EasyChair Preprint

№ 8677

Unsupervised Domain Adaptive Image Semantic Segmentation Based on Convolutional Fine-Grained Discriminant and Entropy Minimization

Xiaohao Zhao, Lihua Tian and Chen Li

August 12, 2022

Unsupervised domain adaptive image semantic segmentation based on convolutional fine-grained discriminant and entropy minimization

Xiaohao Zhao, Lihua Tian and Chen Li
Zhaoxh0099@163.com, lhtian@xjtu.edu.cn and lynnlc@126.com
School of Software Engineering, Xi'an Jiaotong University, Xi'an, China

Abstract. Deep convolutional neural networks have made considerable progress in the field of semantic segmentation of images. However, due to inter-domain differences, even modern networks cannot segment test datasets from different domains very well. To reduce and avoid costly annotation of the source domain training data, unsupervised domain adaptation attempts to provide efficient information transfer from the source domain with detailed annotation to the target domain without annotation. However, most existing methods attempt to align the source and target domains from a holistic view, ignoring the underlying class-level structure in the target domain, along with large noise and ambiguity at the class junctions. In this work, we innovatively employ a fine-grained unsupervised domain adaptation semantic segmentation method with increased entropy certainty, and guide the model for finer-grained feature alignment by adversarial learning, while increasing the pixel certainty near the category boundaries. Our approach is easy to implement and we have achieved good results on both the urban road scene datasets GTA5->Cityscapes and SYNTHIA->Cityscapes.

Keywords: Semantic Segmentation, Unsupervised Domain Adaptation, Class-Level Alignment.

1 Introduction

The goal of image semantic segmentation is to be able to assign the correct category labels to all pixels in an image, so it is suitable for complex image-based scene analysis, which is required for applications such as autonomous driving. The recently adopted convolutional neural networks (CNNs) offer various methods with the best performance for this task [1,2,3]. At the same time, there is a real problem that cannot be ignored, namely, all models depend on a well-trained dataset and its corresponding labels. If all images to be semantically segmented had a test dataset and its complete good label, this task would be a breeze. However, in practice there are two unavoidable problems: first, not all source domain datasets are easily available, e.g., some medical images are not publicly available per se or the number of samples is too small; second, the labels corresponding to the training data are too expensive to obtain, and these labels often require a large number of intensive pixel-level annotations that are done by expensive human labor, which is time-consuming and labor-intensive.

A potential solution to these two problems is to use simulated data, such as selected images from virtual scenes generated by computers or simulators [4,5,6] as source domain data, which has the advantage that a large number of source domain data samples can be easily obtained to solve the problem of insufficient data, while the labeling of these synthetic images is also done by computers,

which is not only very complete and detailed but also fast to save time and cost. However, the models trained with simulated images, no matter how perfect they perform in the simulated data environment, often fail to achieve the expected or satisfactory results once they are replaced by real scene images, and even the accuracy drops drastically. The reason for this degradation is that the two domains (source and target) are taken from different datasets, which have independent data distributions and large inter-domain differences. This phenomenon is commonly referred to as domain drift or domain shift [7]. The cross-domain task [8] needs to overcome this problem and undoubtedly faces a great challenge.

In order to achieve the cross-domain task to solve the domain drift problem, we adopt a domain adaptive approach, i.e., we reduce the domain drift problem to some extent by adjusting the feature distribution of the source domain (virtual image data) and the target domain (real image data) to reduce the distribution difference between them. Specifically we use the idea of adversarial learning to design a segmentation network and a discriminator network as the main framework of the model, and a segmentation network and a discriminator network are designed as the adversarial sides to distinguish the target sample from the source sample by training the domain discriminator, while the segmentation network tries to deceive the discriminator [8,9,10,11,12,13,14,15,16] into making wrong judgments by generating domain invariant features, and the discriminator is responsible for trying its best to identify whether the generated image comes from the source domain or the target domain. As the adversarial training escalates, the images generated by the segmentation network become more and more deceptive closer and closer to the target domain images, which means that the distribution difference between the source and target domains is gradually shrinking and the consistency between the two is getting stronger and stronger, finally achieving the purpose of solving the domain drift.

Despite the impressive progress in domain-adaptive semantic segmentation, most of the previous work has been devoted to the complete global feature distribution without paying much attention to the underlying structure between classes, and there is still a large amount of noise at the junction between classes. This is one of the reasons why the current domain-adapted semantic segmentation is not yet effective enough. As discussed in recent works [17,18], matching the global feature distribution alone does not guarantee that the expected error on the target domain is reduced and the class conditional distributions should be aligned as well. This implies that class-level alignment plays an equally important role in domain-adapted semantic segmentation. Therefore, it is necessary to satisfy both the matching alignment of global features of the image and also to solve the problem about the alignment between semantic classes. The difference between global alignment and class alignment is shown in Fig. 1, where blue indicates the source domain samples and red indicates the target domain samples. Figure (a) shows the results of global feature alignment, where the traditional discriminator can achieve good inter-domain discrimination, and the feature alignment between the two domains is basically good after domain adaptation, but some samples are still mixed together incorrectly. Figure (b) shows the semantic class level alignment by the fine-grained discriminator, which not only achieves the correct classification of the source and target domains but also distinguishes the different semantic classes in the source and target domains.

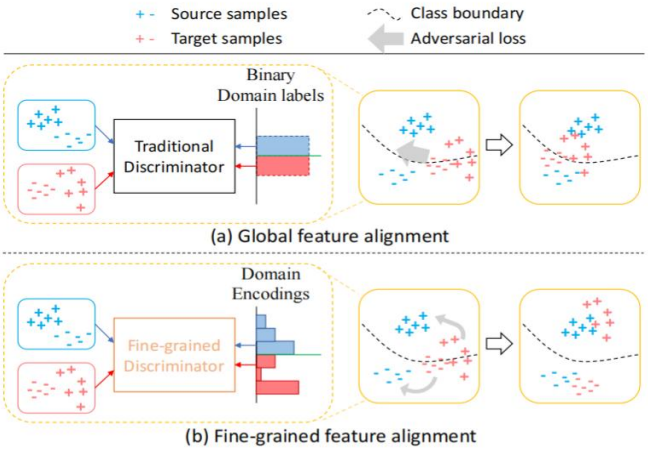

Fig. 1. The traditional approach and adversarial learning based on semantic category fine-grained are illustrated. Traditional adversarial learning pursues edge distribution alignment and ignores the inconsistency of semantic structure between domains. We propose to use fine-grained discriminators to achieve semantic class-level alignment.

