IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 6, 2020, accepted September 6, 2020, date of publication September 25, 2020, date of current version October 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3026684

Semantic Segmentation Using a GAN and a Weakly Supervised Method Based on Deep Transfer Learning

SHUHUAN WEN[1,2,3], WENBO TIAN[1,2], HONG ZHANG[3], (Fellow, IEEE), SHAOKANG FAN[1,2], NANNAN ZHOU[1,2], AND XIONGFEI LI[1,2]

[1]Engineering Research Center of the Ministry of Education for Intelligent Control System and Intelligent Equipment, Yanshan University, Qinhuangdao 066004, China
[2]Key Laboratory of Industrial Computer Control Engineering of Hebei Province, Yanshan University, Qinhuangdao 066004, China
[3]Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada

Corresponding author: Shuhuan Wen (swen@ysu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Project 61673125 and Project 61773333, and in part by the China Scholarship Council (CSC) under Project 201908130016.

ABSTRACT Semantic image segmentation is of crucial importance to many applications, such as autonomous driving, robot vision, and scene understanding. However, the border of a segmented image tends to be rough, and the labeling process is tedious and labor-intensive. Therefore, this study is the first proposing to use a deep generative adversarial network (GAN) with double-layered upsampling based on max-pooling indexed deconvolution. Our proposed upsampling method replaces the bilinear interpolation upsampling method; i.e., we fuse the deep deconvolution method by saving the indices of relative locations of the max weights computed during pooling. Combined with the deep GAN, our upsampling method can improve the extraction of low-resolution features, and compensate for the loss of the image size. To further reduce the whole network's dependence on labeled datasets, a weakly supervised feedback method is proposed. The unlabeled data can improve the generalization ability of the model. Considering the generalization to unseen image domains, we introduce transfer learning based on a deep GAN and a weakly supervised method. The segmentation model using the trained data in the source domain can obtain good segmentation in the target domain using transfer learning. Extensive experiments in various domains demonstrate the advantages of the proposed method compared to the generalization ability of semantic segmentation. This method also significantly decreases the dependence on labeled data and ensures the network accuracy.

INDEX TERMS Semantic segmentation, GAN, deep transfer learning.

I. INTRODUCTION

Traditional approaches such as manually designed features, support vector machines (SVMs) and probability graphs, have been used to build semantic segmentation algorithms. Ren and Malik [7] propose a simple linear iterative clustering (SLIC) algorithm that can result in unstable super pixels, wrong classification, and weak boundary region. This algorithm is difficult to apply in the segmentation of super pixels. With the development of deep learning, many image semantic segmentation methods based on deep learning have been proposed, including image classification [8], [9] and object detection [10]–[13]. Recently, convolutional neural networks (CNNs) have been a common approach for semantic segmentation [14]–[16] since they provide an initial category label for every pixel. A convolutional layer can effectively capture the local features of an image and nest the modules together in a hierarchical manner [17], [18], but the traditional CNN may lose spatial information in the deep layers of the network, and the size of the input picture is fixed. Fully convolutional networks (FCNs) were proposed to handle images of any size by transforming the fully connected layer to a convolutional layer [19]. Generative adversarial networks (GANs) have also been applied to semantic segmentation [20], [21]. A deep GAN can be used to judge real label images and predictive segmentation images,

The associate editor coordinating the review of this manuscript and approving it for publication was Mingjun Dai.

which can reduce the inconsistency between them. However, the detailed information of the final segmentation image is lost, and the segmentation boundaries are rough.

This paper proposes a double-layered upsampling method based on a deep GAN. The discriminator output of a deep GAN uses a supervision signal to feed back the predictive results of the semantic segmentation. Then, the lost detailed information is captured in the samples of the semantic segmentation network, which can improve the quality of the boundaries of the segmented regions. Most of the traditional semantic segmentation networks use fully-supervised CNNs, which require strict training conditions and imply training using labeled data. The labeled data needs manual labor, and the labeled data set also needs to be specially processed. This paper proposes a double-layered upsampling method based on a deep GAN. The classification output of the deep GAN is used to feed back the predicted results of the semantic segmentation network. Then, the lost detailed information is captured during the upsampling process of the bilinear interpolation, which can improve the quality of the boundaries of the segmentation. The weakly supervised segmentation method with feedback is used to train the whole semantic segmentation network, and to avoid the problem of requiring numerous manual labels.

