Source-Free Domain Adaptive Object Detection based on Pseudo-Supervised Mean Teacher

Xing Wei[1,2,3], Ting Bai[1*], Yan Zhai[1], Lei Chen[4], Hui Luo[2], Chong Zhao[2,5] and Yang Lu[1,3]

[1*]School of Computer and Information, Hefei University of Technology, emerald Road420, Hefei City, 230601, Anhui Province, China.
[2]Intelligent Manufacturing Institute of Hefei University of Technology, Hefei University of Technology, emerald Road420, Hefei City, 230051, Anhui Province, China.
[3]Engineering Research Center of Safety Critical Industrial Measurement and Control Technology, Ministry of Education, Hefei City, 230009, Anhui Province, China.
[4]Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei City, 230031, Anhui Province, China.
[5]Engineering Quality Education Center of Undergraduate School, Hefei University of Technology, Hefei City, 230601, Anhui Province, China.

*Corresponding author(s). E-mail(s): baiting@mail.hfut.edu.cn;

Abstract

Domain adaptive object detection refers to training a cross-domain object detector through a large number of labeled source domain datasets and unlabeled target domain datasets and learning the domain invariant features between two domains to reduce or eliminate the domain discrepancy. However, factors such as data privacy protection, limited storage space, and high labor costs often make many source domain labeled samples unavailable in real time situations. In this work, we propose a pseudo-supervised mean teacher model for source-free domain adaptive object detection that alternates between

generating pseudo-labels and fine-tuning the model, and utilizes a pixel-level distillation loss method and the weight regularization module for model adaptation. We use the mean teacher model to assist training to achieve object detection task in the source-free domain. Experiments are carried out on multiple datasets such as Cityscapes, Foggy Cityscapes, and SIM10K. Extensive experiments on multiple domain adaptation scenarios show that our method achieves better performance than the baseline (Faster R-CNN) and multiple state-of-the-art domain adaptation methods which require access to source domain data, demonstrating the effectiveness and robustness of the proposed method.

Keywords: Source-free Object Detection, Transfer Learning, Domain Adaptation

1 Introduction

As an essential and primary computer vision task, Object Detection needs to obtain the position results (rectangular contours and center point coordinates) and category results (tag category and probability) of different objects in the scene image simultaneously. The object detection model based on deep learning has been widely discussed and researched. Focusing on the detection accuracy and speed of the model, researchers have successively proposed RCNN, SPP-Net, YOLO[1] and Faster RCNN[2] etc. A typical example is the Faster RCNN, which structurally integrates the four modules of feature extraction, region recommendation, bounding box regression, and classification into one network, and the overall performance has been significantly improved.

Domain adaptive object detection is an essential direction in cross-domain object detection, formally proposed at the 2018 CVPR[2]. In which, the author draws on the idea of transfer learning, and learns the common domain invariant features of the source and target domain by using labeled data in the source domain and unlabeled data in the target domain. The key to the success of cross-domain tasks is to mine the value of training samples. The solutions can be roughly divided into two categories. The first category is based on the idea of sample generation [3–6], because the training process cannot obtain source domain labeled images, which is not applicable for traditional DA methods[7], but can use generative adversarial ideas to generate labeled samples of like-source domain or like-target domain. This kind of idea is suitable for single-object samples of cross-domain classification networks, but it still has high challenges for the application of multi-object images. The second category is based on pseudo-label generation[8, 9] to generate reliable pseudo-labels during training as much as possible. Meanwhile, researchers have proposed several solutions from different perspectives, such as distribution discrepancy, parallel learning, data reconstruction, hybrid mechanism optimization, etc.[10, 11], which effectively reconcile the contradiction between feature versatility and task specificity.

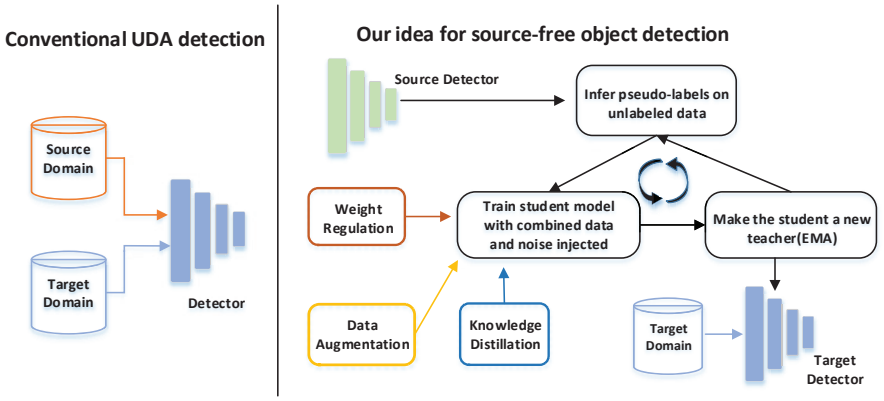

Fig. 1 The comparison between conventional Unsupervised Domain Adaptive(UDA) object detection and our idea of source-free domain adaptive object detection.

Fig.1 shows the comparison between our method and the conventional Unsupervised Domain Adaptive(UDA). The traditional UDA method relies on the training of source data, and needs to access the source domain data in the process of cross-domain learning. However, in some practical scenarios, or because of memory storage requirements, data sharing, privacy issues, or other dataset processing issues, the source domain data cannot be accessed, and the feature space cannot be extracted, which hinders training. Our object detection method no longer relies on source data but instead performs domain adaptation of the network by iteratively generating pseudo-labels, introducing a weight regularization module, and a knowledge distillation module.

At present, UDA has achieved remarkable success in many applications, such as object detection[12, 13], sentiment classification[14, 15], and natural language processing[16]. UDA Either adopts the idea of adversarial learning[4, 17] to encourage the learned source and target features to be indistinguishable from each other, or minimizes the difference in cross-domain distribution by matching the statistical moments of the distribution[12, 18], or uses pseudo-labeling techniques[19] by exploiting the similar distributions of the source and target domains. Due to privacy protection, limited storage space and other issues[20, 21], it may lead to inefficiency and impracticality in practical applications where they have sensitive information. Therefore, the source-free domain adaptation method has achieved a great breakthrough. At present, the application scenarios are mostly reflected in the classification scenarios[4, 8, 22], but there is still a gap in the object detection scene of source-free domain adaptive, and there is a huge room for improvement. In the latest source-free domain adaptive object detection research, [23] utilizes the intermediate domain to avoid the distribution imbalance problem, and [24] utilizes the style information of batch normalization stored in the pre-trained source model to convert the style features into the class source-like domain. In this paper, we propose a simple and effective method, which utilizes the initialization information of the source model and pseudo-label techniques, and

trains the target model with reliable pseudo-labels of target samples in a self-learning manner to reduce the distribution differences between domains, and the effectiveness of our method is shown in experimental comparisons.