There have been some inspiring works [11,19] to try to solve the problem of semantic class-level alignment.Chen et al[19] proposed to use multiple independent discriminators for class-level alignment, but since each discriminator is independent and the effective information of the learned features is not further integrated and optimized, the model may still fail to capture the relationship between individual semantic classes. Luo et al [11] introduced adaptive adversarial loss functions to roughly approximate class-level alignment by applying different weights to each region in the image. In practice, however, they do not explicitly incorporate semantic class information effectively into their approach, which may not facilitate class-level alignment.The work of Haoran Wang[20]et al. is illuminating in that although the labels of the target domain are inaccessible in the unsupervised domain adaptive task, they find that the model predictions on the target domain also contain by semantic class information, and demonstrate that it is possible to use predictions on both domains to supervise the discriminator: i.e., merging semantic class information into the learning process of the adversarial network allows the model to model the semantic inter-class structure, thus enabling fine-grained semantic class-level feature alignment.

We follow this idea by introducing semantic class information in the adversarial learning process and aligning features according to specific classes. We find that this operation also offers the possibility of fine-grained classification, where we integrate semantic class information into the discriminator and encourage it to judge and align at the fine-grained semantic class level by means of an objective optimization function. In addition, we observe a more general problem in the semantic segmentation results between, i.e., different semantic classes always have a large noise at the intersection, or even a segmentation error. As shown in Fig. 2, in the Source only graph it can be seen that the sidewalk is incorrectly segmented into the road, and in the AdaptSegNet graph on the right it is seen that the edge segmentation of the building class and the vegetation class is also not accurate enough.

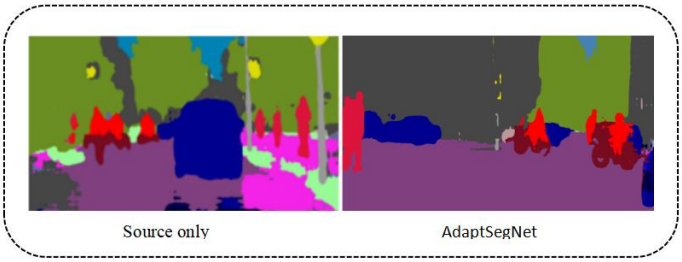

Fig. 2. It illustrates that the pixel certainty of previous domain adaptive semantic segmentation methods for boundary segmentation of semantic categories still needs to be improved.

The study[21]showed that there is some connection between pixel certainty and entropy and also demonstrated that if the model is trained only on the source domain then it tends to produce overconfident (i.e., low entropy) source-like image predictions and underconfident (i.e., high entropy) target-like image predictions. This phenomenon is shown in Fig. 2.

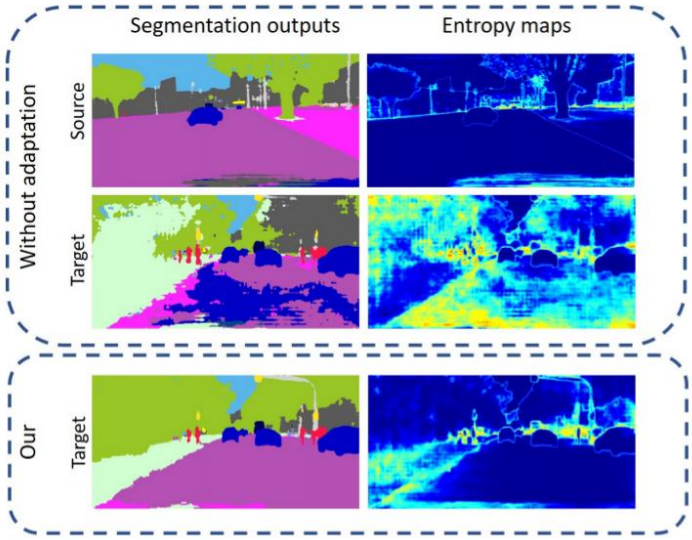

Fig. 3. Unsupervised domain adaptive semantic segmentation based on entropy minimization. The top two rows show the results of the model without domain adaptive training in the source and target domain scenarios. The bottom row shows the results of the model based on entropy minimization domain adaptive training in the same target domain scenario. The left and right columns show the visual semantic segmentation output and the corresponding predicted entropy mapping results, respectively.

On the one hand the predicted entropy map of the scene from the source domain looks like the edge detection results, with only high entropy activations at the boundaries of each semantic category. On the other hand the prediction for the target domain image is uncertain, which leads to a large amount of noise in the image segmentation result, which is represented as some columns of high entropy output in the entropy map. It is easy to argue that one possible way to reduce the difference in distribution between the source and target domains is to reduce the entropy of the pixels in the prediction of the target image, thus increasing the confidence level of the predicted pixels, especially those near the boundaries. Thus, we decided to try to increase the degree of certainty that the target predicted pixels belong to that semantic category while ensuring semantic category alignment. Experiments show that our model outperforms some of the more advanced unsupervised domain-adaptive semantic segmentation methods, and we also tested it on a commonly used publicly available road scene dataset.

- We propose a new mechanism that introduces a target prediction pixel entropy minimization strategy during fine-grained adversarial learning at the semantic class level for images to achieve better semantic segmentation, which results in clearer semantic segmentation outputs and correct recovery of larger blurred regions in images.
- We evaluate our method by comprehensive experiments. Significant progress is achieved on popular domain adaptive semantic segmentation tasks compared to other state-of-the-art methods, including GTA5->Cityscapes and SYNTHIA->Cityscapes.

