Self-paced Supervision for Multi-Source Domain Adaptation

Zengmao Wang[1] , **Chaoyang Zhou**[1] , **Bo Du**[1*] and **Fengxiang He**[2]
[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Institute of Artificial Intelligence, Science and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China
[2]JD Explore Academy, JD.com Inc, China
{wangzengmao, zhouchaoyang, dubo}@whu.edu.cn, fengxiang.f.he@gmail.com

Abstract

Multi-source domain adaptation has attracted great attention in machine learning community. Most of these methods focus on weighting the predictions produced by the adaptation networks of different domains. Thus the domain shifts between certain of domains and target domain are not effectively relieved, resulting in that these domains are not fully exploited and even may have a negative influence on multi-source domain adaptation task. To address such challenge, we propose a multi-source domain adaptation method to gradually improve the adaptation ability of each source domain by producing more high-confident pseudo-labels with self-paced learning for conditional distribution alignment. The proposed method first trains several separate domain branch networks with single domains and an ensemble branch network with all domains. Then we obtain some high-confident pseudo-labels with the branch networks and learn the branch specific pseudo-labels with self-paced learning. Each branch network reduces the domain gap by aligning the conditional distribution with its branch specific pseudo-labels and the pseudo-labels provided by all branch networks. Experiments on Office31, Office-Home and DomainNet show that the proposed method outperforms the state-of-the-art methods.

1 Introduction

The large-scale number of labeled data has promoted the great success of deep learning in many applications, such as object detection and localization [Zhang *et al.*, 2021], medical diagnosis and semantic segmentation [Ouyang *et al.*, 2020]. However, the available of labeled data is usually very limited, since it is very expensive to label the large amount of unlabeled data. Unsupervised domain adaptation has been widely explored to address the scarce of labeled data in machine learning community. Generally, in an unsupervised domain adaptation scenario, it adapts the knowledge in the labeled source domain to the unlabeled target domain by reducing the domain gap between the target domain and source domain. In most unsupervised domain adaptation methods, they mainly focus on minimizing the distribution discrepancy between the target domain and source domain for knowledge adaptation [Jiang *et al.*, 2020; Zhong *et al.*, 2021], and these methods has achieved satisfactory performance with a single source domain by learning the domain-invariant representations. In fact, there always exist many source domains for a specific task in the real world applications, yet there is no effective approach to decide which domain can achieve the optimal adaptation for the task in target domain, since the domain-invariant structures between different domains are various.

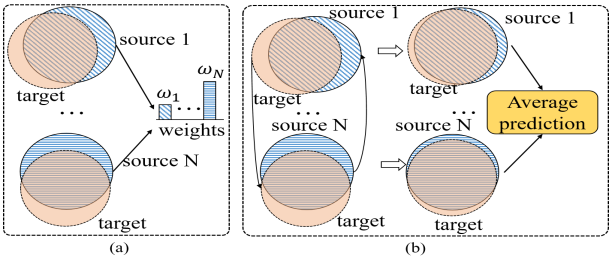


Figure 1: (a) Most multi-source domain adaptation methods attempt to weight the predictions of different domain models. (b) The proposed method attempt to improve adaptation ability of each source domain.

In recent years, to utilize the various knowledge from different domains effectively and guarantee that the model is more practical in the real-world applications, multi-source domain adaptation (MSDA) has attracted great attention in both the academic field and the industrial field [Zhao *et al.*, 2018; Wen *et al.*, 2020]. These methods mainly pay much attention to the importance of each domain by weighting the adaptation ability of different domains. They fail to reduce the domain shifts between certain source domains and target domain. Although the weights can balance the influence of these domains, the poor adaptation ability may still have negative influence on the target prediction. Meanwhile, since the domain shifts of these domains are not effectively reduced, the ability of these domains to assist the target domain may greatly decrease, resulting in that the multi-source domains are hard to further improve the adaptation performance on

*Corresponding Author

target domain. The difference between the proposed method and these methods is illustrated in Figure 1.

To address the above issues, this paper proposes a multi-source domain adaptation method by self-paced supervision, termed as SPS. In the proposed method, we design a deep adaptation network by gradually assigning more and more high-confident pseudo-labels in target domain for conditional distribution alignment with self-paced learning. The proposed architecture has several separate domain branch networks and a ensemble branch network. These branch networks are following a shared bottleneck network to extract the common structures. The domain branch networks are trained with single source domain and target domain respectively while the ensemble branch network is trained with all available domains. Generally, the ensemble branch network trained with all source domains and target domain can learn the discriminative features of different categories more effectively. Hence, we use a discrepancy loss to force the the consistent prediction between domain branches and ensemble branch to guarantee the adaption ability of different domains.