These semantic segmentation methods can only be used in a specific environment. The generalization ability of the segmentation model is low for data outside the specific environment. A highly accurate network that is trained using a specific data set cannot obtain similar performance on other similar data sets which belong to the same kind of scene. In this paper, we present a novel semantic segmentation method using a GAN and weakly supervised segmentation based on deep transfer learning. Our method trains the whole semantic segmentation network using a weakly supervised segmentation method with feedback, which is based on a deep GAN. The proposed method can solve the problem of requiring lots of manually labeled data and simplify the work of obtaining high quality data. We use two kinds of data sets, labeled data and unlabeled data sets, during the training process of the network. The unlabeled data is similar to the labeled data. Weakly supervised training can reduce the dependency on labeled data for the whole network, which further reduces the semantic segmentation network's dependence on the external environment. The unlabeled data samples are used to perform segmentation predictions automatically, which can improve the generalization ability of the model. Transfer learning is combined with the proposed GAN and the weakly supervised segmentation method based on deep learning. The segmentation model that is trained using the data from the source domain can obtain a good segmentation effect in the target domain via transfer learning.

The remainder of this paper is organized as follows. In Section II, we discuss the related works on semantic segmentation in further detail. In Section III, we introduce our novel method to address this problem, focusing on improving the efficiency and accuracy of the semantic segmentation based on deep GAN. The comparative experimental results are described in Section IV, and, finally, Section V summarizes our method and concludes the paper.

II. STATE OF THE ART

Recently, some semantic segmentation methods have been proposed to recognize rich semantic features using pre-trained networks [22]–[25], but these methods have low segmentation accuracy. Pohlen *et al.* [26] propose the ResNet network architecture to obtain accurate segmentation boundaries. Ref. [27] proposes the Border network (BN) to distinguish different adjacent regions of semantic labels with similar forms, which can determine the semantic boundaries and guide the network learning. Dai *et al.* [28] introduce the set of manually labeled image boundaries; and in their method, the convolutional features of super-pixels are extracted from the image domains and used to train a classifier. Reference [29] introduces a GAN to improve the boundary accuracy of segmentation. Krähenbühl and Koltun [30] propose an effective fully-connected conditional random field (CRF) to improve the segmentation and labeling accuracy. The above methods mainly consider the semantic relevance of the object segmentation boundaries at the pixel level rather than focusing on the feature extraction of shallow channels, including boundary textures. Meanwhile, these methods are only used in specific environments or data sets and require manually labeled data. The trained network model is not suitable for similar or different environments. Some studies [31], [32] have reported the use of game engines to fuse image data for automatic driving. This approach can decrease the amount of manual labor and computational requirement.

However, synthetic images and tangible images have considerable errors. Some researchers propose using a model trained using synthetic data to transfer tangible images. Hoffman *et al.* [33] introduce a domain adaptive semantic segmentation method that solves the pixel prediction problem using the first unsupervised GAN method based on the work of [34]. Zhang *et al.* [35] propose a learning method that reduces field gap of the semantic segmentation in city scenes. Huang *et al.* [36] propose a layering unsupervised domain adaptive semantic segmentation method that uses a GAN to adjust the activation distribution. Zou *et al.* [37] propose a UDA framework based on an iterative self-training process and a balanced self-training framework. The above domain transferring networks outperform single domain networks when semantic segmentation is performed. However, these networks are directly transferred to the deep layer of the segmentation network. A shallow network cannot obtain good transfer learning because it is far from the semantic output of the deep layer. We propose transitive domain adaptive transfer learning based on a deep GAN. The proposed method combines i) the multi-level GAN [38] together with ii) Appearance Adaptation Networks [39] and iii) the Shared Domain [40]. Different feature layers are applied to different weight transfer learning processes in the double-layered

upsampling of the segmentation network. The source domain and target domain train the semantic segmentation using the double-layered upsampling of the segmentation network. The source domain, a set of the labeled data, is trained in a fully supervised way, while the target domain, a set of the unlabeled data, is trained in an unsupervised way. In the transfer learning module, according to the GAN training, the data of different spaces is mapped to a certain feature space using the transitive domain adaptive method, and then the distribution of the conditional probability in the feature space becomes similar. The data in the source domain and target domain will be integrated when they cannot be distinguished.