2 Related Works

2.1 Semantic segmentation

Semantic segmentation is the task of predicting the unique semantic class label corresponding to each pixel of an input image, which can also be considered as a pixel-level classification task. With the development of deep convolutional neural networks, computer vision has made tremendous progress in this area. Many excellent models have emerged for the task of image semantic segmentation, which often perform well when sufficient training data and labels are available, but the models do not generalize well enough and show a sharp performance degradation when tested on other discrepant datasets. However, in real-world scenarios, it is not guaranteed that the model has a large amount of well-labeled training data under arbitrary conditions, especially in some unfamiliar and unknown scenarios, and it is also difficult to meet the consistency of the distribution between the source domain data when the model is trained and the target domain data when the model is actually working.

2.2 Domain Adaptation

Domain adaptation, as a representative approach to migration learning, aims to address various types of cross-domain tasks, i.e., for model performance degradation caused by different distributions of the source domain (training data) and the target domain (test data). In recent years, several studies have been proposed to address this problem in image classification tasks [17,12]. Inspired by the existence of risk-theoretic upper bounds in the target domain [22], some pioneering works have suggested feature alignment by optimizing the inter-domain difference measure between the two domains [23,24]. Recently, adversarial training, driven by GAN networks [25], has attracted attention for its leading ability to align features [12,13,19].

2.3 Unsupervised Domain Adaptive Semantic Segmentation

The semantic segmentation of images can in fact be seen as a more detailed pixel-level image classification problem, so theoretically semantic segmentation domain adaptation can fully draw on the existing related research results in the field of image classification. Because the labels of the target domain images are inaccessible, the challenge of unsupervised domain adaptation (UDA) [26,27] is enormous. The aim is to better perform the cross-domain task by transferring the effective information learned by the network model in the labeled source domain dataset to the unlabeled target domain images, thus improving the performance of the model on the target domain. Many UDA methods have been proposed to mitigate the domain drift problem. One common idea is to align the source and target domain distributions [28]. There are several ways to explore this idea in practice.CLAN [11] is an outstanding representative of this: it suggests applying different adversarial weights to different regions, but it does not directly and explicitly merge semantic class information into the model. AdaptSegNet [13] and Advent [21] mitigate domain drift. Another common direction to solve the problem is to align

the input pixels of source and target domain images by generating adversarial networks [10] or Fourier transforms [29]. In recent years, especially in the field of UDA semantic segmentation, pseudolabel refinement in a self-training framework has achieved quite good results. By iteratively training the network with progressively improved target pseudolabels, the performance of the model in the target domain can be further improved. Driven by this motivation, CBST [16] also achieved good results by setting appropriate thresholds for different semantic categories to improve the performance of model self-training.

3 Method

In this section, we propose the domain adaptive semantic segmentation algorithm with convolutional fine-grained discrimination and entropy minimization. To better introduce our model, we will start with the existing convolutional fine-grained adversarial learning and then describe how to introduce the entropy minimization approach and the process of fusing the two.

3.1 Semantic segmentation

The structure of the entire adversarial network can be divided into a generative network and a discriminator network. Traditional adversarial training seeks to align the feature distribution by confusing the binary discriminator; specifically, the generative network makes every effort to generate images that can deceive the discriminator in an attempt to make a wrong judgment; while the binary discriminator tries to correctly identify whether the input image comes from the source or target domain in an attempt to avoid being deceived by the generative network. . The limitation of the traditional binary discriminator is that it can only make simple judgments, i.e., whether the image has a higher probability of belonging to the source domain or to the target domain, which largely limits the segmentation accuracy of the model and falls far short of our requirements. In order to make the discriminator not only focus on the differentiation domain, our idea is to make the discriminator not limited to making simple binary judgments but also focus on the semantic class information, specifically, we use the convolutional fine-grained discriminator to optimize and upgrade the binary discriminator by expanding its original two output channels to K channels, and then encourage it to perform the semantic class level at a finer granularity. adversarial training. Where, K denotes the number of semantic classes to be segmented in the source and target datasets. By this design, the discriminator can fully exploit the role of adversarial learning, so that the discriminator can not only distinguish the domain to which the feature image belongs, but also further distinguish the specific class to which the feature belongs, e.g., whether it is the sky class or the building class in the source domain or the row human class or the vegetation class in the target domain. In other words, the prediction confidence of both source and target domains are represented as confidence distributions over different semantic classes, which enables the new convolutional fine-grained discriminator to model a more complex underlying structure between semantic classes, and thus better perform semantic class-level alignment. After this operation is done, the corresponding binary domain labels of the traditional discriminators are correspondingly overwhelmed and need to be converted into a general form, i.e., domain encoding, to contain semantic class information as well. The domain labels traditionally used for training binary discriminators are the source domain [1,0] and the target domain [0,1], respectively. In contrast, the domain encoding is represented by vectors [a;0] and [0;a] for the two domains, where a is the feature extracted in classifier C, represented by a k-dimensional vector; and 0 is an all-zero k-dimensional vector. When the discriminator believes that an image feature

belongs to the i-th class of the source domain with higher probability, it will set the i-th dimensional vector in [a;0] to 1 and the rest to 0. Similarly, when the discriminator makes a judgment that an image feature belongs to the j-th class of the target domain, it will set the j-th dimensional vector in [0;a] to 1. This achieves the transformation from the traditional binary discriminator to the convolutional fine-grained discriminator. transformation. This allows the discriminator to not only correctly distinguish domains during adversarial learning, but will also learn to model class structure and be able to portray the semantic class-to-class relationships in more detail.