With the above branch networks, we assign pseudo-labels to samples whose predictions between the average of domain branch networks and the ensemble branch network are the same. Then each branch network is trained with self-paced approach to gradually produce more and more high-confident branch specific pseudo-labels. The adaptation ability of each branch network is further improved by aligning the conditional distribution between the source domain and the target domain with pseudo-labels, which includes the pseudo-labels produced by the branch networks and branch specific pseudo-labels. In this way, the adaptation ability of each branch network can be gradually improved with the common knowledge and domain specific knowledge.

We conduct experiments on Office31, Office-Home and DomainNet. Compared with state-of-the-art methods, the proposed method has achieved the best performance for all adaptation tasks on Office31 and Office-Home. It also achieves the best performance at most cases and second best for some adaptation tasks on DomainNet.

The main contributions of the paper can be summarized as:

- We propose a multi-source domain adaptation method to improve the domain adaptation ability of several separate branch networks by gradually producing more high-confident pseudo-labels with self-paced learning for conditional distribution alignment.
- Self-paced learning is successfully applied in multi-source domain adaptation to improve the adaptation ability of each domain. It can learn not only the ensemble knowledge of all domains but also the domain specific knowledge effectively.
- Extensive experiments show that the proposed method outperforms most of the state-of-the-art methods.

2 Related Works

In this section, according to our motivation, we will investigate the works on unsupervised domain adaptation with single source domain and multi-source domain.

2.1 Single Source Domain Adaptation

Due to the expensive cost for the labeling of large-scale unlabeled data, many unsupervised domain adaptation methods have been developed to learn the well adaptation models with single source domain. Maximum mean discrepancy(MMD) is a popular technique for domain adaptation task [Yan *et al.*, 2017; Wang *et al.*, 2017]. [Yan *et al.*, 2017] proposed to introduce class-specific auxiliary weights into the MMD for exploiting the class prior probability on source and target domains. The EM algorithm is adopted by alternating between assigning the pseudo-labels, estimating auxiliary weights and updating model parameters to align the conditional distribution. In recent years, the gradient reversal layer has been widely adopted to align the distribution between source domain and target domain [Zhang *et al.*, 2019]. These methods mainly focus on designing various discrepancy distance between source and target domain, and then minimize the discrepancy to optimize the feature extraction network.

2.2 Multiple Source Domain Adaptation

In many real-world applications, there exist several domains that can be adapted to the target domain. However, it is a challenge to decide which domain can obtain optimal adaptation for the target task. Multi-source domain adaptation utilizes all the available source domains simultaneously for adaptation tasks.

Compared with single source domain adaptation, the efficient MSDA approaches has been not fully developed [Zhao *et al.*, 2020]. [Zhao *et al.*, 2018] proposed a multi-source domain adaptation method with adversarial neural network to learn the feature representation. The invariant features of multi-source domains are obtained by optimizing task-adaptive generalization bounds. [Guo *et al.*, 2020] claimed that different measures can only provide specific estimates of domain similarities and each measure has its pathological cases. Therefore, they consider the mixture of several measures to minimize the distribution between target and source domains. [Zhu *et al.*, 2019] attempted to align the distribution between target and each source domain respectively with several branch networks. Then the discrepancy between different branch networks on target domain is minimized simultaneously. [Wang *et al.*, 2020] constructed knowledge graph on the prototypes of various domains to realize the information propagation within the semantic structures, and then a relation alignment loss is proposed to promote the feature of intra-class invariance and inter-class separately. [He *et al.*, 2021] used the pseudo-labels online generated by an ensemble model in target domain to update the multiple adaptation models. In this way, the performance of the multiple adaptation models are further improved. In [Zhao *et al.*, 2021], it proposed a multi-source adversarial domain aggregation network to make different adapted domains more closely aggregated with domain adversarial training.

These methods mainly focus on learning the adaptation network with the well adapted source domains and the influence of domains with poor adaptation is ignored, resulting in that the different knowledge of different domains are not effectively aggregated.