III. PROPOSED METHOD

In this section, we first provide a method used in our computationally efficient semantic segmentation model. Then, we provide a detailed explanation of the method that fuses double-layered upsampling and weakly supervised learning in order to reduce the dependence on labeled data and improve the accuracy of the semantic segmentation. In order to improve the generalization ability of the segmentation model, this paper combines transfer learning with the proposed GAN and weakly supervised learning based on deep learning.

A. SEMATIC SEGMENTATION BASED ON DOUBLE-LAYERED UPSAMPLING AND WEAKLY SUPERVISED LEARNING

The detailed boundary information of a segmented image will experience losses when using bilinear interpolation upsampling because this method can result in an inaccurate reconstruction of the nonlinear structure of an object boundary of the segmented image. Therefore, we propose a double-layered upsampling method based on the deep GAN network. The deep GAN network refers to a deep generative adversarial network. The deep GAN network uses two independent sub-neural networks, which are called the ''generator'' and the ''discriminator''. During the training process, these two sub-networks perform the minimum and maximum value mechanisms. The generator outputs a sample of the target data distribution with a random vector, and the discriminator distinguishes the sample generated by the generator from the target sample. The generator obfuscates the discriminator through backward propagation, and thus the generator generates samples similar to the target sample. We propose a double-layered upsampling method based on a dense upsampling convolution structure [41] and the idea of saving the indices of relative locations of the max weights computed during convolution pooling in a SegNet network [42]. The relative position of maximum weights is the position information of the maximum value in the maximum pooling process, that is, the relative position information of the brown squares in figure 11. In the process of deep deconvolution upsampling, the downsampled sparse feature map is compensated by the segmentation network. The discriminator output from the deep GAN is used as a supervisory signal that feeds back to the predictive results of the semantic segmentation network. Our proposed upsampling method can replace the bilinear interpolation upsampling method; i.e., we fuse the deep deconvolution method with saving the indices of relative locations of the max weights computed during pooling. Combined with the deep GAN, our upsampling method can improve the extraction of low-resolution features, and compensate for the loss of the image size. The network structure is shown in Fig. 1.

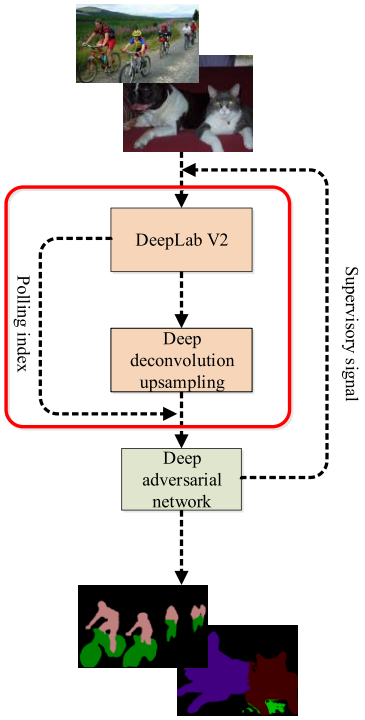


FIGURE 1. Deep deconvolution upsampling of a semantic segmentation network.

In figure 1, the proposed double-layered upsampling method replaces the bilinear interpolation upsampling method by fusing the deep deconvolution method with saving the indices of relative locations of the max weights computed during pooling. The discriminator output from the deep GAN is used as a supervisory signal that feeds back to the predictive results of the semantic segmentation network.

Our semantic segmentation network model uses the DeepLab v2 network without multi-scale fusion as the baseline network. We use the ResNet-101 model pre-trained on ImageNet. Atrous spatial pyramid pooling (ASPP) is used for the final classification. Finally the double-layered upsampling method is used to output a classification prediction with the same size as the input image. The discriminator network uses 5 full convolutional layers. The generator net contains convolutional layers.

First of all, the original image is input, and the final output of the semantic segmentation network is the initial segmentation prediction map that maps with the original image. The deep anti-neural network serves as a component of the

discriminator, and the discriminator network is trained with the real labeled image; then the semantic segmentation is performed. The initial segmentation prediction map output by the network is input to the discriminator. If the pixel-level label in the segmentation prediction map matches the pixel-level label in the real marked image in the discriminator, then the discrimination is true, otherwise, the discrimination is false, and finally the deep adversarial neural network will output a discriminated probability map. The probability graph is used as the supervising signal of the semantic segmentation network to train again. After many iterations, it can achieve the effect of deeply resisting the indistinguishability of the neural network. According to the discriminator network proposed by Yu *et al.* [27], the anti-loss function and the standard cross-entropy loss function are combined through the semantic segmentation network to improve the effect of semantic segmentation.