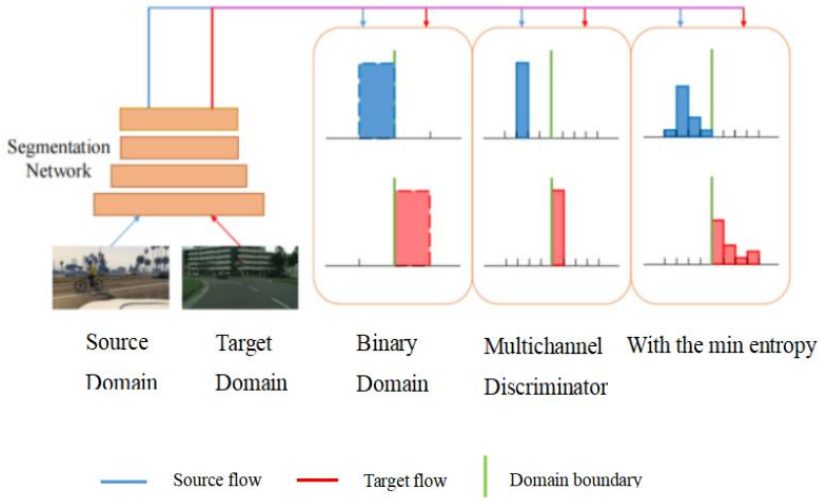

Fig. 4. Different strategies for generating domain labels, as shown in the figure, the traditional binary discriminator is only able to determine which domain the feature pixel comes from, while the updated multi-channel discriminator is able to perform a more fine-grained determination of semantic categories, and then when the entropy minimization strategy is added, it can be seen that the degree of certainty that each pixel belongs to its semantic category also increases significantly.

The network structure is shown in Fig. 4. We divide the whole network structure into three parts: the generative network or called semantic segmentation part, the adversarial learning network, and the entropy minimization network that increases the pixel certainty.

In which, the segmentation network G consists of feature extractor F and classifier C. Firstly, some images are randomly selected from the source and target domains and fed into the segmentation network. After the feature extraction and classification by the extractor and classifier, the feature maps of the source and target domains are obtained. On the one hand, the segmentation loss is calculated by comparing the source domain feature map with its corresponding source domain label, and the segmentation loss is continuously reduced in the process of adversarial training to help the segmentation network generate more deceptive images, i.e., the consistency of the source and target domain images in the generated images of the segmentation network is getting higher and higher, i.e., the gap between domains is decreasing.

On the other hand, after obtaining the feature maps, the semantic feature maps of the two domains are then input to the convolutional fine-grained discriminator and enter into the discriminator work part. At this point the discriminator uses the domain encoding processed from the sample prediction and tries to distinguish the domain information and class information of the features on the fine-grained semantic classes, calculates the probability that the feature pixel belongs to the domain and the class and makes a corresponding judgment, and calculates the loss of the discriminator by comparing the

number of correct and incorrect judgments of the discriminator. In the adversarial training, the parameters are continuously iterated and updated to reduce the loss function, thus improving the ability of the discriminator to make correct judgments to continue the confrontation with the generative network. Also, to increase the degree of confidence of a pixel in the semantic class to which it belongs and to reduce the large amount of noise present in image segmentation, especially near the junction between semantic classes. We transform the hard-to-express degree of pixel certainty into pixel entropy that can quantify the output, by adding the process of calculating the target pixel entropy minimization for the feature output, so that it can be learned adversarially with the discriminator to compensate for the wrong judgments made by the convolutional fine-grained discriminator. Therefore, we integrate the loss function of the convolutional fine-grained discriminator with the loss function of entropy minimization, and during the iterative process of adversarial learning, as the adversarial loss is continuously reduced, our segmented images will become more and more accurate, and also avoid much noise in the segmentation, being the boundary more clear.

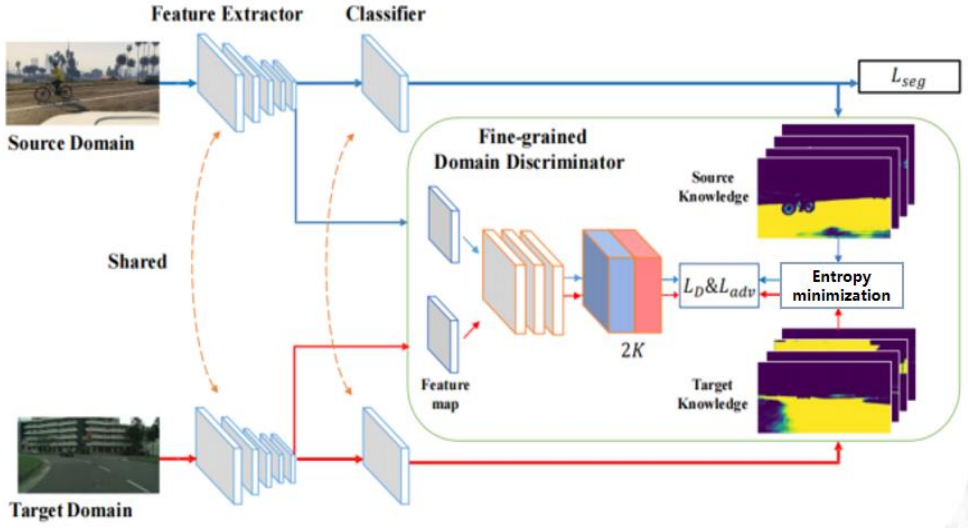

Fig. 5. The overall architecture of the network.

In order to better constrain the learning process of the network, we design three loss functions in the whole network structure, including: segmentation loss function L_{seg}, discriminator loss function L_D, and adversarial loss function L_{adv}.

The segmentation loss function is as follows.

$$L_{seg} = -\sum_{(h,w)\in n_s}\sum_{k=1}^{K} y_s^{(h,w)} \, log \, P_{x_s}^{(h,w)} \tag{1}$$

The main role of the segmentation loss function is to guide the segmentation network to generate semantic segmentation images with finer accuracy, which is achieved by training the reduction on the source domain dataset while training the reduced adversarial loss on the target domain dataset together with continuously updating the feature extractor and classifier. Where, is the source domain label, is the probability confidence that the segmentation network predicts that the source domain sample x belongs to the kth semantic class.