3 Methodology

In this section, we will introduce the proposed SPS in detail. Suppose there are N source domains with K classes and we define these source domains as $\{S_i\}_{i=1}^{N}$. Meanwhile, the target domain is denoted as T. In SPS, there are $N+1$ branch networks following a shared bottleneck network. For each branch network, it is designed with a feature extraction network and a classification network. The first N branch networks are domain branch networks, which are trained with one of the source domains and the target domain, i.e. the i^{th} domain branch network is trained with S_i and T. While the $N+1^{th}$ branch network is the ensemble branch network, which is trained with all source domain and the target domain. For the i^{th} branch network, we define the feature extraction network as F_i and the classification network as C_i. Then the i^{th} branch network can be represented as $F_i \circ C_i$ and all the branch networks in SPS can be represented as $\{F_i \circ C_i\}_{i=1}^{N+1}$. $\circ$ represents the composition of two functions. The shared bottleneck network is defined as F.

3.1 Pseudo Labeling

To improve the performance of unsupervised domain adaptation, pseudo-labels are usually adopted to align the distribution between source domain and target domain [Jiang *et al.*, 2020]. Generally, the high-confident pseudo-labels are selected with high prediction probability over a threshold[Zheng and Yang, 2021]. However, the wrong prediction may have high prediction probability while the correct prediction may have small prediction probability, since the domain gap exists and the adaptation model may be not very strong. In the proposed method, we select the high-confident labels adaptively based on the branch networks which are trained with the different domains. For example, the i^{th} domain branch network can be trained with the i^{th} source domain by

$$\mathcal{L}_{cls}^{S_i} = \min_{F_i, C_i} \sum_{x_j \in S_i} \ell((C_i \circ F_i)(F(x_j)), y_j^{S_i}) \tag{1}$$

where ℓ is the crossentropy loss, $y_j^{S_i}$ is the label of j^{th} sample in S_i. $\circ$ represents the composition of two functions, for example, $(g \circ f)(x)$ is equal to $g(f(x))$. For the ensemble branch network $C_{N+1} \circ F_{N+1}$, it is trained with all the source domain, which can be represented as

$$\mathcal{L}_{cls}^{S} = \min_{F_{N+1}, C_{N+1}} \sum_{i=1}^{N} \sum_{x_j \in S_i} \ell((C_{N+1} \circ F_{N+1})(F(x_j)), y_j^{S_i}) \tag{2}$$

Then with all the source domains, the proposed architecture can be trained with

$$\mathcal{L}_{cls} = \min_{F, \{C_i \circ F_i\}_{i=1}^{N+1}} \sum_{i=1}^{N} \mathcal{L}_{cls}^{S_i} + \mathcal{L}_{cls}^{S} \tag{3}$$

We should note that the optimal adaptation is that all the branch networks can predict the target domain precisely. However, the adaptation ability of each source domain is different due to the different distribution shifts between the multi-source domains and target domain. In the proposed method, the domain branch networks learn the invariant feature from single source domain and the target domain while the ensemble branch network learns the invariant feature from all source domains and the target domain. We treat the ensemble branch network as strong classifier, since it can guarantee that the invariant features are discriminative for the multi-source domains. To guarantee the adaptation ability of each source domain, we force the predictions of target domain from domain branch networks to be consistent with that from ensemble branch network and a discrepancy loss is adopted[Saito *et al.*, 2018]. We adopt the the absolute values of the difference between the probabilistic outputs of the ensemble branch network and the domain branch networks as discrepancy loss, and it can be represented as

$$\mathcal{L}_{dis} = \min_{F, \{C_i \circ F_i\}_{i=1}^{N+1}} \sum_{i=1}^{N} \sum_{x_j \in T} |p_j^{S_i} - p_j^{N+1}| \tag{4}$$

where $p_j^{S_i}$ is the probabilities of x_j in target domain from the i^{th} domain branch network, which can be formulated as $p_j^{S_i} = softmax((C_i \circ F_i)(F(x_j)))$. $softmax$ is a softmax function to normalize the outputs of the classification networks. p_j^{N+1} represents the probabilities of x_j in target domain from the ensemble branch network. Then the proposed adaptation network can be trained with

$$\min_{F, \{C_i \circ F_i\}_{i=1}^{N+1}} \mathcal{L}_{cls} + \lambda \mathcal{L}_{dis} \tag{5}$$

where λ is a trade-off parameter. Based on the proposed method, $N+1$ predictions can be obtained for each sample in the target domain. Since each source domain has different distributions, we first use the ensemble of domain branch networks to generate high-confident predictions, which can be represented as

$$p^e = \frac{p^1 + p^2 + ... + p^N}{N} \tag{6}$$

We define the pseudo-labels of target domain T as $\hat{y}^e$ based on p^e. Meanwhile, we define the predictions that obtained from ensemble branch network as $\hat{y}^{N+1}$. Then for a sample x_i in target domain, if its prediction label $\hat{y}_i^e$ and $\hat{y}_i^{N+1}$ are the same, we assign the prediction label to x_i. We denote these samples in T as T^E and their pseudo-labels as $\hat{y}^E$.