The entire network optimizes the objective function. It combines the traditional standard cross-entropy loss function with the confrontation loss function. This confrontation mechanism motivates the semantic segmentation network to generate prediction labels. Since the deep adversarial neural network can evaluate the joint configuration of multiple label variables, it can enforce various forms of higher-order consistency. This kind of consistency cannot be performed by paired terms or cross-entropy losses of per pixel are measured. The adversarial training method enhances the continuity of spatial labeling without increasing the complexity of the model used in the test. Moreover, the adversarial model can flexibly detect mismatches in a large range of high-order statistics between the model prediction and the real image without manual labeling. The entire training process is a classic game idea, improving the network's ability mutually, refining the segmentation accuracy and enhancing the discriminating ability.

The probability map in the network shows the regional quality of the predicted labels output in the semantic segmentation network, so that the semantic segmentation network can automatically identify which regions are judged to be true labels and which regions are judged to be the predicted labels output by the segmentation network during the training process. A loop iteration of network training is performed on the predicted label regions that meant to be the output of the segmentation network, and the result of the segmentation prediction map is maximized, which is close to the real labeled image.

The double-layered upsampling method uses the method of saving the indices of relative locations of the max weights computed during the SegNet network pooling process. During the upsampling process, each maximum weight position will be saved after the maximum pooling in the entire segmented network is restored. The position where the largest weight is located and the weights of the other positions are 0, that is, we get the feature map after depooling. In addition, the input feature map is subjected to deep deconvolution upsampling, and a deep deconvolution method is used to increase the number of channels. The depth deconvolution network graph is shown in Fig. 2.

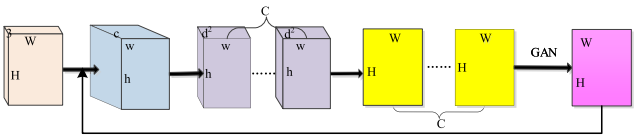


FIGURE 2. Depth deconvolution network graph in double-layered upsampling semantic segmentation network.

In figure 2, when the size of the input image is $(H, W, 3)$, the size of the feature map of the entire model before prediction is (h, w, c), where c is the number of feature channels, $H/d = h$, $W/d = w$, d is the downsampling factor, and the channel of the output feature map (h, w, c) is converted to $(h, w, d^2 \times C)$. C is the number of semantic categories of the segmented object. Then, the feature map is enlarged to (H, W, C) through dimensional conversion, which obtains the feature map after deep deconvolution. Finally, the feature map after de-pooling and the feature map after deep de-convolution are superimposed. The feature map obtained by deep de-convolution is used to fill the missing content of the de-pooled feature map, and finally the label prediction map is obtained by the segmentation network output. The segmented and predicted image (H, W, C) is input to a deep GAN, and finally the discriminant probability map $(H \times W \times 1)$ is output through network discrimination. The probability map is used as the supervised signal to perform self-learning by combining it with the supervisory signal of the discriminator. Through network iterations, the output of the segmentation network is continuously optimized to obtain an accurate semantic labeled map. The yellow box is the segmented and predicted image. The pink box is the discriminant probability map output through network discrimination.

The semantic segmentation network uses the DeepLabv2 network without multi-scale fusion as the baseline network, uses the ResNet-101 pre-training model on ImageNet, sets the stride of the last two convolutional layers to 1, and sets the dilation settings of the 4th and 5th convolutional layers. For 2 and 4, the final layer uses porous spatial pyramid pooling (ASPP) for final classification, uses a two-level merge upsampling method, and finally outputs a classification prediction with the same size as the input image. The deep adversarial neural network uses 5 full convolutional layers, kernel_size is set to 4, stride is set to 2, the number of channels is {64,128,256,512,1}, in addition to the input layer, BN layer is added after the convolution of each layer. Each of the first 4 convolutional layers is followed by a leaky Relu layer to prevent gradient sparseness. Its parameter is 0.2, and the last convolutional layer is followed by an upsampling layer. The BN layer is not used in the output layer of the semantic segmentation network and the input layer of the deep anti-neural network. The BN layer added after the remaining layers are convolved to prevent the semantic segmentation

network from converging all segmentation prediction results to one point.

