaptation problem under a federated setting.

To extract values from vast amount of images and videos recorded by surveillance systems, object detection is the first step for any further applications. With the development of deep learning, object detection has become one of the most thriving fields in computer vision. Although domain adaptation has been widely studied for image classification[(1)], using it in object detection is more challenging because the detection involves both classification and regression problems. An unsupervised domain adaptation method was first proposed in DAF[(2)], proposed an idea by adapting in two different levels, image level and instance level.

To the best of our knowledge, multi-source domain adaptive object detection (MSDAOD) is an emerging research problem and most related works has just been published within the last two years[3][4][5]. Recent works effectively utilize all of the source data information simultaneously to train a model with better performance. However, because of privacy preserving settings, data cannot be revealed to anyone except the data owner. Thus, unlike MSDAOD approaches, the data owner can only share models instead of data for collaborative training.

To resolve the foregoing problem, a federated architecture is proposed, as shown in Fig. 1, wherein different data owners (clients) upload only their model instead of data to the server, while the server collects and aggregates different models provided by clients. In traditional federated learning approaches, such as FedAvg[6], performance will decrease considerably after aggregating models at the early stage of training. Thus, exchanging model weights between server and clients for many rounds can reduce the diversity between different models and gradually obtain a more stable and domain-invariant global model.

Recently, many researchers have adopted teacher-student architectures for the domain adaptation problem[18], and obtaining quality pseudo labels is critical for the final results. For filtering false pseudo labels, an extra hyper-parameter such as threshold may be required. In this paper, a threshold-free probabilistic teacher technique is adopted on the client side to train a local model with labeled local data (source data) and unlabeled target data. As clients only have unlabeled target data, this is regarded as a self-training method, which typically relies on the pseudo labels generated by a teacher model to update the student model. However, the pseudo label generated from the teacher model usually contains a substantial number of errors and false positives because of the large domain gap between the labeled source data and unlabeled target data. We apply Weak-Strong augmentation[31] to increase the variety of input images for the student model while suppressing the false positives pseudo labels generated by the teacher model.

We evaluate our method by using several commonly used datasets for benchmarking object detection tasks, including Cityscapes[8], KITTI[9] and BDD100k[10]. The experiments were conducted on multiple real-world domain discrepancy cases, such as adapting from Cityscapes, KITTI to BDD100k. According to the experimental results, the proposed method can maintain good performance under the privacy preserving restriction.

Contributions of this paper include:

- A novel scenario of federated multi-source domain adaptive object detection is stated.
- A federated architecture leveraging probabilistic Teacher-Student Mutual Learning and weak-strong augmentation in cross-domain object detection is proposed.
- Effectiveness of model aggregation algorithms in server site and different domain adaptive object detectors in client sites with empirical experiments are evaluated.

2. Related Work

Unsupervised Domain Adaptive Object Detection (UDAOD). Domain adaptation has been researched for many years and most approaches try to reduce the domain gap by minimizing the distance between similar image features obtained

in two different domains. The application on object detection is more challenging than the classification problem as the latter contains both classification and localization parts. The first work(2) applying domain adaptation on object detection adopts both image alignment and instance alignment to diminish the domain gap. Many domain adaptive object detection solutions(11)(12)(13) try to use adversarial feature learning to reduce the domain gap between source and target domains.

Pseudo label based self-training(14)(15) and mean teacher training(16)(17) are other popular types of UDAOD approach, while the former focus on how to generate more reliable pseudo label and the latter utilize unlabeled data to improve model generalization by progressively training a detector in a student teacher framework. Although the number of UDAOD solutions has been growing recently, and is still developing, most of those works train both source and target data together as the privacy preserving issue has yet to be a major concern.

Multi-source domain adaptation (MSDA). Multi-source unsupervised domain adaptation algorithms have been proposed with early theoretical analysis(19) along with recent developments based on deep learning(20)(21)(22). Although more information is obtained for adapting to the target domain, along with locally available multiple source data, domain shifts between multiple sources (20)(21)(23) need to be reduced whenever possible. However, some approaches cause performance decay after adopting such methods(21)(22), while others ignore the loss of the discriminating ability of image feature when aligning different domains.