3.2 Self-paced Training

In the self-paced learning, it considers a weighted loss term for all samples and a general self-paced regularizer with respect to sample weights, which can be represented as

$$\min_{f, v \in [0,1]^n} \sum_{i=1}^{n} (v_i \ell(y_i, f(x_i)) + \Omega(v_i, \lambda) \tag{7}$$

where λ is the age parameter for controlling the learning space, and $\Omega(v, \lambda)$ represents the self-paced regularizer. The model parameter f and the latent weight v are alternatively optimized with gradually increasing age parameter. Then more samples can be adopted for training from easy to complex in a purely self-paced way.

In the proposed method, we hope each branch network can learn the branch specific samples to assign pseudo-labels. Hence, we train each branch network with a self-paced strategy. Since the samples in T^E are labeled with different branch networks and their pseudo-labels are very confident, the self-paced way are trained from easy to complex with the samples T/T^E. The self-paced training procedure for the i^{th} branch network can be represented as

$$\begin{aligned}\mathcal{L}^i_{sp} = &\min_{F,F_i,C_i,\hat{y},v^i} \sum_{x_j \in T^E} \ell(C_i \circ F_i(F(x_j)), \hat{y}^E_j) \\ &+ \sum_{x_j \in T/T^E} (v^i_j \ell(C_i \circ F_i(F(x_j)), \hat{y}_j) - \lambda^i v^i_j)\end{aligned} \tag{8}$$

where v^i_j denotes the weight of x_j in T/T^E with the i^{th} domain branch network. λ^i is the age parameter controlling the training scale in each iteration with respect to the i^{th} domain branch network. Based on the self-paced learning, we can obtain the easy training samples in target domain for a specific branch work. To be convenient, we denote the training samples and their pseudo-labels obtained by self-paced learning during the training of the i^{th} branch network as T^i and $\hat{y}^{T_i}$ respectively.

3.3 Distribution Alignment

The conditional distribution matching between target domain and source domain is very effective to improve the adaptation ability. In the proposed method, we propose to use the pseudo-labels to align the conditional distribution between target domain and each source domain respectively with MMD, which is a popular technique to align the distribution in domain adaptation. For the alignment between each source domain and target domain, we not only use the samples that are assigned with pseudo-labels based on the ensemble of branch networks but also the branch specific pseudo-labels obtained by the self-paced learning. In this way, the distribution alignment with MMD to optimize the i^{th} domain branch network can be represented as

$$\begin{aligned}&\mathcal{L}^i_{align} = \min_{F,F_i} \\ &|| \sum_{k=1}^{K} [\frac{1}{\sum_{y_j \in \hat{y}^E \cup \hat{y}^{T_i}} \delta(y_j, k)} \sum_{x_t \in T^i \cup T^E} \delta(\hat{y}^i_t, k) \phi(F_i \circ F(x_t)) \\ &- \frac{1}{\sum_{y_j \in y^{S_i}} \delta(y_j, k)} \sum_{x_s \in S_i} \delta(y^{S_i}_j, k) \phi(F_i \circ F(x_s))] \, ||^2_{\mathcal{H}}\end{aligned} \tag{9}$$

where $\delta(a, b)$ is a function to indicate whether a is equal to b. If a is equal to b, $\delta(a, b) = 1$, otherwise $\delta(a, b) = 0$. The same optimization is adopted to align the features of the target domain and source domain for the other domain branch networks.

We should note that the ensemble branch network is trained with all the source domains. Hence, the alignment in the $(N+1)^{th}$ branch network should use the pseudo-labels to align with each source domain respectively.