The loss function of the discriminator is as follows.

$$L_D = -\sum_{i=1}^{n_s}\sum_{k=1}^{K} a_{ik}^{(s)} \log P(d=0, c=k|f_i) - \sum_{j=1}^{n_t}\sum_{k=1}^{K} a_{jk}^{(t)} \log P(d=1, c=k|f_j) \tag{2}$$

During the training process, the discriminator not only tries to distinguish between source or target domains, but also will learn to model the semantic class structure. Where $a_{ik}^{(s)}$ and $a_{jk}^{(s)}$ is the kth class feature of sample i in the source domain and sample j in the target domain.
The adversarial loss function is as follows.

$$L_{adv} = -\sum_{j=1}^{n_t}\sum_{k=1}^{K} a_{jk}^{(t)} \log P(d=0, c=k|f_j) \tag{3}$$

The main role of L_{adv} is to deceive the discriminator and guide the generative network to generate feature maps with consistent distribution between domains, in other words to maximize the probability of target domain features being used as source domain features without compromising the relationship between features and semantic classes.

3.2 Entropy minimization

In unsupervised domain adaptive semantic segmentation, the target domain is unlabeled, and there is no way to use the labels to compute the segmentation loss function to guide the model training as in the source domain. So we propose to use a constrained model to make it produce high confidence predictions. Here we do not use high confidence pseudo-labeling based on the lack of memory capacity of the graphics card, because end-to-end training is not worth the memory occupied by pseudo-labeling. Instead, we generate semantic segmentation images with a high confidence level by minimizing the prediction pixel entropy. Specifically, we use Shannon entropy [30] to accomplish this task. Given a target domain input image x_t, the entropy mapping $E_{x_t} \in [0,1]^{H\times W}$ consists of independent pixel-level entropies normalized to the range [0,1] as follows.

$$E_{x_t}^{(h,w)} = -\frac{-1}{\log(K)}\sum_{k=1}^{K} P_{x_t}^{(h,w,k)} \log P_{x_t}^{(h,w,k)} \tag{4}$$

Then, the entropy loss can be defined as the sum of all pixelated normalized entropies as follows.

$$L_{ent}(x_t) = \sum_{h,w} E_{x_t}^{(h,w)} \tag{5}$$

Similarly, in the training process, we jointly optimize the cross-entropy loss of the supervised segmentation of the source domain samples and the unsupervised entropy loss of the target domain samples. The loss function of minimum entropy can be obtained and expressed as:

$$\min_{G} \frac{1}{|x_s|}\sum_{x_s} L_{seg}(x_s, y_s) + \frac{\lambda}{|x_t|}\sum_{x_t} L_{ent}(x_t) \tag{6}$$

3.3 Fusion and Adversarial Learning

By carefully observing and studying the prediction results of semantic segmentation, we found that if only the adversarial learning of convolutional fine-grained discriminator is performed, although the modeling of semantic inter-class relations can be improved, the confidence of the pixels in the classes to which they belong still needs to be improved, and intuitively the semantic segmentation pixels in the target domain still have a large ambiguous part, especially at the locations where classes intersect with each other. Therefore, we introduce the entropy minimization of the predicted pixels on top of the convolutional fine-grained discriminator. The specific measures we take are as follows: first, to find the right time for the entropy minimization to be added, here we add it in the adversarial network, and the fine-grained discriminator piece by piece in the adversarial learning to continuously constrain the segmentation network to generate feature images with higher segmentation accuracy. We do not recommend adding this operation after the classifier in the segmentation network, as it has been found

experimentally that the effect of entropy minimization in improving pixel confidence will be greatly reduced if the adversarial learning process is lost. In addition, since the adversarial network contains both the convolutional fine-grained discriminator and the entropy minimization, we also modify the adversarial loss. Then the adversarial loss function of the whole network becomes:

$$L_{adv} = -\sum_{j=1}^{n_t}\sum_{k=1}^{K} a_{jk}^{(t)} \log P(d=0, c=k|f_j) + \lambda \sum_{h,w} E_{x_t}^{(h,w)} \quad (7)$$

where is the weight of the sum of all pixel-wise normalized entropies.

In the training process, we then jointly optimize the supervised segmentation loss of the source domain data samples and the unsupervised adversarial loss of the target domain data samples. The final optimization problem is formulated as follows.

$$\min_{G} L_{seg} + \lambda_{adv} L_{adv} \quad (8)$$

In this way, the entropy minimization of the predicted pixels can well compensate for the lack of pixel confidence in the fine-grained discriminator, and together with the convolutional fine-grained discriminator, through adversarial learning, continuously promote the improvement of the network segmentation image accuracy, and finally reach the effect that the model is trained in the source domain and can also achieve good performance in the target domain.

4 Experiments

4.1 Datasets

We performed a comprehensive evaluation of our proposed method on two popular unsupervised domain adaptive semantic segmentation datasets, GTA5->Cityscapes and SYNTHIA->Cityscapes.

Cityscapes Cityscapes[31] is a large-scale dataset for autonomous driving model training, focusing on some road scenes of urban life, with a high diversity of videos and images sampled from different urban centers, as well as in multiple seasons. set contains 2975 images, the validation set 500 images and the test set 1525 images. Following standard protocols [10,8,13], we use 2975 images from the Cityscapes training set as the unlabeled target domain training set and evaluate our model on 500 images from the validation set.

SYNTHIA SYNTHIA[5] is a large synthetic dataset of images obtained from scene renderings of virtual cities. We selected its subset SYNTHIA-RAND-CITYSCAPES, which shares 16 semantic classes with Cityscapes, as the source domain. In total, 9400 images from the SYNTHIA dataset were used as the source domain training data for this task.

GTA5 GTA5[4] is another synthetic dataset that shares 19 semantic classes with Cityscapes. The dataset was rendered from the modern computer game Grand Theft Auto V, which has labels fully compatible with Cityscapes. 24,966 images of urban scenes were collected and used as source training data.