While most MSDA works focus on image classification, DMSN(3) is the first to introduce MSDA into object detection while MTK(4) and TRKP(5) are proposed later. In DMSN, feature alignment among sources and pseudo subnet learning are developed for their weighted combination. However, its temporary domain discrepancy measurement leads to a local optimum. In MTK, a network is designed to align features from both source-to-source and source-to-target pairs. Nevertheless, the network scale may increase with respect to the number of source domains. In TRKP, it proposed a framework which can collaboratively train a model with less domain specific information and preserve more target-relevant knowledge from different source domains. However, the above approaches require multi-domain data in the training process, which violates a privacy protection setting.

Federated Learning. Federated learning (FL) provides a promising privacy-preserving solution for obtaining a collaborative model across multiple clients(6), which helps clients keep their data locally during the training process. As different methods of model aggregation may significantly affect system performance, quite a few approaches have been proposed, such as FedAvg(6), FedSGD(6) and FedMA(7). In FADA(24), federated learning is first treated as a domain adaptation problem, and tries to resolve the disentangle problem between different source domains. In FedDG(25), a domain generalization model on medical image segmentation is obtained by exchanging the amplitude spectra of Fourier transformed images between different sites. Nonetheless, most of previous works focus on image classification or segmentation, rather than object detection, because

of the simplicity.

Recently, different methods have been developed to adopt a knowledge distillation (KD) technique, a teacher-student architecture, for FL to reduce communication cost(26), and to integrate more knowledge from different sites(27). Other works employ KD only on client sites(28)(29), which do not decrease the communication. In FD(30), only soft labels are transferred to reduce the communication cost, but with the performance compromises. Nonetheless, these methods are developed to apply simple classification applications that cannot migrate directly to handle the object detection problem. To the best of our knowledge, applying federated learning to a domain adaptive object detection problem has never been previously mentioned.

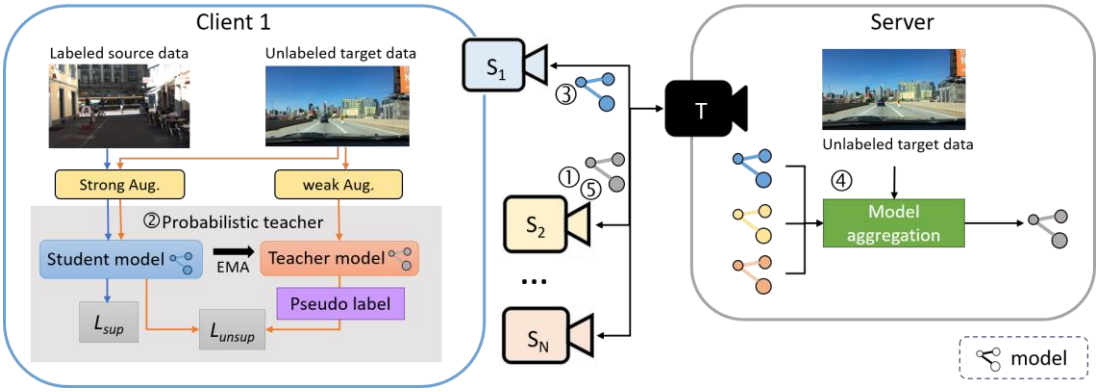

Fig. 1 System architecture of the proposed federated scenario. The server is responsible for model aggregation and sends a global model to clients. The clients feed both local dataset (source data) and global data (target data) to Weak-Strong augmentation as the inputs of a Mutual Learning Teacher-Student Framework which adopts probabilistic teacher technique on object detector. The supervised and unsupervised losses are used to update the Student model weight while the Teacher model weight is gradually updated by exponential moving average (EMA) technique. Clients send trained models back to the server and this procedure repeat for *R* rounds.