3.4 The Proposed Method

In the proposed SPS, we force the various distribution shifts between the source domains and target domain to be reduced respectively, and the invariant representations between each source domain and target domain are learned effectively. Generally, the samples around the classification boundaries are easy to be misclassified. In SPS, the discrepancy loss by measuring the absolute values of the difference between the probabilistic outputs of the ensemble branch network and the domain branches network is adopted. It can guarantee that all the branch networks can predict the target domain with similar predictions without considering the distribution shifts. To further improve the domain adaptation ability of all the source domains, we learn each domain branch network with self-paced strategy to gradually produce more high-confident pseudo-labels. Then the distribution alignment is adopted to training each branch network. Meanwhile, the entropy loss for the target domain is adopted to further improve the classification ability of each branch network, which can be represented as

$$\mathcal{L}_{ent} = \min_{F,\{C_i \circ F_i\}_{i=1}^{N+1}} \sum_{i=1}^{N+1} -\frac{1}{n_t} \sum_{x_t \in T} p^i_t log(p^i_t) \tag{10}$$

where p^i_t is the prediction probabilities of x_t obtained by the i^{th} branch network.

The total loss of SPS is to minimize the classification loss, self-paced loss, the entropy loss, discrepancy loss and the distribution alignment loss. We represent the total loss of SPS as

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{ent} + \mathcal{L}_{sp} + \lambda \mathcal{L}_{dis} + \beta \mathcal{L}_{align} \tag{11}$$

where λ and β are two trade-off parameters. For testing, we use the average of domain branch networks and the ensemble branch network for prediction. The final prediction of an image x_i can be represented as

$$p_i = \frac{1}{2} \left[\frac{(p^1_i + p^2_i + ... + p^N_i)}{N} + p^{N+1}_i \right] \tag{12}$$

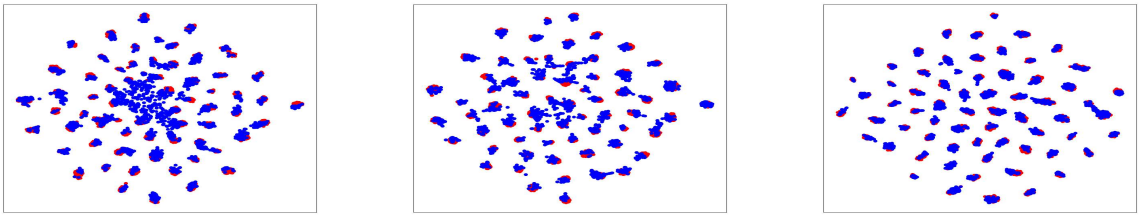
where p^k is the prediction probability obtained with $C^k \circ F^k$.

4 Experiments

In this section, we will verify the effectiveness of the proposed method with three popular datasets in domain adaptation, including Office-31, Office-Home and DomainNet[Zhu *et al.*, 2019; Li *et al.*, 2021]. For each dataset, one domain in it is treated as target domain while the other domains are treated as the source domains.

4.1 Implementation Details

To verify the effectiveness of the proposed method, several baseline and state-of-the-art methods are compared. These methods are set under three scenarios. 1) Single Best(SB), it shows the best result among the adaptation tasks with different source domain and the same target domain. 2) Source Combine(SC), it treats all source domains as a single domain to perform single source domain adaptation. 3) Multiple Source(MS), the methods are developed for multi-source



(a) Source Only (b) MSFAN (c) SPS(ours)

Figure 2: The t-SNE visualization on the task $\rightarrow$ *product* in Office-Home dataset.

Standards	Methods	Office31				Office-Home				
		$\rightarrow$ A	$\rightarrow$ D	$\rightarrow$ W	Avg	$\rightarrow$ Ar	$\rightarrow$ Pr	$\rightarrow$ Cl	$\rightarrow$ Rw	Avg
Single Best	Source only	62.5	99.3	96.7	86.2	65.3	79.7	49.6	75.4	67.5
	DAN(ICML'15)	66.7	99.5	96.8	87.7	68.2	80.3	56.5	75.9	70.2
	D-CORAL(ECCV'16)	65.3	99.7	98.0	87.7	67.0	80.3	53.6	76.3	69.3
	DANN(ICML'15)	68.2	99.4	96.8	88.1	67.9	80.4	55.9	75.8	70.0
	MCD(CVPR'18)	69.7	**100.0**	98.5	89.4	69.1	79.6	52.2	75.1	69.0
Source Combine	DAN(ICML'15)	67.6	99.6	97.8	88.3	68.5	79.0	59.4	82.5	72.4
	D-CORAL(ECCV'16)	67.1	99.3	98.0	88.1	68.1	79.5	58.6	82.7	72.2
	DANN(ICML'15)	67.6	99.7	98.1	88.5	68.4	79.5	59.1	82.7	72.4
	MCD(CVPR'18)	68.5	99.4	99.3	89.0	67.8	79.2	59.9	80.9	71.9
Multi-Source	MFSAN(AAAI'19)	72.7	99.5	98.5	90.2	72.1	80.3	62.0	81.8	74.1
	MDDA(AAAI'20)	56.2	99.2	97.1	84.2	66.7	79.5	62.3	79.6	71.0
	SImpAI(NIPS'20)	70.6	99.2	97.4	89.0	70.8	80.2	56.3	81.5	72.2
	MADAN(IJCV'21)	63.9	99.4	98.4	87.2	66.8	78.2	54.9	81.5	70.4
	SPS(ours)	**73.8**	**100.0**	**99.3**	**91.0**	**75.1**	**84.4**	**66.0**	**84.2**	**77.4**