To achieve this more challenging task of domain adaptive object detection, this paper proposes a"pseudo-supervised" mean teacher model for source-free domain adaptive object detection, which only needs a pre-trained object detection model in the source domain and an unlabeled target domain dataset, uses the pseudo-label as the "ground-truth" to complete the "pseudo-supervised" learning process of the student model for training and generates a detector suitable for the target domain. This paper introduces the MT algorithm framework[18, 25, 26], which uses two identical models to participate in the training together. The teacher model weight parameters are frozen, and the teacher model parameters are fine-tuned only by the student model's exponential moving average (EMA) method. To overcome the model deviation of the MT model in the cross-domain object detection method, this paper introduces the weight regularization module to learn the feature discrepancy between domains, realizes the feature space alignment of the above two models in pixel-level adaptation through the knowledge distillation method, and penalizes the inconsistency of the predictions of the two models to encourage model robustness.

The following are the main contributions of our work:

1. We provide a source-free domain adaptive object detection model and its training mechanism, that is, the pre-trained source domain model and target domain unlabeled data are trained to obtain a detector that achieves precise positioning and classification in the target domain.

2. We use the method of entropy minimization for threshold tuning, iteratively filter the generated pseudo-labels, reduce the entropy of the prediction results, complete the "pseudo-supervised" learning of the student model, and optimize the weight parameters of the mean teacher model.

3. We propose a weight regularization method to reduce the domain discrepancy between domains and add a distillation loss mechanism to adaptively realize the feature space alignment of the Mean Teacher model at the pixel level, optimize the network structure, and improve model performance.

The feasibility and effectiveness of the proposed model are verified in the dataset testing in Cityscapes, Pascal VOC and other fields, and the experimental analysis is used as the verification indicator. The organizational structure of this paper is as follows: Section 2 introduces related work, Section 3 presents our method and introduces the model, and Section 4 describes our experiments on different datasets and gives the experimental results of hyperparametric analysis and quantitative analysis. Section 5 gives our concluding conclusion.

Table 1 Summary of related references

Research Topics	Summary of Related References
Domain Adaptation Object Detection	[2][12][25][9][27][18][34][35]
Source-Free Domain classification network	[4][8][22]
Source-Free Domain Object Detection	[7][36][24][23][30]
Knowledge Distillation	[37][7][38][39][40][41]

2 Related work

2.1 Domain Adaptation and Object Detection Domain Adaptation

As a branch of transfer learning, domain adaptive methods are widely used in deep learning, such as object detection[2, 9, 25, 27], sentiment classification[14, 15], natural language processing[16] and semantic segmentation[28, 29]. In the early related work, the idea of adversarial generation was proposed, and the domain adversarial network framework DANN was introduced to achieve feature-level adaptation between domains. Jinhong et al.[18] combined feature-level adversarial training to learn domain-invariant representations to generate object-like images of different styles in cross-domain scenarios. Meanwhile, the discrepancy method was proposed. The domain adaptive methods can be summarized into two categories. One is the Generative Adversarial Net(GAN), Ganin et al.[30] proposed the idea of the gradient reversal layer (GRL) and introduced the domain adversarial network framework DANN to achieve feature-level adaptation between domains in early related work. Later, a series of methods were derived based on DANN (for example, conditional domain adversarial network(CDAN)[31], maximum classifier discrepancy(MCD)[32]). The second category is based on the discrepancy, such as the classic DDC[33] method for unsupervised DA, which uses a kernel function of MMD (Maximum mean discrepancy) to minimize the distance between the source domain and the target domain. The deep adaptive network DAN was developed based on DDC, using multi-core MMD (MK-MMD) to prove a better experimental effect. Table 1 summarizes the related references for each research topic in this paper. Deng et al.[18] first proposed a cross-domain object detection method and designed two domain adaptive modules to eliminate the domain discrepancy between the image-level and instance-level. A consistent regularization method is proposed to learn domain invariant features and complete end-to-end training to improve detection accuracy. Zheng et al.[12] proposed an adaptive feature method from coarse-grained to fine-grained, gradually and accurately aligning the depth features to achieve two-stage cross-domain object detection. Inspired by the MMD method, our work uses pixel-level distillation and a regularization loss strategy to minimize the source and target domain distribution distances, achieve feature alignment, and improve accuracy.

2.2 Knowledge Distillation

In recent years, knowledge distillation[7, 37] has succeeded in computer vision tasks, mainly due to its ability to extract feature knowledge and realize multiple model compression and acceleration techniques. In recent work, the knowledge distillation strategy can extract knowledge to guide the lightweight student model from the heavyweight teacher model, which has received wide attention and provides new ideas and challenges for computer vision tasks. The mean teacher model, which was first applied in the semi-supervised field, consists of two networks with the same structure (the teacher model and the student model). The student model uses labeled information for supervised training, and the teacher model updates the weights and passes parameters through the exponential moving average (EMA) method of the student model. Therefore, each sample prediction of the teacher model can be considered a combination of the current student model and the earlier version, enhancing its robustness and stability. Passalis et al.[38, 39] transferred knowledge by matching the probability distribution in the feature space to transfer teacher knowledge more efficiently. Seo et al.[40] proposed that the structural information between samples would be better than the characteristic structure of a single sample and introduced intermediate representation to optimize the performance of the student model. Xie et al.[41] applied a distillation strategy to semantic segmentation tasks, which is similar to our method. After the feature extraction, the category probability is assigned to the feature map pixel by pixel. The feature alignment is achieved at the pixel level to better balance experimental efficiency and accuracy.

2.3 Source-Free Domain Adaptation

In the traditional domain adaptive method, the source domain dataset contains a large amount of label information to perform the supervisory task, and the target domain is a dataset with no labels or only a few labels. The target domain is guided by supervised learning of the source domain knowledge. A huge amount of existing work uses source domain labeled datasets, which leads to a massive increase in the annotation of training datasets and the illegal access of private information. So unsupervised DA applications are proposed. The related survey found that the application scenarios of unsupervised DA are primarily applied to classification networks. Liang et al.[8] proposed a novel self-supervised pseudo-label method to enhance the representation learning of the target domain and introduced the idea of hypothesis transfer learning (HTL), based on "shared classifier parameters in the target domain and the source domain" as a whole. Kurmi et al.[4] used a conditional generative adversarial network combined with pre-trained classifier to provide a generative framework to solve the problem of the domain without source datasets. The classifier used pseudo samples with labeled information to adapt to the target domain. Yang et al.[22] proposed a cooperative conditional generative adversarial network (3C-GAN) without the source domain dataset,

Table 2 Glossary of symbols

Symbols	Description
D_t	The Object detection in target domain
x_i^S	Group of samples on student model with noise
x_i^T	Group of samples on teacher model with noise
$Y_{i,k}^T$	Corresponding pseudo-labels for the k-th iteration
G_x^S	Relational graph on student model
G_x^T	Relational graph on teacher model
L_{dist}	Distillation loss
$h_{optimal}$	The optimal threshold
L_{Re}	Region-level prediction loss
L_{Et}	Inter-graph prediction loss
L_{In}	Intra-graph prediction loss
L	The total loss

while the generator and prediction model are collaboratively enhanced, and the classification performance is improved.