3. Proposed Method

3.1 System Architecture and Overview

An overview of our framework is presented in Fig. 1. Our framework consists of two roles, server and clients. The server assigns a global model to clients and aggregates different models trained by clients, while each client applies probabilistic teacher domain adaptive technique on local dataset to generate a local object detector. To map our scenario to traditional unsupervised domain adaptive problem, we treat public global data as unlabeled target data and private local data as labeled source data.

The flow of our framework can be summarized as follows: (1) The server sends an initial global model to clients. (2) Each client independently trains a Teacher-Student model on the private local data (source) and a public global data (target). (3) Clients send Student models back to the server. (4) The server aggregates models from different clients to get a new global model. (5) The server sends a new global model to clients and repeats (2) ~ (4) for *R* rounds.

3.2 Teacher-Student Mutual Learning Framework

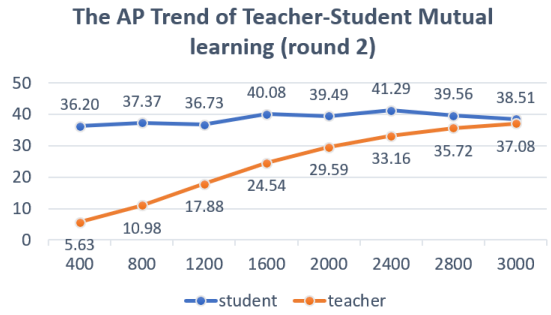

Fig. 2 The AP trend of the Teacher and Student model (round 2) by using KITTI as source data to show the process that Teacher model is gradually updated by Student Model.

On the client side, we adopted a Teacher-Student Mutual Learning Framework which contains two models of identical architecture: Student model and Teacher model. In each round of training, client copies global model weights to both Student and Teacher models. The Teacher model provides pseudo-labels to update the weights of the Student model as shown in Fig. 2. On the other hand, the Teacher model is updated with the exponential moving average (EMA) technique. Here the Teacher model can also be regarded as a temporal ensemble of Student models in different time steps since it copies the weights of the Student model temporally. The EMA can be updated with:

$$\theta_t^i \leftarrow \alpha\theta_t^{i-1} + (1-\alpha)\theta_s^{i-1}, \qquad (1)$$

where θ_t^i and θ_s^i denote the weights of the Teacher and Student models in the *i*-th iteration, respectively, and α is the EMA rate.

While most of previous works focused on inter-domain alignment, it is visually proved in PT[32] that *intra-domain gap* is the main bottleneck which restricts the performance of UDAOD. Thus, to lower the false negatives caused by different anchor sizes, we feed both source and target images with strong data augmentation as inputs of the Student model. To avoid generating too many false pseudo labels, the Teacher model adopt weak augmentation instead of strong augmentation on unlabeled target data.

For multi-source domain adaptation, assume there are *N* labeled source domains and one unlabeled target domain. Suppose a source image I^S, is annotated with *M* bounding boxes $B = \{b_j\}_{j=1}^{M}$ as well as their class labels $C = \{c_j\}_{j=1}^{M}$. The *i*-th source domain and target domain can be represented as $S_i = \left\{\left(I_j^{S_i}, B_j^{S_i}, C_j^{S_i}\right)\right\}_{j=1}^{n_{S_i}}$ and $T = \{(I_j^T)\}_{j=1}^{n_t}$, with n_{S_i} and n_T denoting the total numbers of source and target images, respectively.