Table 1: The accuracy(%) of different adaptation tasks on Office31 and Office-Home. The best performance is emphasized in bold.

domain adaptation. In the experiments, we compare the proposed method with several popular domain adaptation methods, i.e., DAN [Long *et al.*, 2015], D-CORAL [Sun and Saenko, 2016], DANN [Ganin and Lempitsky, 2015], and MCD [Saito *et al.*, 2018]. These methods are widely focused in the domain adaptation field with deep learning. For the multi-source domain adaptation methods, we compare SPS with MFSAN [Zhu *et al.*, 2019], MADAN [Zhao *et al.*, 2021], MDDA [Zhao *et al.*, 2020], SImpAI [Venkat and Kundu, 2020], M^{3}SDA [Peng *et al.*, 2019], T-SVDNet [Li *et al.*, 2021] and FPDA [Fu *et al.*, 2021], which are the state-of-the-art methods for multi-source domain adaptation tasks.

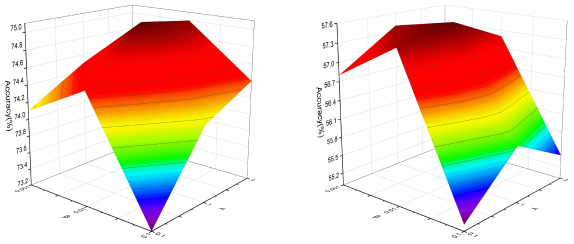
For the compared method, the Resnet-50 is adopted as the bottleneck network and we report their results according to [Venkat and Kundu, 2020] or the experiments are conducted with the settings in their original paper. In SPS, the same bottleneck network is adopted and its learning rate is set to be 10 times than that of the other layers. We set the learning rate as 0.001 while the optimizer and the learning schedule are set same with [Zhu *et al.*, 2019]. Meanwhile, there are two trade-off parameters λ and β. We set β as 0.01 in all experiments and set λ as 0.1 in Office31 while set it as 1 in Office-Home and DomainNet for practical application. Meanwhile, at the initialization, the deep network is pre-trained without domain alignment at the first some iterations,which is 2000 for Office31, 1000 for Office-Home and 10000 for DomainNet. We set a random seed 8 over 5 runs in the experiments, and the average results are reported.

Multi-Source Methods	$\rightarrow$ Clp	$\rightarrow$ Inf	$\rightarrow$ Pnt	$\rightarrow$ Qdr	$\rightarrow$ Rel	$\rightarrow$ Skt	Avg
M^{3}SDA(ICCV'19)	58.6	26	52.3	6.3	62.7	49.5	42.6
MDDA(AAAI'2020)	59.4	23.8	53.2	12.5	61.8	48.6	43.2
SImPAI$_{101}$(NIPS'20)	66.4	26.5	56.6	18.9	68.0	55.5	48.6
T-SVDNet(ICCV'21)	66.1	25.0	54.3	16.5	65.4	54.6	47.0
PFDA(CVPR'21)	64.5	29.2	57.6	17.2	67.2	55.1	48.5
SPS(ours)	**70.8**	24.6	55.2	**19.4**	67.5	**57.6**	**49.2**

Table 2: The accuracy(%) of different adaptation tasks on DomainNet. The best performance is emphasized in bold.