As far as object detection is concerned, Xiong et al.[23] designed an approximation method based on the Law of Large Numbers to obtain the domain perturbation, thereby constructing a super target domain, and using the learning alignment from the super target domain to the target domain to avoid the imbalance problem in cross-domain object detection to a certain extent. Zhang et al.[24] proposed a new vision that utilizes the style information of batch normalization stored in a pre-trained source model, converted into source-like style features, to challenge the cross-domain object detection task. However, by analogy with the classification method, the biggest problem is that it cannot directly generate labeled source domain samples, and the idea of source-free joining adversarial generation[30] is challenging to realize. Since each picture in object detection contains multiple labels, it is difficult to use the pre-trained model to generate labeled source domain pictures from noise through the GAN network. Therefore, our work is different from that, instead of directly generating pseudo-samples in the source domain. We propose a source-free domain adaptive object detection based on pseudo-supervised mean teacher, use the mean-teacher model by dynamically updating the confidence threshold, filter the pseudo-labels generated by the teacher model, and "pseudo-supervised" train the student model, while aligning the predictions from the two models through the consistency regularization between teachers and students[7] to complete the object detection task. The research on domain adaptive object detection without source data is still in its infancy, and there is enormous room for performance improvement.

3 Proposed Method

The overall framework uses the pre-trained source model to initialize the training model(see Fig.2). Table 2 is the glossary of symbols proposed in the paper. First, we provide a set of unlabeled target domain data samples, defined as $D_t = \{x_i\}_{i=1}^{N_t}$, where x_i represents the ith picture in the target domain, and

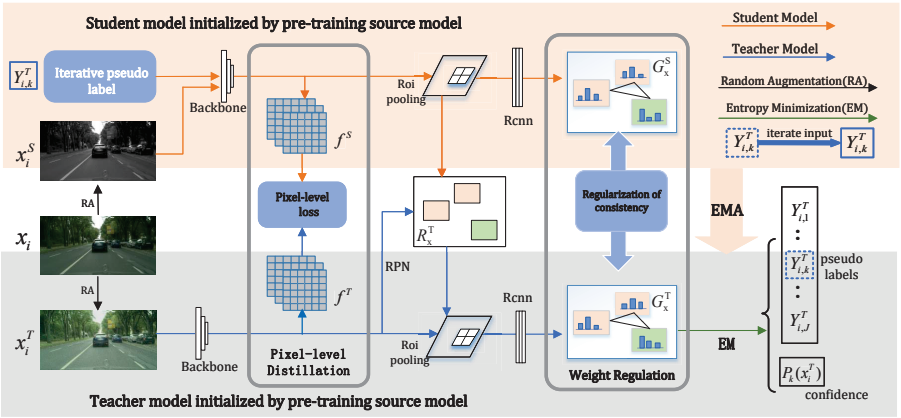

Fig. 2 Overview of the framework: The overall framework consists of the Teacher model (grey part) and the Student model (yellow part), using the pretrained source model to initialize the student network and the teacher network, and using threshold tuning for entropy minimization to generate iteratively updated pseudo-labels. After a pixel-level knowledge distillation module and a weight regularization module, the final training generates a detector suitable for the target domain.

N_t represents the number of pictures in the target domain. Two noisy pictures x_i^T and x_i^S are generated by Random Augmentation (RA) and sent to the teacher model and the student model respectively. The teacher model generates pseudo-labels through threshold tuning filtering of entropy minimization. During the following iterative training process of the student model, the x_i^S with the pseudo-label $Y_{i,k}^T$ is used as the input of the student model S_{model}, and the two models share the same region proposals generated by the teacher model's RPN to achieve feature extraction and generate feature maps f^T and f^S, where the backbone is based on the ResNet101 network. We use the Weight Regulation strategy to generate a graph relationship(G_x^S, G_x^T) to update the weight parameters of the student model of the current layer. The weights of the teacher network are frozen, and the weight parameters are only updated by the exponential moving average(EMA) on the student side. Unlike traditional DA, this training model does not provide the source samples with labeled information but uses a pre-trained source model as supervision. The task takes the pre-trained source model as the training benchmark. Finally, it generates the detector suitable for the target domain to improve the detection performance of the training model in the target domain.

3.1 Pixel-level Distillation Module

In the target domain, we take the image x_i as the target domain, generate two noisy pictures x_i^T and x_i^S by Random Augmentation (RA), input the backbone to realize feature extraction and generate feature maps $f^T$and f^S. We use the knowledge distillation mechanism to align the class probability of each pixel of the student network at the pixel level. The student model is guided by

distillation loss, and the weight parameters are updated by backpropagation to help the student network improve performance.

We define knowledge distillation as the task of assigning category labels to each pixel in the image and achieving feature alignment. The total number of categories is defined as C, and the RGB image I of size $W \times H \times 3$ is used as input. After the feature is extracted, the size of the feature map is calculated, which is defined as $W' \times H' \times N$, where N is the number of channels.

Here we treat this task as a collection of independent pixel labeling, and use knowledge distillation to align the class probabilities of each pixel generated by these two networks. The calculation formula of distillation loss is as follows:

$$L_{dist} = \frac{1}{W' \times H'} \cdot \sum_{i \in R} D_{KL}\left(p_i^S \parallel p_i^T\right) \tag{1}$$

Where p_i^S and p_i^T represent the category probability of the ith pixel in the feature map in the student model and the teacher model respectively, $R = \{1, 2, 3, ..., W' \times H'\}$ represents the collection of all pixels, and $D_{KL}(\cdot)$ refers to the Kullback-Leibler Divergence between the two class probabilities.

3.2 Entropy Minimization training

In object detection tasks, identifying label errors and characterizing label noise are important but easily overlooked tasks. In the training process, there may be a large number of negative samples mixed with positive samples. If false-positive samples are hardly removed, the learning of the student model will be misled, which could result in a failure to achieve good performance. Secondly, choosing a too high or too low confidence threshold will affect training performance due to label noise. Therefore, we need to choose an appropriate confidence threshold to help the model filter the generated pseudo-labels, leaving only the reliable parts to reduce the impact on training performance.

Information entropy is a measure of the degree of disorderliness of a system, measuring the uncertainty contained in the information, and the expression is defined as $H(x) = \sum_x p(x) \log \frac{1}{p(x)}$. We use information entropy to evaluate the quality of pseudo-labels. The lower the entropy value is, the lower the portion of false-positive samples is, and the higher the reliability value of the pseudo-label is. During the iterative training process of the entire dataset, a reasonable lower entropy value is continuously selected, the confidence threshold is dynamically updated, and reliable pseudo-labels are generated to participate in the training.