In this paper, we employ a two-stage object detector, Faster R-CNN[34], as the base detector. The total loss of our framework can be written as:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_{usp}\mathcal{L}_{usp} \qquad (2)$$

where $\mathcal{L}_{sup}$ is the supervised loss on labeled source data, and $\mathcal{L}_{usp}$ is an unsupervised loss on unlabeled target data. λ_{usp} is the hyper-parameter used to control the weighting of the loss from target domain. The supervised loss for training the Student

model can be defined as:

$$\mathcal{L}_{sup} = \mathcal{L}_{cls}^{rpn}(B^S, C^S; I^S) + \mathcal{L}_{reg}^{rpn}(B^S; I^S) + \mathcal{L}_{cls}^{roi}(B^S, C^S; I^S) + \mathcal{L}_{reg}^{roi}(B^S; I^S), \quad (3)$$

where RPN loss $\mathcal{L}^{rpn}$ is the loss for learning the Region Proposal Network (RPN), which is designed to generate candidate proposals, while Region of Interest (ROI) loss $\mathcal{L}^{roi}$ is for the prediction branch of ROI head, with both RPN and ROI perform bounding box regression (*reg*) and classification (*cls*). The original Faster R-CNN uses L1 loss for *reg* and cross-entropy for *cls*, but we adopt binary cross-entropy loss for both *reg* and *cls*.

3.3 Probabilistic Teacher

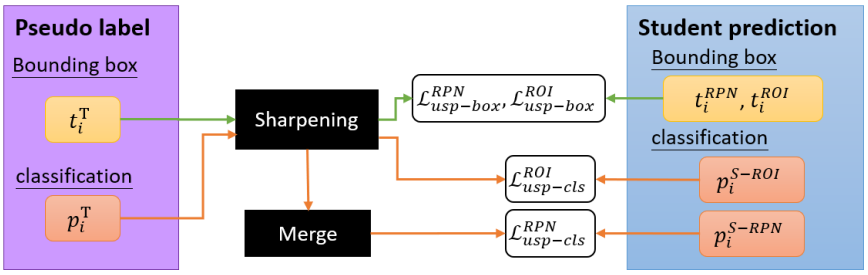

Fig. 3 The flow of computing the unsupervised loss on unlabeled target data. The Teacher model generates pseudo labels for target data, including probability distributions of classification and bounding box coordinates, and passes them to a sharpening function followed by a merging operation, to guide the Student training.

Probabilistic Teacher is a threshold-free technique for object detection which maps each set of bounding box coordinates to a single Gaussian model. Therefore, the bounding box regression loss can be implemented by a cross-entropy function between the ground-truth distribution and the predicted one. Moreover, using variance of Gaussian model, rather than foreground score, to estimate IoU (intersection of union) will be more accurate. (Details of the proof can be found in PT[32].)

In Teacher-Student Mutual Learning Framework, we feed the weak augmented unlabeled target data to the Teacher model to generate pseudo labels for optimizing the 2nd term of Eq. (2), i.e., unsupervised loss $\mathcal{L}_{unsup}$. Similar to supervised loss in Eq. (3), unsupervised loss also consists of four training losses wherein the two *cls* losses can be formulated by the following probability distribution:

$$\mathcal{L}_{usp-cls}^{RPN} = \frac{1}{N_{cls}^{RPN}} \sum_i \mathcal{H}\left(\mathcal{M}\left(S(p_i^T, \tau)\right), p_i^{S-RPN}\right), \quad (4)$$

$$\mathcal{L}_{usp-cls}^{ROI} = \frac{1}{N_{cls}^{ROI}} \sum_i \mathcal{H}\left(S(p_i^T, \tau), p_i^{S-ROI}\right),$$

where p_i^T, p_i^S are the *i*-th classification probability distribution predicted by the

Teacher and Student model, respectively. p_i^{S-RPN} and p_i^{S-ROI} indicate the prediction in RPN and ROI head, with *S* and τ being a sharpening function and a temperature factor, respectively. In addition, *H* denotes the cross-entropy function and *M* is the merging operation for summing up all foreground category probabilities to achieve foreground/background probability distributions to guide the training of RPN, while N_{cls}^{ROI} and N_{cls}^{RPN} are the batch size in ROI head and RPN, respectively.