4.2 Experimental Results

We report the results of SPS and the comparison methods on Office31 and Office-Home in Table 1 while the results on DomainNet in Table 2. From the results in Table 1, we can directly observe that SPS has achieved the best performance. Compared with SB and SC, the multi-source domain adaptation methods achieve better performance at most cases. This demonstrates that the multi-source domain adaptation methods can improve the adaptation ability well. MFSAN and SImpAl are two state-of-the-art methods to reduce the domain gap with distribution alignment and implicit alignment respectively. MMD is adopted to minimize the distri-



(a) $\rightarrow Ar$ (b) $\rightarrow skt$

Figure 3: The results of SPS with different pair weights (λ, β) on the tasks $\rightarrow$ Ar in Office-Home dataset and $\rightarrow skt$ in DomainNet dataset.

bution shift between each source domain and target domain in MFSAN while the pseudo-labels are adopted in SImpAI by enforcing classifier agreement. However, we note that SImpAI performs a little worse than MFSAN. The reason may be that the pseudo-labels in SImpAI are not very confident and the classifier is influenced by the noise labels. In SPS, we provide different pseudo-labels to align with each source domain. The performance of SPS is significantly improved. This demonstrates that the proposed method can provide more high-confident pseudo labels with self-paced learning and improve the adaptation ability of each source well.

We also show the distribution alignment in feature space on the task $\rightarrow product$ in Figure 2, and it also directly shows that the distribution alignment of SPS is much better than source only and MSFAN. In Table 2, we also can observe that SPS performs much better at most cases than the state-of-the-art methods. Overall, SPS is very effective to align the target domain with multi-source domains and the self-paced supervision is very useful to produce much more high-confident pseudo-labels to improve the domain adaptation ability of multi-source domains.

4.3 Trade-off Parameters

There are two parameters λ and β. We report the results on the tasks $\rightarrow Ar$ and $\rightarrow Skt$ by choosing λ from the candidate set $\{0.1, 1, 2\}$ and β from the candidate set $\{0.001, 0.01, 0.1\}$ respectively in Figure 3. From Figure 3, we can observe that the performance of the proposed method changes obviously when β is set with different values. Hence, the proposed method is sensitive to β. This demonstrates that the alignment with self-paced supervision is very important to improve the performance of the proposed method. When β is fixed and λ is set with different values, we can observe that the results of the proposed method almost show the similar performance at most cases. This indicates that the predictions for the target domain with different source domain are similar, and the proposed method can align the conditional distribution between multi-source domains and target domain well. The results also support that the proposed method can improve the adaptation ability for each source domain effectively. Generally, we can set β as 0.01 and λ as 1 for practice.

4.4 Ablation Study

To verify the components in the proposed method, the ablation study is reported in Table 3 on task $\rightarrow Ar$ and $\rightarrow Skt$. Cross sign (✗) represents SPS is trained without the corresponding component while Check sign (✓) represents SPS is trained with the corresponding component. When any component in the proposed model is ignored, the performance degrades $0.3\% \sim 1.0\%$ in terms of average accuracy. This strongly illustrates that all components are essential in improving performance. Meanwhile, to demonstrate that the effectiveness of the pseudo-labels provided by the self-paced strategy, we show the results of each branch network in SPS on Office-Home dataset. S1, S2 and S3 represents the domain branch networks while Ensemble represents the ensemble branch network. Avg is the average results of the branch networks. Compared with single best and single worse, we can observe that the adaptation ability of each domain is significantly improved. Hence, the self-paced supervision can improve the adaptation ability of different sources well.

$\mathcal{L}_{sp}$	$\mathcal{L}_{ent}$	$\mathcal{L}_{align}$	$\mathcal{L}_{dis}$	$\rightarrow$ Ar	$\rightarrow$ Skt
✗	✓	✓	✓	74.6	57.3
✓	✗	✓	✓	74.2	56.6
✓	✓	✗	✓	74.3	57.0
✓	✓	✓	✗	74.3	57.3
✓	✓	✓	✓	**75.1**	**57.6**

Table 3: The ablation study on the task $\rightarrow Ar$ in Office-Home dataset and $\rightarrow Skt$ in DomainNet dataset.

Methods	$\rightarrow$ Rw	$\rightarrow$ Cl	$\rightarrow$ Pr	$\rightarrow$ Ar
Single Best	74.1	46.2	78.3	65.8
Single Worst	64.8	40.9	62.8	53.3
S1	83.5	65.5	84.1	74.6
S2	83.9	65.2	84.4	75.1
S3	83.2	65.7	84.6	75.0
Ensemble	83.5	66.2	84.3	75.3
Avg	83.5	65.6	84.3	75.0
SPS	84.2	66.0	84.4	75.1

Table 4: The accuracy(%) of the branch networks in SPS on Office-Home.