4.2 Network Architecture

We use Deeplab-V2[2] as the basic semantic segmentation architecture, and apply a void space pyramidal pooling (ASPP) on the feature output of the last layer in order to better capture the scene context. The sampling rate is fixed to {6,12,18,24}, similar to the ASPPL model in [2]. We perform experiments on the base deep CNN architecture ResNet-101[32]. After [2], we modify the step size and expansion rate of the last layer to produce denser feature maps and larger perceptual fields. To further

improve the performance of ResNet-101, we also adapt the multilevel outputs from the conv4 and conv5 features[13].

4.3 Implementation details

We use the PyTorch[33] implementation. For a fair comparison, we used DeeplabV2, as the segmentation base network. All models are pre-trained on ImageNet[34]. For the convolutional fine-grained discriminator, we used a simple structure consisting of 3 convolutional layers with {256,128,2K} channels, 3 convolutional kernels, and a step size of 1. Each convolutional layer except the last one is followed by a Leaky-ReLU[35] parameter with a value of 0.2. To train the segmentation network, we use a stochastic gradient descent (SGD) optimizer, where the momentum is 0.9 and the weights decay to. The learning rate was initially set and decreased with a poly learning rate of power of 0.9. The discriminator is trained using the Adam optimization algorithm, = 0.9, = 0.99, with an initial learning rate of. The same poly learning rate strategy was used. It was set to 0.001. Regarding the training process, the network is first trained on the source data for 20k iterations and then fine-tuned for 40k iterations using our framework. The batch size is 2. One of them is the source image and one is the target image. Some data enhancement methods are used, including random flips and color changes, to prevent overfitting.

4.4 Experimental results

Our approach is tested with respect to domain adaptation on both datasets, and the experimental results show that our algorithm achieves excellent results. The experiments use the mean intersection-to-merge ratio (mIoU) as an evaluation metric, which is the most important evaluation metric in semantic segmentation tasks. Our algorithm model achieves 49.5 mIoU in GTA5->Cityscapes experiments and 45.4 in SYNTHIA->Cityscapes experiments.

Table 1. GTAV-to-Cityscapes results.

Method	Road	Side	Building	Wall	Fence	Pole	Light	Sign	Vegetation	terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
Source Only	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.6	25.3	36.0	36.6
Cycada	79.1	33.1	77.9	23.4	17.3	32.1	33.3	31.8	81.5	26.7	69.0	62.8	14.7	74.5	20.9	25.6	6.9	18.8	20.4	39.5
AdaptSeg	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
SIBAN	88.5	35.4	79.5	26.3	24.3	28.5	32.5	18.3	81.2	40.0	76.5	58.1	25.8	82.6	30.3	34.4	3.4	21.6	21.5	42.6
CLAN	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
CBST	86.8	46.7	76.9	26.3	24.8	**42.0**	**46.0**	**38.6**	80.7	15.7	48.0	57.3	27.9	78.2	24.5	49.6	17.7	25.5	45.1	45.2
DISE	**91.5**	47.5	82.5	31.3	25.6	33.0	33.7	25.8	82.7	28.8	82.7	**62.4**	30.8	85.2	27.7	34.5	6.4	25.2	24.4	45.4
ADVENT	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	**38.5**	44.5	1.7	31.6	32.4	45.5
PyCDA	90.5	36.3	84.4	32.4	28.7	34.6	36.4	31.5	**86.8**	37.9	78.5	62.3	21.5	85.6	27.9	34.8	**18.0**	22.9	**49.3**	47.4
Ours	90.8	**49.8**	**85.0**	**39.5**	**28.7**	33.3	35.5	18.1	86.7	**39.7**	**85.6**	61.1	**35.9**	**86.7**	31.8	**49.9**	0.0	**35.5**	46.8	**49.5**

As can be seen from Table 1, all domain adaptive methods significantly outperform Source Only methods, i.e., models trained on images of synthetic scenes are directly applied to images of real scenes.

This shows that domain adaptive methods are necessary for semantic segmentation tasks with different feature distributions. Comparing some current more advanced unsupervised domain adaptation methods, the model designed in this paper is able to achieve optimal results with a mIoU of 49.5, which is a significant improvement of 12.9 over the Source Only model trained on the source domain. firstly, for the baseline method AdaptSegNet, which uses a traditional unsupervised domain adaptation method based on adversarial discriminations from The disadvantage of AdaptSegNet is that matching the global distribution of source and target domains may lead to some classes whose distributions in the feature space are already matched being disrupted after migration instead, resulting in some classes not performing as well as the Source Only model, i.e., a negative migration phenomenon. Specifically, for the classes "fence", "pole" and "bike" in Table 1, the performance of AdaptSegNet method on these three classes is even The performance of the AdaptSegNet method on these three classes is not even as good as that of the model without the domain adaptation method (i.e., Source Only). In contrast, the algorithm model proposed in this paper aligns the joint distribution of classes in the source and target domains at the class level, so that it can handle each class well and significantly outperforms the baseline method AdaptSegNe by 8.1 mIoU. More specifically, the algorithm in this study outperforms the Source Only model for almost all classes, i.e., there is no negative migration phenomenon. It should be noted that the simulator-generated dataset GTAV is taken from the in-game city scenes, and for the "train" category, its effect on the model is not considered for the time being because of the low stability of the training samples in this category in the source and target domains. Then, the algorithm of this study outperforms the unsupervised domain-based adaptive method CBST by 4.3 mIoU, and outperforms ADVENT and PyCDA by 4 mIoU and 2.1 mIoU, respectively. Overall, the algorithm model proposed in this paper outperforms other models, and the effectiveness of the model is verified by rich comparison tests.

Table 2. SYNTHIA-to-Cityscapes results.