For two *reg* losses, they can be formulated as:

$$\mathcal{L}_{usp-box} = \frac{1}{N_{box}} \sum_i \sigma \mathcal{H}(S(t_i^T, \tau), t_i), \quad (5)$$

where t_i^T, t_i are the *i*-th bounding box coordinate probability distributions predicted by Teacher and Student model, respectively. σ is a sign function to indicate whether the predicted bounding box is matched to the region proposal. A schematic diagram for computing an unsupervised loss mentioned above is illustrated in Fig. 3.

3.4 Model Aggregation on Server

Algorithm 1 shows the details of FedAvg model aggregation algorithm. For every federated round *r*, *N* clients download the same initial model M_{r-1} from the server and perform probabilistic teacher domain adaptation to update the model on local data. On the client side, the training process minimize the loss *L* over local mini-batches *b* for *E* iterations before the local model being sent back to the server. The server then averages the model weights collected from all clients to get a new model weight M_r. After the model aggregation finished, this new global model will be transmitted to all clients again. The above procedure will repeat for *R* rounds.

Algorithm 1 FedAvg.

Input: *N* source domains $\{S_i\}_{i=1}^N$; a target domain *T;* detectors from *N* sources $\{M^1, M^2, \dots, M^N\}$*;* total rounds *R;* local iterations *E* and weight control parameter λ_{usp}.
Output: A well-trained model M^t
Initialization: Server initializes and then sends federated model M_0 to *N* clients.

Server:
for *r* = 1, ...,*R* **do**
  for *i* = 1, ...,*N* **do** //local computation at clients
    Adopt M_{r-1} as initial model
    M_r^i=ClientUpdate(*i*, M_{r-1})
  $M_r = \sum_{i=1}^{N} M_r^i$
return M_r
ClientUpdate(*i*, M^{global}): //Run on client *i*

for j = 1, ..., E **do**
Sample mini-batch from $\{(I_b^{s_i}, B_b^{s_i}, C_b^{s_i})\}_{b=1}^{n_{s_i}}$
Compute object detection loss:
$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_{unsup}\mathcal{L}_{unsup}$$
update model M^{local} according to loss function
return M^{local}

4. Experimental Results

4.1 Datasets

In this section, we introduce all datasets used in the experiments, including Cityscapes[8], KITTI[9] and BDD100k[10].

Cityscapes. The Cityscapes dataset[8] collects data by capturing images from outdoor street scenes in normal weather conditions from 50 cities and include diverse scenes. There are 2,975 images for training and 500 images for validation with dense pixel-level labels. All of the labels are transformed to bounding box annotations.

KITTI. The KITTI dataset[9] is collected by an autonomous driving platform, containing street scenarios taken in cities, highways, and rural areas. It contains 14,999 images and 80,256 bounding boxes. Only 7,481 training images are used as source images here.

BDD100k. The BDD100k dataset[10] is a large-scale dataset containing 100,000 images, including 70,000 training images and 10,000 validation images with bounding box annotations. Images of the dataset are captured at different times of a day, and we assign daytime as the target domain in cross camera adaptation.

4.2 Implementation Details

We adopted VGG16[33] as the backbone for the Faster R-CNN[34] detection network, and follow the most common setting[2] of UDAOD. Besides, we use the pre-trained weights of ImageNet[34] for the initial global model for both the Teacher and Student models in the clients. The batch size of each dataset was set to 16 for 3 rounds of training, while each round contains 3,000 iterations with the learning rate set to a fixed value of 0.016 during the entire training stage. The optimizer of the network is based on Stochastic Gradient Descent (SGD) with a momentum of 0.9 and weight decay of 0.0001. The parameter α in the exponential moving average (EMA) for updating the Teacher model was set to 0.9996, while Detectron2 is used in the implementation. Moreover, λ_{unsup} was set to 1 for the total loss computation, and τ was set to 0.5, for simplicity.

4.3 Comparison with Existing Approaches

In this paper, our method will be compared (in the next subsection) with previous state-of-the-art approaches, which include:

1. **Source-only setting:** We apply Faster R-CNN[34] as base detector to train a model on source data and directly test on target dataset without domain adaptation technique.
2. **Domain adaptation approaches:** We utilize different state-of-the-art

domain adaptive object detectors to adapt single source or all sources to target dataset, which include SW[11], CRDA[35], UMT[17], and UBT[31].