Based on this task, we introduce the entropy minimization method in the teacher network, use the input sample $D_t = \{x_i\}_{i=1}^{N_t}$ after data augmentation to generate x_i^T as the input of the teacher model T_{model}, which T_{model} is initialized by the pre-trained model parameters, and then generates pseudo-labels and calculates the corresponding confidence, the formula is as follows:

$$\left\{Y_{i,k}^T, P_k(x_i^T)\right\}_{i=1}^{N_t} = \left\{T_{model}(x_i^T | h, \Re)\right\}_{i=1}^{Nt} \tag{2}$$

where $Y_{i,k}^T$ represents the pseudo-label generated by training the teacher model at the kth iteration, x_i^T represents the unlabeled target domain samples of the input teacher model, and $\Re$ represents the weight parameter of the teacher model at the kth iteration. $P_k(x_i^T)$ represents the confidence calculated by the kth iteration training, which is output by the softmax of the classification branch, the pseudo-label $Y_{i,k}^T$ is determined by the argmax of the foreground class probability, if the value is above the confidence threshold h, the corresponding box is assigned as the class label with the largest score. Otherwise, it is defined as the background class. h represents the confidence threshold, in which the first training uses a given confidence threshold to generate pseudo-labels, and then iterative training takes the local minimum entropy value to obtain the optimal threshold h.

$$H(D_t) = -\frac{1}{N_t}\sum_{i}^{N_t}\left(\frac{1}{n_c}\sum_{c}^{n_c} P_k^c(x_i^T)\log(P_k^c(x_i^T))\right) \tag{3}$$

$$h_{optimal} = \arg\min(H(D_t)) \tag{4}$$

where $P_k^c(x_i^T)$ represents the confidence of a category at the kth iteration, n_c represents the total number of categories, and c represents the category.

Based on our model, the teacher outputs reliable pseudo-labels. In the next iterative training, the generated pseudo-labels are used as"ground-truth" to complete the "pseudo-supervised" learning of the student model. Finally, the weight parameters of the teacher model are fine-tuned through EMA, new pseudo-labels are generated by filtering the confidence threshold obtained in the previous iteration, and the final labels and detectors suitable for the target domain are generated iteratively.

3.3 Weight Regulation Module

In this module, we introduce a weight regularization method to generate two relational graphs (G_x^S, G_x^T) respectively, and propose three levels of regularization loss. After data augmentation processing of target domain images, we input two different target domain samples as the premise to pursue image-level consistency, guide the student model by backpropagation, and ensure the consistency of prediction between the teacher model and the student model to optimize our cross-domain object detection model. Under the condition of source-free domain data participating in training, the task of cross-domain object detection is accomplished.

3.3.1 Region-Level Consistency

Regional-level consistency is introduced to reduce local instance discrepancies, such as light intensity, random noise, and scale, to align regional-level predictions between the vertices of the teacher and student images that share the same spatial location.

For a given unlabeled target domain image, images x_i^T and x_i^S with domain noise are generated after random augmentation processing. Subsequently, the two pictures are fed into the two models as input terminals to generate feature maps f^T and f^S respectively. It should be noted that the student model needs to share the region proposals generated by the teacher's RPN network, which is represented by R_x^T here, to ensure that the region-level can be measured consistently by the mutual learning between the teacher model and the student model. When the region mapping f_r^T and f_r^S of each region are obtained, we can obtain the corresponding probability distributions $d_r^S = F_{RCNN}^S(f_r^S)$ and $d_r^T = F_{RCNN}^T(f_r^T)$ of each region, so that the entire detection results of the model are: $V_{x_t}^S = \left\{d_r^S\right\}$ and $V_{x_t}^T = \left\{d_r^T\right\}$.

We define the prediction of regional consistency as a measure of the distance between the prediction results of the teacher model and the student model. Firstly, perform region is preprocessed and the confidence threshold is set to filter out all low-confidence foreground regions and background regions. The average of the mean square error is adopted to calculate the region-level prediction loss of the two models:

$$L_{\mathrm{Re}} = \frac{1}{|R_x^T|} \cdot \sum_{r \in R_x^T} \left\| d_r^S - d_r^T \right\|_2^2 \tag{5}$$

where R_x^T represents the region proposals generated by the teacher's RPN network, d_r^S and d_r^T represent the detection probability of the teacher model and student model respectively for the region.

3.3.2 Inter-graph consistency

Regional consistency of the above-mentioned aims to achieve feature alignment between the regions of the two models. We hope not only to achieve feature alignment from a macro perspective but also to reduce the differences caused by the structure in the graph. Therefore, the idea of regularization of graph structure is born.

Firstly, the picture is input into the model and the relationship diagram $G_{x_i}^S = \left\{V_{x_i}^S, a_x^S\right\}$ is obtained through calculation, where $V_{x_i}^S$ represents the probability set of all regions, a_x^S represents the graph's affinity matrix obtained by the student model. The similarity $\left(a_x^S\right)_{r_i r_j}$ between every two regions in the figure is obtained, where $r_i, r_j \in R_x^S$, which represents two different regions in the figure. Region features are extracted through the ROI-pooling layer and represented by $f_{r_i}^S$, and the corresponding feature is converted into a fixed-size dimension. The calculation formula of cosine similarity is introduced to obtain the similarity between every two regions, and generate an affinity matrix of the student model.

$$\left(a_x^S\right)_{\mathrm{r}_i r_j} = f_{r_i}^S \cdot f_{r_j}^S / (\left\| f_{r_i}^S \right\|_2 \cdot \left\| f_{r_j}^S \right\|_2) \tag{6}$$

where $f_{r_i}^S$ and $f_{r_j}^S$ respectively represent the feature map with fixed dimensions obtained by the student model through the ROI-pooling layer processing, and $\left(a_x^S\right)_{r_i r_j}$ represents the calculated similarity between every two regions. The calculation method of the affinity matrix of the teacher model is the same as above, and two affinity matrices can be obtained through the above calculation. Therefore, the consistency difference between graphs can be defined as the mean square error between two affinity matrices, and the calculation formula is as follows:

$$L_{Et} = \frac{1}{|R_x^T|^2} \cdot \left\| a_x^S - a_x^T \right\|_2^2 \tag{7}$$

where a_x^S and a_x^T represent the affinity matrix of the student model and the teacher model respectively, avd R_x^T represents the region proposals generated by the teacher's RPN network.

3.3.3 Intra-graph consistency

In traditional domain adaptive methods, some of the work introduced pseudo-labels to participate in semi-supervised learning, aiming to provide more accurate guidance and improve model performance. Inspired by this work, we hope to apply the principle of consistency to guide the student model through the pseudo-labels generated by the teacher model. First, we use the initial pre-training source model as benchmark to learn knowledge and generate pseudo-labels $l_r = \arg\max_{k \in C} \left(d_{r_k}^T\right)$ for each region proposal $\mathrm{r}_x^T \in R^T$ on the teacher-side, where C represents the total number of categories in the dataset, k represents a specific category, and $d_{r_k}^T$ represents the probability of the predicting. Then, the supervision matrix M_x^T of $(|\mathrm{R}_\mathrm{x}^T| \times |\mathrm{R}_\mathrm{x}^T|)$ can be generated in the following way to determine whether every region belongs to the same category:

$$\left(M_x^T\right)_{i,j} = \begin{cases} 1 \text{ if } l_{r_j} = l_{r_i} \\ 0 \text{ otherwise} \end{cases} \tag{8}$$

Where l_{r_i} and l_{r_j} represent the pseudo-labels generated in a specific area, and $\left(M_x^T\right)_{i,j}$ represents the generated supervision matrix with i rows and j columns.

$$L_{In} = \frac{\sum\limits_{1 \le i,j \le \left|\mathrm{R}_\mathrm{x}^T\right|} \left(M_x^T\right)_{i,j} \cdot \left(1 - \left(a_x^S\right)_{r_i r_j}\right)}{MAX\left(1, \sum\limits_{1 \le i.j \le |R_x^T|} \left(M_x^T\right)_{i,j}\right)} \tag{9}$$

where $\left(a_x^S\right)_{r_i r_j}$ represents the calculated similarity between every two regions in the student model, $\left(M_x^T\right)_{i,j}$ represents the generated supervision matrix with ith rows and jth columns, and $1 \le i, j \le \left|\mathrm{R}_\mathrm{x}^T\right|$ represent each region in the teacher model.