Method	Road	SideWalk	Building	Wall	Fence	Pole	Light	Sign	Vegetation	Sky	Person	Rider	Car	Bus	Motor	Bike	mIoU
Source only	55.6	23.8	74.6	**9.2**	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5
SIBAN	82.5	24.0	79.4	-	-	-	16.5	12.7	79.2	82.8	**58.3**	18.0	79.3	25.3	17.6	25.9	-
AdaptSegNet	84.3	**42.7**	77.5	-	-	-	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	-
CLAN	**88.5**	35.4	79.5	-	-	-	**32.5**	**18.3**	81.2	76.5	58.1	**25.8**	82.6	34.4	21.6	21.5	-
AdaptPatch	84.5	38.0	78.6	8.7	**0.6**	26.0	3.9	11.1	75.5	**84.6**	53.5	21.6	71.4	32.6	19.3	31.7	40.0
ADVENT	62.4	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2
Ours	84.3	40.5	**82.2**	8.6	0.2	**31.3**	18.4	18.2	**85.5**	83.4	54.5	18.7	**85.1**	**47.8**	**22.6**	**44.8**	**45.4**

It is obvious from Table 2 that all unsupervised image semantic segmentation algorithms using domain adaptation outperform Source only methods, which means that domain adaptation methods play an important role in reducing the domain gap between the source and target domains and improving the performance of the model on the target domain. It can be seen that the algorithm designed in this paper even improves the mIoU by 11.9 compared to Source only. Because algorithms SIBAN, AdaptSegNet and CLAN have individual semantic classes that are not correctly identified and segmented, their mIoU

results are not calculated and further compared. In addition, compared to algorithms AdaptPatch and ADVENT, the algorithms in this paper improved 5.4 mIoU and 4.2 mIoU, respectively.

Table 3. Ablation experiments.

GTAV →Cityscapes				
Source Only	AdaptSegNet	Convolutional fine-grained discriminator	entropy minimization	mIou
√				36.6
	√			41.4
	√	√		48.4
	√	√	√	49.5

The results of the ablation experiments are shown in Table 3. The results of the semantic segmentation model trained on the source domain without domain adaptation and tested directly on the target domain, i.e., the Source Only method, are shown first. The model without domain adaptation method shows a relatively poor result of 36.6 mIoU on the target domain data aggregation. Then the results of the unsupervised domain adaptation method based on traditional adversarial discrimination for semantic segmentation, i.e., the AdaptSegNet method, are shown. The global alignment of the traditional adversarial discriminant-based method is significantly improved compared to the unsupervised domain adaptive model, with a result of 41.4 mIoU. After that, the discriminant network is modified based on the AdaptSegNet method, i.e., the original traditional binary discriminator is transformed into a multi-category convolutional fine-grained discriminator. The binary discriminator can only distinguish the source domain or target domain, and further make the judgment that the features may belong to a semantic category in the source or target domain, so that each semantic category can be well aligned. Then, on top of this, the target prediction pixel entropy minimization method is introduced, and the entropy loss function and adversarial training are used to increase the certainty of the predicted pixels, especially the pixels near the category boundaries, by calculating the entropy map of the target domain prediction image, so that the boundary features of the semantic categories are clearer, and the final model is improved again by 1.1 mIoU, and finally the mIoU of the algorithm model in this paper reaches 49.5 on the validation set of the target domain data.

The results of some randomly selected visualized adaptive semantic segmentation are shown in Fig. 6. It is obvious from the figure that the visualization results of the algorithm model in this paper are closer to the real image semantic labels than the Source Only model without domain adaptation, which can not only identify some rare categories, such as "pole" and "street light" It can not only identify some rare categories, such as "pole" and "street light", but also has no significant noise for the intersection boundary of different semantic categories.

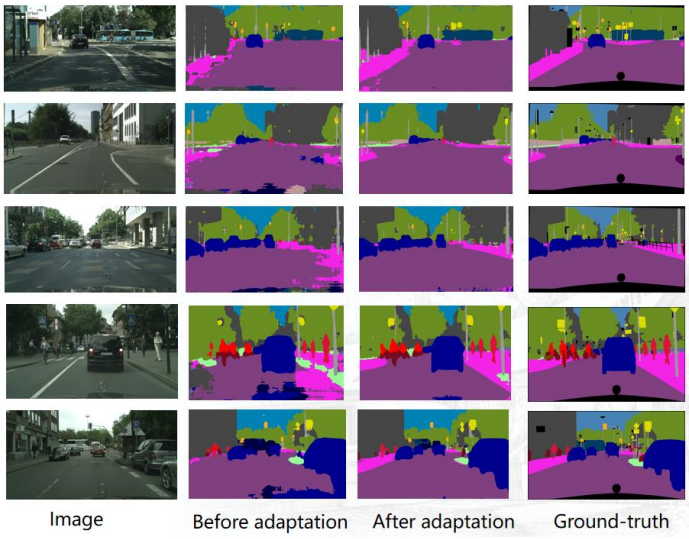

Fig. 6. GTA5->Cityscapes Semantic segmentation results

The comparison focuses on the traffic sign class within the white box in the target domain image. For the segmentation results of the model without any domain adaptation processing (Source Only), the traffic sign classes can be correctly segmented, although the global segmentation results are poor. This phenomenon indicates that some classes are initially aligned in the distribution of source and target domain data features even though they are not processed by any domain adaptation method. The adversarial discrimination-based domain adaptive method AdaptSegNet, as the baseline method in this paper, uses a global-level alignment of the distribution of the output features of the source and target domains after the semantic segmentation network, and although the overall segmentation effect of the adversarial learning-based domain adaptive method is better than that of the model without any domain adaptation, the segmentation result for traffic signs is poor, even inferior to that of the model without any domain adaptive model. This is because the global alignment strategy favors some common classes with a large percentage of pixels and tends to make conservative predictions for the features. This results in some uncommon features being predicted to other common classes, causing a negative migration phenomenon for those uncommon features, even though these classes are well aligned in the initial state.