3. **Multi-source domain adaptation (MSDA) methods:** we adopt MDAN[22], M^{3}SDA[21], DMSN[3], and TRKP[5] for MSDA in object detection. As privacy issue is not considered in these approaches, the trainer is able to collect information from all datasets.

4. **Privacy-preserving MSDA:** For a privacy-preserving setting, we simply use different model aggregation algorithms, such as FedAvg[6] and FADA[24], to merge different models trained individually by single source without revealing the source dataset on the server.

5. **Oracle:** We use fully labeled target images, e.g., BDD100k, to train an object detector as an estimated upper bound.

4.4 Privacy-preserving Cross-Camera Adaptation

Datasets captured by different devices often have different camera setups and specifications, such as angle, type, resolution and quality, etc., which cause a strong domain shift between these image sources. By following the DMSN[3] setting, we use KITTI and Cityscapes as source domains, while the daytime of BDD100k is treated as a target domain. Thus, the experiment corresponding to an adaptation from small-scale datasets to a large-scale dataset.

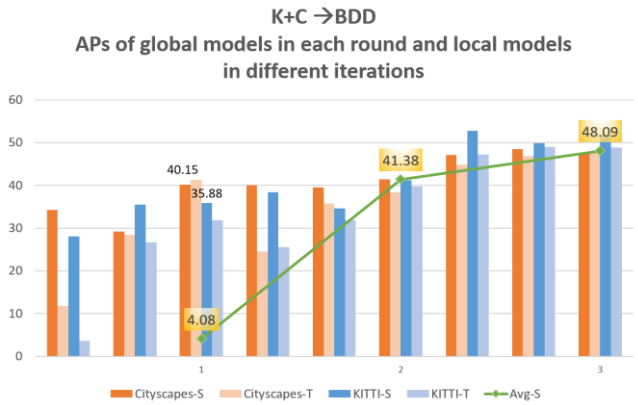

Fig. 4 The trend of APs for K+C→BDD. The orange bars indicate the results of a client model using Cityscape in local iterations while the blue bars use the KITTI dataset. The darker and lighter bars represent the results of the Student and Teacher model, respectively. The green line gives the APs of the global model after aggregating the client models.

Table 1 shows the experimental results evaluated on the common category, *car*, in terms of the widely used average precision (AP) for Cityscapes+KITTI to BDD100k. Whether or not adopting domain adaptation, the results of Cityscapes are much better than those in KITTI or the source-combined cases because the data distribution between Cityscapes and BDD100k is more similar than that between KITTI and BDD100k. In addition, merely by combining the sources does not help to bridge the domain discrepancy between different sources. In multi-source domain adaptation, TRKP[5] achieves the best performance because such method not only obtains the complete source datasets, but also restrains knowledge degradation between sources.

However, under the constraint of accessibility of source data, it is difficult to maintain the performance for data from an unseen domain. As shown in Fig. 4, the APs of the Student model in each client before the 1st round of model aggregation are 35.88% and 40.15% for KITTI and Cityscapes, respectively. However, after the 1st round of aggregation, the AP of the global model drop to 4.08%, which is extremely low. Nonetheless, the APs do increase afterwards, which may be interpreted as an indication of the diverse behaviors of the Student models at the early stage of training. Similar performance drop also occurs in FedAvg and FADA, but the AP values can only reach 43.3% and 43.2%, respectively, at the end, which is worse than the results of single source domain adaptation. On the other hand, the performance of our model is better than all other privacy preserving setting, as the AP of global model in the 3rd round can already reach to 48.1%. As the performance gap between our method and Oracle still exists, more effective model aggregation methods are required for keeping the domain invariant part between different models, which will be investigated in the near future.