Algorithm 1 Algorithm for source-free adaptation object detection task
Parameter: Database D_t represents the unlabeled target domain dataset. $G(\cdot)$ represents data random augmentation. θ_t represents the weight parameter of the teacher model, initialize the trade-off parameters α, weight smoothing parameter λ. Pre-train parameter $\Re$. $S_{\text{model}}, T_{\text{model}}$ represent the student model and teacher model respectively. $\left\{Y_{i,k}^T\right\}_{i=1}^{N_t}$ represents the pseudo-labels generated by the teacher at the *kth* iteration. $P_k(x_i^T)$ represents the confidence threshold.

Input: Unannotated data $D_t = \{x_i\}_{i=1}^{N_t}$
Output: Detector model parameter C, Predictions

Initialize the student model S_{model} and teacher model T_{model} with $\Re$
Freeze *autograde* for T_{model}
for epoch=1 to N **do**
  for i=1 to n **do**
    Generate samples via $x_i^T, x_i^S \leftarrow G(x_i)$ randomly
    Use x_i^S as the input of S_{model}, and x_i^T as input of T_{model}
    Generate R_x^T with $T_{\text{model}}{}'s$ RPN and shared with S_{model}
    Update S_{model} with Eq 11
    Update T_{model} from S_{model} with Eq 10
    if starting adaptation **then**
      Train S_{model} with pseudo-labels $\left\{Y_{i,k}^T\right\}_{i=1}^{N_t}$
      Update S_{model} via $\theta_s \leftarrow Adam\left(\nabla_{\theta_s}\left(\alpha L_{dist} + \beta\left(\mathrm{L_{Re}} + L_{Et} + L_{In}\right)\right), \theta_s, \alpha, \beta\right.$
      Update T_{model} via $\theta_t = \lambda\theta_{t-1} + (1-\lambda)\theta_s$, λ
    end if
  end for
  Generate pseudo-labels $\left\{Y_{i,k}^T\right\}_{i=1}^{N_t}$ for the next iteration
  Update $P_k(x_i^T)$ via Eq 3 and Eq 4
end for
Get trained model

3.4 Overall Objective Function

Based on the framework model proposed, we introduce weight regularization and knowledge distillation mechanisms to learn domain-invariant representations, and introduce entropy minimization to iteratively filter false pseudo-labels for "pseudo-supervised" learning. First, the weights of the teacher model are frozen, and there is only one source for the weight update of the teacher model, which is the exponential moving average of the student network. Therefore, the teacher model's network weight can be considered a combination of the current student model and the earlier version. The network parameter update of the teacher model is obtained by Eq 10:

$$\theta_t = \lambda\theta_{t-1} + (1-\lambda)\theta_s \tag{10}$$

where θ_{t-1} and θ_t represent the upper training of the teacher model and the training parameters of this layer respectively, θ_s represents the currently updated network parameters of the student model, and λ represents the weight smoothing parameter, which is set to 0.99.

Secondly, the invariant features of the domain are learned from the perspective of the student model, and the network parameters are optimized to reduce the model deviation, including one distillation loss and three regularization losses. One introduces a knowledge distillation strategy to achieve pixel-level feature alignment. The other one implements region-level feature alignment, inter-graph structure matching and intra-graph feature alignment for the relation diagrams generated in the Faster RCNN framework. The algorithm description is shown in Alg. 1. The overall training loss is expressed as follows:

$$L = \alpha L_{dist} + \beta \left(L_{\mathrm{Re}} + L_{Et} + L_{In} \right) \tag{11}$$

where α, β is the tuning parameter.

4 Experiments

4.1 Dataset and Experimental Settings

4.1.1 Experimental parameter settings

We use Faster RCNN as the fundamental network for object detection, which the experimental setup is consistent with [2]. The source domain data is only used in the pre-training step. The experimental setup follows Faster RCNN [2], and all source data participate in training. For the domain adaptation phase, the student and teacher models are first initialized by the pre-trained source domain model. In the experiment, batchsize is set to 1, and ResNet101 is added as the backbone of the network model to verify the model's performance. In all experiments, we use 4 GTX 1080 Ti with 11GB memory for training, the initial learning rate is set to 0.0001, the momentum is set to 0.9, the overall training loss α is set to 0.2, β is set to 0.999. The teacher model is the final model used for testing. We give the average precision AP and mAP of all categories in the dataset, where AP is defined as $AveragePrecision = \sum Precision/N\,(TotalImage)$ and mAP is defined as $MeanAveragePrecision = \sum AveragePrecision/N\,(Classes)$.

To verify the feasibility of the proposed method, we present the experimental results of ablation experiments and evaluate the experimental effects of the developed modules quantitatively.

4.1.2 Dataset Setting and Enhancement

In the experiment, we use Cityscapes[42], Foggy Cityscapes, SIM10K and other datasets to participate in the training, and verify the effectiveness of our method on three datasets with different styles. Fig.3 is a sample diagram of 3 types of data augmentation, which realize cross-domain object detection in three different scenarios. In the experiment, several basic data augmentation

Fig. 3 Sample diagram of dataset augmentation, the first row represents the original image of the data sample, and the following is the data augmentation image.

strategies are used to process the input samples of the target domain, including techniques such as color jitter, gaussian blur and grayscale. This method adds noise by randomly adjusting the sample parameters. The input is the same image, but the input in different epochs is constantly changing, that is, the input-output mapping is not fixed, which is equivalent to the consistent regularization of the same unlabeled image between different epochs. The model does not easily overfit the noise in the pseudo-labels in this case. At the same time, the student model learns from the images after data enhancement processing, which increases the difficulty of learning. It can learn more abundant representations through additional information, and improve the robustness of the model.