In order to simplify the work of obtaining high-quality data, the weakly supervised method is applied to semantic segmentation, and the deep GAN network is used to achieve the weakly supervised learning of image segmentation. Traditional image semantic segmentation networks require a large number of manually labeled datasets for training, and each pair of accurately labeled images takes about one hour to process. In order to simplify the work of obtaining high-quality data, the weakly supervised method is applied to semantic segmentation, and the deep GAN is used to perform the weakly supervised learning of the image segmentation [43]. The deep GAN using unsupervised training can be widely used in the fields of unsupervised learning and weakly supervised learning. Compared to other models, the deep GAN can produce clearer and more realistic samples. The structure of the semantic segmentation network trained using the weakly supervision method is shown in Fig. 3. For the weakly supervised learning, a few labeled dataset samples are used for network training, which can reduce the demands on the number of manually labeled samples in the preparation process of the dataset and save considerable resources.

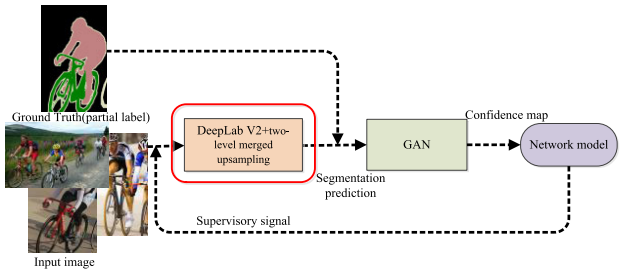


FIGURE 3. Weakly supervised semantic segmentation algorithm structure.

The weakly supervised method [43] is used to train the whole segmentation network, and the input of the given image is the labeled and unlabeled datasets. The semantic segmentation network combines the cross-entropy loss function L_{seg} and the deep GAN loss function L_{adv} to generate the segmentation prediction graph S(ξ) which is similar to the real labeled image in a high-order form by stimulating the semantic segmentation network. We use the same definitions of L_{adv}, L_{semi} and L_{seg} as ref. [43].

The total loss function L_o is defined as follows:

$$L_o = L_{seg} + \lambda_1 L_{adv} + \lambda_2 L_{semi} \tag{1}$$

where L_{seg} represents the segmentation loss function, L_{adv} represents the adversarial loss function, L_{semi} represents the weakly supervised loss function, and $\lambda 1$, $\lambda 2$ are two weights for minimizing the proposed multi-task loss function.

The final goal is to minimize the segmentation loss function in the segmentation network and maximize the probability that the label prediction graph is regarded as the real label graph in the deep GAN discriminator. It can be expressed as

$$\max_D \min_G L_o \tag{2}$$

Polynomial decay is used for network training to decrease the learning rate in this paper. The learning rate will attenuate to 0 when the maximum number of iterations is reached. The formula is defined as follows:

$$lr = base\_lr \bullet (1 - \tau/N)^{power} \tag{3}$$

where $power = 0.9$, lr is the learning rate, $base\_lr$ is the initial learning rate, τ is the current iteration number, and N is the maximum iteration number.

To evaluate the segmentation accuracy of the proposed method, we use the following evaluation metrics.

$$IOU = \frac{SR \cap GT}{SR \cup GT} \tag{4}$$

where SR is the segmentation result, and GT is the Ground Truth.

The proposed method adds a double-layered upsampling method to the weakly supervised method [43] segmentation network, which can obtain better segmentation results for small objects. As shown in Fig. 3, we use the DeepLab v2 network as the baseline network, and use the ResNet-101 trained using ImageNet as the pre-trained model. The stride of the last two convolution layers is 1, and the dilations of the fourth and fifth convolution layers are 2 and 4, respectively. In the last layer, atrous spatial pyramid pooling (ASPP) is used for the final classification. The double-layered upsampling method is used to output a classification prediction with the same size as the input image. The deep GAN uses 5 full convolutional layers, where the kernel size is 4, the stride is 2, and the numbers of channels are {64,128,256,512,1}. A BN layer is added to each convolution layer except the input layer. Each layer of the first 4 convolutional layers is followed by a leaky ReLU layer and its parameter is 0.2. The last convolutional layer is the upsampling layer. The weakly supervised method randomly iterates using the labeled dataset and the unlabeled dataset. When randomly selecting labeled data and unlabeled data, different random seeds are used for selection to ensure the robustness of the overall network. In order to prevent the model from being affected by the initial noise mask, the segmentation network starts weakly supervised training after 5000 labeled data set training sessions. Compared with deeplab v2-adv, we can highlight that our method has a better segmentation effect.