Table 1. Results of adaptation from Cityscapes and KITTI to BDD100k (*daytime*). Averate precision (AP, %) on *car* category in target domain is evaluated.

Setting	Source	Method	AP on car
Single Source	C	FRCNN[34]	44.6
		SW[11]	45.5
		CRDA[35]	46.5
		UMT[17]	47.5
		UBT[31]	48.4
Single Source	K	FRCNN[34]	28.6
		SW[11]	29.6
		CRDA[35]	30.8
		UMT[17]	35.4
		UBT[31]	33.8
Source-combined DA	C+K	FRCNN[34]	43.2
		SW[11]	41.9
		CRDA[35]	43.6
		UMT[17]	47.0
		UBT[31]	47.6
Multi-source DA	C+K	MDAN[22]	43.2
		M3SDA[21]	44.1
		DMSN[3]	49.2
		TRKP[5]	58.4
Privacy-preserving Multi-source	C+K	FedAvg[6]	43.3
		FADA[24]	43.2
		FedPT (Ours)	**48.1**
Oracle	BDD100K	FRCNN[34]	60.2
Oracle	BDD100K	FRCNN[34]	60.2

5. Conclusions

In this paper, we proposed a federated learning problem and solved it as a multi-source domain adaptive object detection scenario which can train a global model without disclosing the source data of data owner. On the client side, more stable pseudo labels for the target domain are generated via Mutual Learning Teacher-Student Framework while Weak-Strong augmentation helps in increasing more reliable intra-domain anchors and reduces false positives. In addition, a domain adaptive object detector, Probabilistic Teacher is adopted for achieving better performance without extra threshold setting. Experimental results show that our approach outperforms other

privacy-preserving methods. In the future, we expect to explore effectiveness of different domain adaptation methods on client site, as well as further improvements of model aggregation on server site.

Acknowledgment

We thank to National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

References

(1) Duan, Lixin, Ivor W. Tsang, and Dong Xu : “Domain transfer multiple kernel learning.”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 3, pp. 465-479, 2012.
(2) Chen, Yuhua and Li, Wen and Sakaridis, Christos and Dai, Dengxin and Van Gool, Luc : “Domain adaptive Faster R-CNN for object detection in the wild”, Proceedings of the IEEE CVPR, 2018.
(3) Yao, Xingxu and Zhao, Sicheng and Xu, Pengfei and Yang, Jufeng : “Multi-source domain adaptation for object detection”, Proceedings of the IEEE/CVF ICCV, 2021.
(4) Zhang, Dan and Ye, Mao and Liu, Yiguang and Xiong, Lin and Zhou, Lihua : “Multi-source unsupervised domain adaptation for object detection”, Elsevier Information Fusion, Vol. 78, pp. 138-148, 2022.
(5) Wu, Jiaxi and Chen, Jiaxin and He, Mengzhe and Wang, Yiru and Li, Bo and Ma, Bingqi and Gan, Weihao and Wu, Wei and Wang, Yali and Huang, Di : “Target-Relevant Knowledge Preservation for Multi-Source Domain Adaptive Object Detection”, Proceedings of the IEEE/CVF Conference on CVPR, pp. 5301-5310, 2022
(6) McMahan, Brendan, et al. : “Communication-efficient learning of deep networks from decentralized data”, Artificial intelligence and statistics, PMLR, 2017.
(7) Wang, Hongyi, et al. : “Federated learning with matched averaging”, arXiv preprint arXiv:2002.06440, 2020.
(8) Cordts, Marius, et al. : “The cityscapes dataset for semantic urban scene understanding”, Proceedings of the IEEE conference on CVPR, 2016.
(9) Geiger, Andreas, Philip Lenz, and Raquel Urtasun : “Are we ready for autonomous driving? the kitti vision benchmark suite”, 2012 IEEE conference on CVPR, 2012.
(10) Yu, Fisher, et al. : “Bdd100k: A diverse driving video database with scalable annotation tooling”, arXiv preprint arXiv:1805.04687, 2018.
(11) Saito, Kuniaki, et al. : “Strong-weak distribution alignment for adaptive object detection”, Proceedings of the IEEE/CVF Conference on CVPR, 2019.
(12) Vs, Vibashan, et al. : “Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection”, Proceedings of the IEEE/CVF Conference on CVPR, 2021.
(13) Wang, Yu, et al. : “Domain-specific suppression for adaptive object detection”, Proceedings of the IEEE/CVF Conference on CVPR, 2021.
(14) Jiang, Junguang, et al. : “Decoupled Adaptation for Cross-Domain Object