4.2 Comparison Results

Our experiments complete three sets of adaptive results. The "baseline" in the experiment means that no DA adjustment is made, and the Faster RCNN network without an alignment module is used to train the model on the source domain dataset to obtain the detection result in the target domain. the "oracle" represents the accuracy of training a model using the target domain image with label information and detecting the target domain image. We introduce DA Faster-RCNN[2], BDC-Faster[27], MAF[35] and SFOD[36] traditional domain adaptive methods for comparison in the task. Experiments examine the average

Table 3 *Cityscapes → Foggy Cityscapes adaptation* The average precision of all categories under different cross-domain object detection methods (mAP)

Methods	Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP
baseline	24.2	23.0	34.2	15.0	26.4	14.3	15.7	28.1	22.6
DA-Faster[2]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
BDC-Faster[27]	26.4	37.2	42.4	21.2	29.2	12.3	22.6	28.9	27.5
Selective-Faster[34]	33.5	38.0	48.5	16.5	39.0	23.3	28.0	33.6	33.8
MAF[35]	28.2	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
SFOD(SED)[36]	11.8	25.3	40.4	34.3	21.7	**34.5**	32.6	44.0	30.6
SFOD(Ideal)[36]	22.3	**44.0**	38.2	31.4	15.1	25.7	**34.6**	36.8	31.0
SMT[24]	**35.5**	12.1	44.4	**39.6**	**34.5**	33.0	21.6	**47.3**	**33.5**
Ours	34.3	30.1	**45.1**	30.6	28.8	28.7	25.1	41.5	33.0
Oracle	34.2	50.1	55.3	38.0	46.9	32.4	39.4	53.2	43.7

accuracy of these methods under various categories and compare them with our source-free domain adaptive method. Although the detection effect of our method is slightly lower than the object detection accuracy of partial domain adaptation, it has good adaptability and can well solve the defect problem of detection in the source-free domain.

4.2.1 Normal weather to Foggy weather

In this experiment, the Cityscapes[42] dataset is used as the source domain, we use it to pre-train a source model and provide it to our student model. Foggy Cityscapes[43] is an unlabeled target domain dataset, which contains pictures of outdoor street scenes from 50 different cities. There are 5000 high-quality pixel-level annotated images, including 2975 images in the training dataset, 500 images in the verification dataset, and 1525 images in the test dataset, including 19 categories. The Foggy Cityscapes dataset is based on the Cityscapes image, it is a synthetic foggy dataset that simulates real scenes. There are three levels of fog density representation, and the file name suffix will explicitly represent the density level. Although there is a one-to-one correspondence between the Cityscapes dataset and the Foggy Cityscapes dataset, our model does not use this correspondence during the training process, but randomly obtains picture information and sends it into the model to participate in training.

We obtained the object detection results in the adaptive process from Cityscapes to Foggy Cityscapes (Table 3), including the average precision of each category (AP) and the average precision of all categories (mAP). The experimental results show that the baseline accuracy is only 22.6%, and our model is 10.4% higher than the baseline. Due to the similarity between the source domain and the target domain, the optimization effect of the model is remarkable, and its training effect exceeds some traditional domain adaptation methods, which proves the effectiveness of the method.

Table 4 The average precision (AP) of the "Car" category in the cross-domain detection of the *SIM10K → Cityscapes adaptation*

Methods	AP on Car
baseline	33.1
DA-Faster[2]	38.5
SW-Detection[27]	40.1
MAF[35]	41.1
AT-Faster[9]	42.1
SOAP[23]	40.8
SFOD(SED)[36]	42.3
SFOD(Ideal)[36]	42.5
Ours	**42.8**
Oracle	56.9

4.2.2 Adaptation from Virtual to Real Images

In this experiment, the SIM10K dataset is used as the pre-training source model, and Cityscapes[42] is used as the unlabeled target domain dataset. The SIM10K dataset contains synthetic images in 10k computer game rendering. It contains 10,000 annotated images with the category "Car", so we only chose the category "Car" to participate in the experiment.

Table 4 compares the performance of cross-domain detection from the synthetic dataset to the real dataset. We only use the category "Car" to participate in the experiment. As can be seen from the table, our model's AP reaches 42.8%, which is 4.3% higher than DA-Faster, and 0.3% higher than the source-free object detection method SFOD. Therefore, experiments show that style transfer in virtual-reality can help the model better capture the discrepancies between graphs and improve the detection accuracy.

4.2.3 Adaptation from Real to Artistic Images

In this experiment, the Pascal VOC dataset is used as the pre-training source model, which contains 20 common real-world categories. We use Pascal VOC's 5011 training set (2501 training set, 2510 validation set) in the experiment, and the Watercolor2k[44] dataset is used as the target domain to be tested. It contains six categories of watercolor-style artistic pictures, totaling 2000 pictures. In the experiment, we only use the six categories shared by the two datasets for training.

As shown in Table 5, we report the detection results of the dataset under 6 categories and compare them. Our model is 5.7% higher than the baseline, better than the traditional domain adaptive object detection method, and the source-free object detection accuracy of SOAP, proving the feasibility and applicability of our method in different application scenarios.

4.3 Ablation experiment

We performed several ablation experiments to examine the contribution of each module to the performance of this object detector. Our model consists of 3 modules. While ensuring that all settings are exactly the same, we add each

Table 5 The average precision (mAP) of the cross-domain detection of the *Pascal VOC → Watercolor adaptation*.

Methods	Bike	bird	car	cat	dog	person	mAP
baseline	74.6	48.4	45.1	28.9	22.0	53.1	45.4
DA-Faster[2]	75.2	40.6	48.0	31.5	20.6	60.0	46.0
BDC-Faster[27]	68.6	48.3	47.2	26.5	21.7	60.5	45.5
SOAP[23]	**77.7**	43.2	40.1	**48.2**	**38.8**	55.4	50.6
Ours	76.5	**48.9**	**45.2**	43.2	36.2	**56.5**	**51.1**

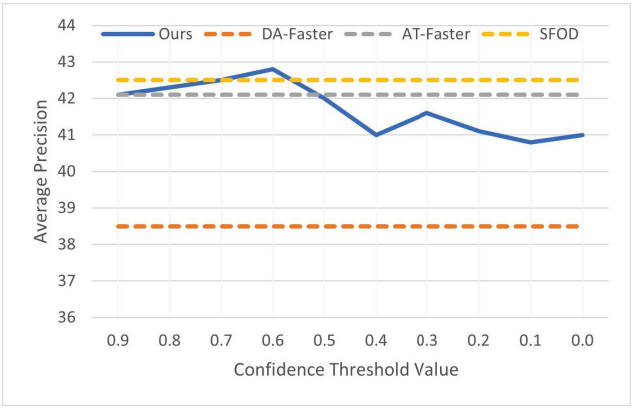

Fig. 4 Detection accuracy curves of pseudo-label generation under different confidence thresholds. In the SIM10K to Cityscapes domain adaptation experiment, the detection accuracy AP varies with the threshold

module in turn to study the effectiveness of each module. Table 6 shows the domain adaptive results of each module from Cityscapes to Foggy Cityscapes, where WR represents Weight Regulation, KD represents Knowledge Distillation, and EM represents Entropy Minimization. We can observe that each module can improve the performance of the model. The KD module implements the pixel-level class probability prediction of the feature map, combines the other two modules respectively, trains and tests their influence on the model. It can be seen from Table 6 that the baseline without any DA adjustment is only 22.6%, and the accuracy is increased to 4.4% after adding the EM module. Obviously, iteratively filtering pseudo-labels provides more accurate guidance for the mean teacher framework, prompting the generation of more valuable pseudo-labels in subsequent iterations and improving the overall performance of the model.