5 Conclusion

In this paper, a multi-source domain adaptation method is proposed with self-paced supervision to produce more high-confident pseudo-labels for domain alignment. The pseudo-labels are assigned adaptively based on the consistent predictions between the domain branch networks and the ensemble branch network. Meanwhile, much more pseudo-labels are produced based self-paced learning. Then the adaptation ability of each source domain can be improved with the conditional alignment. The extensive experiments show that the proposed method outperforms several state-of-the-art multi-source domain adaptation methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62006176, 62141112, 41871243, the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170, the Major Science and Technology Innovation 2030 "New Generation Artificial Intelligence" key project under Grant 2021ZD0111700, and the Natural Science Foundation of Hubei Province under Grants 2020CFB241. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[Fu *et al.*, 2021] Yangye Fu, Ming Zhang, Xing Xu, Zuo Cao, Chao Ma, Yanli Ji, Kai Zuo, and Huimin Lu. Partial feature selection and alignment for multi-source domain adaptation. In *IEEE CVPR*, pages 16654–16663, 2021.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015.

[Guo *et al.*, 2020] Han Guo, Ramakanth Pasunuru, and Mohit Bansal. Multi-source domain adaptation for text classification via distancenet-bandits. In *AAAI*, pages 7830–7838, 2020.

[He *et al.*, 2021] Jianzhong He, Xu Jia, Shuaijun Chen, and Jianzhuang Liu. Multi-source domain adaptation with collaborative learning for semantic segmentation. In *IEEE CVPR*, pages 11008–11017, 2021.

[Jiang *et al.*, 2020] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *ICML*, pages 4816–4827, 2020.

[Li *et al.*, 2021] Ruihuang Li, Xu Jia, Jianzhong He, Shuaijun Chen, and Qinghua Hu. T-svdnet: Exploring high-order prototypical correlations for multi-source domain adaptation. In *IEEE CVPR*, pages 9991–10000, 2021.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015.

[Ouyang *et al.*, 2020] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *ECCV*, pages 762–780, 2020.

[Peng *et al.*, 2019] X. Peng, Q. Bai, X. Xia, Z. Huang, and B. Wang. Moment matching for multi-source domain adaptation. In *IEEE CVPR*, 2019.

[Saito *et al.*, 2018] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE CVPR*, pages 3723–3732, 2018.

[Sun and Saenko, 2016] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450, 2016.

[Venkat and Kundu, 2020] Naveen Venkat and Jogendra Nath Kundu. Your classifier can secretly suffice multi-source domain adaptation. In *NeurIPS*, 2020.

[Wang *et al.*, 2017] Zengmao Wang, Bo Du, Lefei Zhang, Ruimin Hu, and Dacheng Tao. On gleaning knowledge from multiple domains for active learning. In *IJCAI*, pages 3013–3019, 2017.

[Wang *et al.*, 2020] Hang Wang, Minghao Xu, Bingbing Ni, and Wenjun Zhang. Learning to combine: Knowledge aggregation for multi-source domain adaptation. In *ECCV*, pages 727–744, 2020.

[Wen *et al.*, 2020] Junfeng Wen, Russell Greiner, and Dale Schuurmans. Domain aggregation networks for multi-source domain adaptation. In *ICML*, pages 10214–10224, 2020.

[Yan *et al.*, 2017] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *IEEE CVPR*, pages 2272–2281, 2017.

[Zhang *et al.*, 2019] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *ICML*, pages 7404–7413, 2019.

[Zhang *et al.*, 2021] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE TPAMI*, 2021.

[Zhao *et al.*, 2018] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *NeurIPS*, 31:8559–8570, 2018.

[Zhao *et al.*, 2020] Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. In *AAAI*, pages 12975–12983, 2020.

[Zhao *et al.*, 2021] Sicheng Zhao, Bo Li, Pengfei Xu, Xiangyu Yue, Guiguang Ding, and Kurt Keutzer. Madan: multi-source adversarial domain aggregation network for domain adaptation. *IJCV*, 129:2399–2424, 2021.

[Zheng and Yang, 2021] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, 129(4):1106–1120, 2021.

[Zhong *et al.*, 2021] Li Zhong, Zhen Fang, Feng Liu, Jie Lu, Bo Yuan, and Guangquan Zhang. How does the combined risk affect the performance of unsupervised domain adaptation approaches? In *AAAI*, pages 11079–11087, 2021.

[Zhu *et al.*, 2019] Yongchun Zhu, Fuzhen Zhuang, and Deqing Wang. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *AAAI*, pages 5989–5996, 2019.