Detection", arXiv preprint arXiv:2110.02578, 2021.
(15) Li, Xianfeng, et al. : "A free lunch for unsupervised domain adaptive object detection without source data", Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 10, 2021.
(16) Cai, Qi, et al. : "Exploring object relation in mean teacher for cross-domain detection", Proceedings of the IEEE/CVF Conference on CVPR, 2019.
(17) Deng, Jinhong, et al. : "Unbiased mean teacher for cross-domain object detection", Proceedings of the IEEE/CVF Conference on CVPR, 2021.
(18) Li, Yu-Jhe, et al. : "Cross-Domain Adaptive Teacher for Object Detection", Proceedings of the IEEE/CVF Conference on CVPR, 2022.
(19) Mansour, Yishay, Mehryar Mohri, and Afshin Rostamizadeh : "Domain adaptation with multiple sources", Advances in NIPS, 2008.
(20) Zhao, Sicheng, et al. : "Multi-source domain adaptation for semantic segmentation", Advances in NIPS, 2019.
(21) Peng, Xingchao, et al. : "Moment matching for multi-source domain adaptation", Proceedings of the IEEE/CVF ICCV, 2019.
(22) Zhao, Han, et al. : "Adversarial multiple source domain adaptation", Advances in NIPS, 2018.
(23) Venkat, Naveen, et al. : "Your classifier can secretly suffice multi-source domain adaptation", Advances in NIPS, pp. 4647-4659, 2020.
(24) Peng, Xingchao, et al. : "Federated adversarial domain adaptation", arXiv preprint arXiv:1911.02054, 2019.
(25) Liu, Quande, et al. : "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space", Proceedings of the IEEE/CVF Conference on CVPR. 2021.
(26) Sattler, Felix, et al. : "Cfd: Communication-efficient federated distillation via soft-label quantization and delta coding", IEEE Transactions on Network Science and Engineering Vol.9, No. 4, pp. 2025-2038, 2021.
(27) Lin, Tao, et al. : "Ensemble distillation for robust model fusion in federated learning", Advances in NIPS, pp. 2351-2363, 2020.
(28) Jiang, Donglin, Chen Shan, and Zhihui Zhang : "Federated learning algorithm based on knowledge distillation", 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE), IEEE, 2020.
(29) Wu, Chuhan, et al. : "Fedkd: Communication efficient federated learning via knowledge distillation", arXiv preprint arXiv:2108.13323, 2021.
(30) Jeong, Eunjeong, et al. : "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data", arXiv preprint arXiv:1811.11479, 2018.
(31) Liu, Yen-Cheng, et al. : "Unbiased teacher for semi-supervised object detection", *arXiv preprint arXiv:2102.09480*, 2021.
(32) Chen, Meilin, et al. : "Learning domain adaptive object detection with probabilistic teacher", arXiv preprint arXiv:2206.06293, 2022.
(33) Simonyan, Karen, and Andrew Zisserman : "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014.
(34) Ren, Shaoqing, et al. : "Faster r-cnn: Towards real-time object detection with region proposal networks", Advances in NIPS, 2015.
(35) Deng, Jia, et al. : "Imagenet: A large-scale hierarchical image database", 2009

IEEE conference on CVPR, 2009.
(36) Xu, Chang-Dong, et al. : “Exploring categorical regularization for domain adaptive object detection”, Proceedings of the IEEE/CVF Conference on CVPR, 2020.