Fig.4 shows the detection accuracy curves of pseudo-label generation under different confidence thresholds. During iterative training on the dataset, we train the network with iteratively generated pseudo-labels. To verify the effect of entropy changes on accuracy, we use the confidence threshold as a controllable variable to conduct experiments on the SIM10K to Cityscapes datasets, and obtained the detection accuracy curves under different confidence thresholds, verifying the effect of entropy changes on the experimental results.

Table 6 Analysis of ablation experiments from Cityscapes to Foggy Cityscapes.

WR	EM	KD	mAP
			22.6
√			24.6
√		√	26.8
	√		27.0
	√	√	31.3
√	√	√	33.0

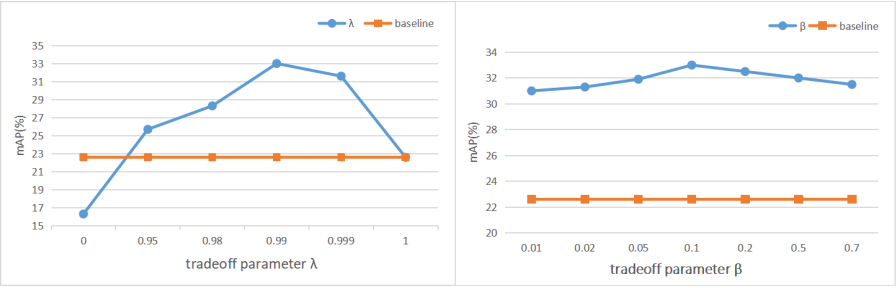

Fig. 5 Hyperparameter Analysis

4.4 Hyperparametric Analysis

To analyze the influence of parameters on the proposed method, we performed a parameter analysis in the Cityscapes to Foggy Cityscapes scene adaptation. Under the condition of ensuring that all settings are exactly the same, the experimental analysis results of parameters λ and β are given. For the parameter λ, we have done 5 sets of experiments from 0 to 1.0. It can be observed that the left graph of Fig.5 shows the line graph of the detection results changing with the parameter λ. Obviously, when the value of λ is 0, the parameters of the teacher model will be completely changed with the parameters of the student model and become a single model, which will affect the performance of our method. When λ is 0.99, we are most affected by the smoothing coefficient parameter λ, which is the best performance. When λ is close to 1, the teacher model is difficult to train, so we take a value of 0.99 in our experiments. Fig.5 on the right is the analysis result of the value of parameter β. The value range is set from 0.01 to 0.7, the overall trend is stable and the detection accuracy is good. Considering the experimental results of the parameter β, we take it as 0.1 in our experiment. We employ the same parameter settings in domain adaptation experiments in other scenarios, showing the robustness and effectiveness of our method.

4.5 Visualization of results

Fig.6 shows two groups of cross-domain object detection results in 3 different scenarios. The three rows of images from top to bottom represent Cityscapes to Foggy Cityscapes, SIM10K to Cityscapes and Pascal VOC to Watercolor. Among the three sets of experimental scenarios we applied, it can be observed

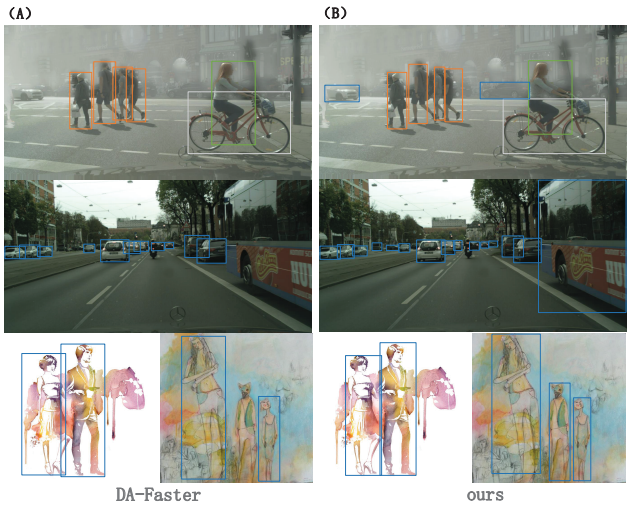

Fig. 6 Cross-domain object detection results of three different datasets, two sets of experiments(A and B) show the visualization results of the DA-Faster model and our model on the dataset under different scenarios, respectively.

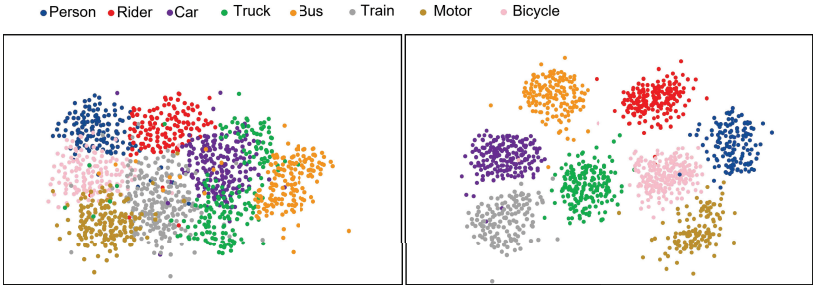

Fig. 7 Visualizations of feature map using t-SNE algorithm. The scene is from Cityscapes to Foggy Cityscapes. Left: the result obtained with the source-only model. Right: the result obtained with our proposed model. It can be seen that the classification distributions of the source and target domains can be well aligned, and the problem of uneven sample distribution and unclear boundaries is optimized.

that the detection accuracy of each category is improved, and the probability of misprediction is reduced. Compared with the domain adaptive method, the experimental results are comparable, and can effectively demonstrate the effectiveness of our method on object detection under source-free domain conditions.

Fig.7 shows the visualizations using the t-SNE algorithm. The scene of the experiment is from Cityscapes to Foggy Cityscapes, with a total of 8 different colors to label the categories. The left is the source-only model, the right is the training result of our method, it can be seen the distribution gap between the source domain and the target domain. After model training, the problem of uneven sample distribution and unclear boundaries is optimized.

5 Conclusion

In our work, in order to solve the problems of data privacy protection, limited storage space or high labor costs in the process of experimental processing and training. We propose a model that uses entropy minimization for threshold tuning, iteratively filters the generated pseudo-labels to make the model more influenced by the correct pseudo-labels during training, and uses the knowledge distillation mechanism and the weight regularization module to generate the suitable target domain detector. We use the mean average precision(mAP) as a measure to verify the performance of our method on different styles of datasets, and use experiments to prove the feasibility of our proposed method. Although our method outperforms many source-based DA models, it has to be admitted that in the case of no source domain dataset, the dataset of a single source domain scenario cannot fully cover the target domain features, and it is still very important to choose an appropriate source model. In future work, we will continue to study the application of DA methods in the source-free domain, hoping to combine all source and target domain datasets to improve the robustness of the model in the multi-source-free DA.