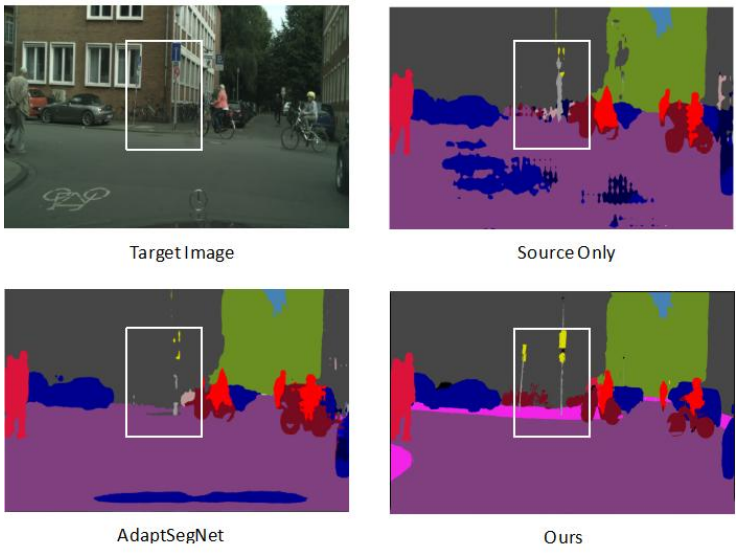

Fig. 7. Comparison of several algorithms

5 Conclusion

In this work, we solve the task of unsupervised domain adaptive semantic segmentation, and propose a convolutional fine-grained discriminant and entropy minimization algorithm. Specifically, the discriminant used in traditional confrontational training can only judge whether features come from the source domain or the target domain, which seriously damages the identification between semantic categories. Different from the traditional binary discriminator, the convolutional fine-grained discriminator expands the channel and keeps consistent with the number of semantic categories in the datasets of the target domain, so it can not only distinguish the source domain or the target domain, but also further make a judgment that the feature belongs to a class in the source domain or a class in the target domain. In addition, the entropy of the target pixel is calculated and reduced by adversarial training to increase the determination of the predicted pixel, especially the position of the junction between different semantic categories in the image. Finally, the effectiveness of the model in this task was verified by a large number of comparative experiments, and the contribution of each module to the model was verified by detailed ablation experiments. Our model achieves good results from two challenging synthetic datasets to real datasets.

References

1. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015. 1, 2, 5, 6.
2. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2018. 1, 5, 6 .
3. Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Macro-micro adversarial network for human parsing. In European Conference on Computer Vision, pages424–440. Springer, 2018. 1,2.
4. Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In European Conference on Computer Vision, pages 102–118. Springer, 2016. 1, 2, 6.
5. German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3234–3243, 2016. 1, 6.
6. Yawei Luo, Tao Guan, Hailong Pan, Yuesong Wang, and Junqing Yu. Accurate localization for mobile device using a multi-planar city model. In ICPR, 2016. 1.

7. Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference, 90(2):227–244, 2000. 1.
8. Hoffffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation (2016).
9. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive fast rcnn for object detection in the wild. In: Computer Vision and Pattern Recognition (CVPR) (2018).
10. Hoffffman, J., Tzeng, E., Park, T., Jun-Yan Zhu, a.P.I., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle consistent adversarial domain adaptation. In: International Conference on Machine Learning (ICML) (2018).
11. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019).
12. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classififier discrepancy for unsupervised domain adaptation. arXiv preprint arXiv:1712.02560 (2017).
13. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018).
14. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: The IEEE International Conference on Computer Vision (ICCV). vol. 2, p. 6 (Oct 2017).
15. Zhang, Y., Qiu, Z., Yao, T., Liu, D., Mei, T.: Fully convolutional adaptation networks for semantic segmentation. CoRR abs/1804.08286 (2018).
16. Zou, Y., Yu, Z., Kumar, B.V., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the Europan Conference on Computer Vision (ECCV). pp. 289–305 (2018).
17. Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T., Huang, J.: Progressive feature alignment for unsupervised domain adaptation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019).
18. Kumar, A., Sattigeri, P., Wadhawan, K., Karlinsky, L., Feris, R., Freeman, B., Wornell, G.: Co-regularized alignment for unsupervised domain adaptation. In:Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 9345–9356. Curran Associates, Inc. (2018),http://papers.nips.cc/paper/8146-co-regularized-alignment-for-unsupervised-domain-adaptation.pdf.
19. Chen, Y., Chen, W., Chen, Y., Tsai, B., Wang, Y.F., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2011–2020 (2017).
20. Haoran Wang, Tong Shen, Wei Zhang, Lingyu Duan, and Tao Mei. Classes matter: A fifine-grained adversarial approach to cross-domain semantic segmentation. In The European Conference on Computer Vision (ECCV), August 2020. 5, 6, 17, 18.
21. Vu, T.H., Jain, H., Bucher, M., Cord, M., P´erez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR (2019).
22. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine Learning 79(1), 151– 175 (May 2010). https://doi.org/10.1007/s10994-009-5152-4, https://doi.org/ 10.1007/s10994-009-5152-4.

23. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37. pp. 97–105. ICML'15, JMLR.org (2015), http://dl.acm.org/citation.cfm?id=3045118.3045130.
24. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: Hua, G., J´egou, H. (eds.) Computer Vision – ECCV 2016 Workshops. pp. 443– 450. Springer International Publishing, Cham (2016).
25. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair,S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. pp. 2672–2680. NIPS'14, MIT Press, Cambridge, MA, USA (2014), http://dl. acm.org/citation.cfm?id=2969033.2969125.
26. Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks, 22(2):199–210, 2011. 2.
27. Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. IEEE signal processing magazine, 32(3):53–69, 2015. 2.
28. Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Proceedings of International Conference on Machine Learning (ICML), pages 1180–1189, 2015. 2.
29. Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4085–4095, 2020. 2, 7, 8.
30. C. E. Shannon. A mathematical theory of communication. Bell system technical journal, 1948. 3.
31. 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: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).
32. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016. 5.
33. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alch´e-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp.8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/ 9015-pytorch-an-imperative-style-high-performance-deep-learning-library. Pdf.
34. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large Scale Hierarchical Image Database. In: CVPR09 (2009).
35. Maas, A., Hannun, A., Ng, A.: Rectifiifier nonlinearities improve neural network acoustic models. In: Proceedings of the International Conference on Machine Learning. Atlanta, Georgia (2013).