Acknowledgments. This work was supported by Joint Fund of Natural Science Foundation of Anhui Province in 2020 (2008085UD08), Anhui Provincial Key R&D Program (202004a05020004), Open fund of Intelligent Interconnected Systems Laboratory of Anhui Province (PA2021AKSK0107), Intelligent Networking and New Energy Vehicle Special Project of Intelligent Manufacturing Institute of HFUT (IMIWL2019003, IMIDC2019002).

Declarations

- Availability of data and materials

  Data openly available in a public repository.

Cityscapes: https://www.cityscapes-dataset.com/downloads/

Pascal VOC: http://host.robots.ox.ac.uk/pascal/VOC/

SIM10K: https://fcav.engin.umich.edu/projects/driving-in-the-matrix

References

[1] Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)

[2] Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3339–3348 (2018)

[3] Hou, Y., Zheng, L.: Visualizing adapted knowledge in domain transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13824–13833 (2021)

[4] Kurmi, V.K., Subramanian, V.K., Namboodiri, V.P.: Domain impression: A source data free domain adaptation method. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 615–625 (2021)

[5] Li, R., Jiao, Q., Cao, W., Wong, H.-S., Wu, S.: Model adaptation: Unsupervised domain adaptation without source data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9641–9650 (2020)

[6] Kim, Y., Cho, D., Han, K., Panda, P., Hong, S.: Domain adaptation without source data. IEEE Transactions on Artificial Intelligence **2**(6), 508–518 (2021)

[7] Xu, C.-D., Zhao, X.-R., Jin, X., Wei, X.-S.: Exploring categorical regularization for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11724–11733 (2020)

[8] Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International Conference on Machine Learning, pp. 6028–6039 (2020). PMLR

[9] He, Z., Zhang, L.: Domain adaptive object detection via asymmetric tri-way faster-rcnn. In: European Conference on Computer Vision, pp. 309–324 (2020). Springer

[10] Han, X.-F., Laga, H., Bennamoun, M.: Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. IEEE transactions on pattern analysis and machine intelligence **43**(5), 1578–1604 (2019)

[11] Ben-Nun, T., Hoefler, T.: Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. ACM Computing Surveys (CSUR) **52**(4), 1–43 (2019)

[12] Zheng, Y., Huang, D., Liu, S., Wang, Y.: Cross-domain object detection through coarse-to-fine feature adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13766–13775 (2020)

[13] Xie, R., Yu, F., Wang, J., Wang, Y., Zhang, L.: Multi-level domain

adaptive learning for cross-domain detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0 (2019)

[14] Sadr, H., Nazari Soleimandarabi, M.: Acnn-tl: attention-based convolutional neural network coupling with transfer learning and contextualized word representation for enhancing the performance of sentiment classification. The Journal of Supercomputing **78**(7), 10149–10175 (2022)

[15] Thakkar, A., Mungra, D., Agrawal, A., Chaudhari, K.: Improving the performance of sentiment analysis using enhanced preprocessing technique and artificial neural network. IEEE Transactions on Affective Computing (2022)

[16] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964 (2020)

[17] Dong, J., Cong, Y., Sun, G., Liu, Y., Xu, X.: Cscl: Critical semantic-consistent learning for unsupervised domain adaptation. In: European Conference on Computer Vision, pp. 745–762 (2020). Springer

[18] Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4091–4101 (2021)

[19] Wang, Q., Breckon, T.: Unsupervised domain adaptation via structured prediction based selective pseudo-labeling. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6243–6250 (2020)

[20] Wang, L., Xu, S., Wang, X., Zhu, Q.: Eavesdrop the composition proportion of training labels in federated learning. arXiv preprint arXiv:1910.06044 (2019)

[21] Wang, L., Xu, S., Wang, X., Zhu, Q.: Towards class imbalance in federated learning (2020)

[22] Yang, S., Wang, Y., van de Weijer, J., Herranz, L., Jui, S.: Unsupervised domain adaptation without source data by casting a bait. arXiv preprint arXiv:2010.12427 (2020)

[23] Xiong, L., Ye, M., Zhang, D., Gan, Y., Li, X., Zhu, Y.: Source data-free domain adaptation of object detector through domain-specific perturbation. International Journal of Intelligent Systems **36**(8), 3746–3766 (2021)

[24] Zhang, D., Ye, M., Xiong, L., Li, S., Li, X.: Source-style transferred mean

teacher for source-data free object detection. In: ACM Multimedia Asia, pp. 1–8 (2021)

[25] Liu, Y.-C., Ma, C.-Y., He, Z., Kuo, C.-W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480 (2021)

[26] He, R., Lee, W.S., Ng, H.T., Dahlmeier, D.: Adaptive semi-supervised learning for cross-domain sentiment classification. arXiv preprint arXiv:1809.00530 (2018)

[27] Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6956–6965 (2019)

[28] Chen, X., Pan, S., Chong, Y.: Unsupervised domain adaptation for remote sensing image semantic segmentation using region and category adaptive domain discriminator. IEEE Transactions on Geoscience and Remote Sensing (2022)

[29] Tasar, O., Happy, S., Tarabalka, Y., Alliez, P.: Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. IEEE Transactions on Geoscience and Remote Sensing **58**(10), 7178–7193 (2020)

[30] Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning, pp. 1180–1189 (2015). PMLR

[31] Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. arXiv preprint arXiv:1705.10667 (2017)

[32] Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)

[33] Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3723–3732 (2018)

[34] Zhu, X., Pang, J., Yang, C., Shi, J., Lin, D.: Adapting object detectors via selective cross-domain alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 687–696 (2019)

[35] He, Z., Zhang, L.: Multi-adversarial faster-rcnn for unrestricted object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6668–6677 (2019)

[36] Li, X., Chen, W., Xie, D., Yang, S., Yuan, P., Pu, S., Zhuang, Y.: A free lunch for unsupervised domain adaptive object detection without source data. arXiv preprint arXiv:2012.05400 (2020)

[37] Liu, Y., Shu, C., Wang, J., Shen, C.: Structured knowledge distillation for dense prediction. IEEE transactions on pattern analysis and machine intelligence (2020)

[38] Passalis, N., Tefas, A.: Learning deep representations with probabilistic knowledge transfer. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 268–284 (2018)

[39] Passalis, N., Tzelepi, M., Tefas, A.: Probabilistic knowledge transfer for lightweight deep representation learning. IEEE Transactions on Neural Networks and Learning Systems **32**(5), 2030–2039 (2020)

[40] Seo, H., Park, J., Oh, S., Bennis, M., Kim, S.-L.: Federated knowledge distillation. arXiv preprint arXiv:2011.02367 (2020)

[41] Xie, J., Shuai, B., Hu, J.-F., Lin, J., Zheng, W.-S.: Improving fast segmentation with teacher-student learning. arXiv preprint arXiv:1810.08476 (2018)

[42] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)

[43] Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**(9), 973–992 (2018)

[44] Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5001–5009 (2018)