Table 1 is the evaluation results of our proposed method compared with SmallFov-light [21] and DeepLab v2-adv [43] using the supervised and weakly supervised processes with 25% labeled data after 20,000 iterations. The results show that our method has much better segmentation accuracy than the other methods.

TABLE 1. MIOU evaluation of different algorithms after 20000 iterations.

Methods	MIOU/%
Lue P[21]	72.00
Segmentation Network[43]	74.90
Segmentation Network + Dense upsampling	76.80
Weakly-supervised Network[43]	72.10

B. TRANSFER LEARNING BASED ON THE DEEP GAN NETWORK

The overall network structure is shown in Fig. 4. This structure mainly includes two semantic segmentation networks based on source domain data, target domain data and a multi-threaded transfer GAN. The semantic segmentation network acts as a generator, and the transfer GAN acts as a discriminator. The features of the shallow level in the segmentation network cannot well adapt to the network because they are far from the deep level of the output labels. In order to solve this problem, according to the multi-layer strategy of adversarial learning composed of different feature layers in a segmentation model which was proposed in [39], adversarial learning is added in the shallow layer and the final output layer of the network. In order to make the output target prediction closer to the source prediction, the discriminator network is used to distinguish whether the input is an image from the source domain or the target domain. Then, the adversarial loss is computed based on the output of the target prediction and back propagated to the segmentation network. After several iterations, domain adaptation segmentation is achieved.

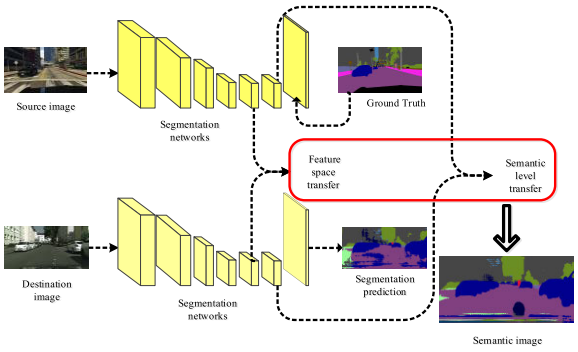


FIGURE 4. Semantic segmentation structure based on the multi-threaded transfer GAN.

In Figure 4, the yellow box indicates the training of the semantic segmentation on the source and target domain data via the double-layered upsampling semantic segmentation network. The red box indicates that the data of different spatial distributions are mapped to a feature space through the domain adaptation method, and the conditional probability distribution in the feature space becomes increasingly closer via adversarial training. The features learned by the network are equally applicable to the source and target domain tasks rather than just a specific segmentation task, which makes the learned features generalizable. Finally, the probability of the segmentation network prediction based on the target domain and the source domain approaches is maximized, and this completes the transfer task from the source model to the target domain.

In this paper, the alignment of the inherent pixel-level and feature space structures in the two domains is included in the GAN of each thread to improve the domain distribution alignment problem between the synthesized data and the real data [40].

It is supposed that each picture is divided into $m \times m$ areas, where $m = 1, 2, 3, \ldots\ldots N$; I_m is the index of the sub-domain in each image; Γ_m^s is the activation function of domain I_m where the image in the source domain is located; Γ_m^t is the activation function of domain I_m where the image in the target domain is located. Then, the loss function of the spatial adaptation is

$$L_{sa} = \min \sum_{m=1}^{N} L_{da}\left(I_m^s, I_m^t\right) \tag{5}$$

where L_{da} is the loss function of the domain adaptation, which is used to measure the difference in the feature domains of the two domains. L_{sa} is the loss function of the spatial adaptation. The training loss function of the domain classifier is defined as:

$$L_D(\xi^s, \xi^t) = \min \frac{1}{\xi} \sum_{x \in \xi} L(f(\xi), y) \tag{6}$$

where y is the tag data, ξ is the sum of the training data, and $L(.)$ is the cross-entropy loss function. The equation is defined as:

$$L\left(f\left(\xi\right), y\right) = -\sum\nolimits_{n=1}^{k} \left(y_i = k\right) \bullet logf\left(\xi\right)_n \tag{7}$$

$y_i = k$ represents the indicator function and $f(\xi)$ is the prediction classifier.

In the segmentation network, the shallow layer of the network generally extracts the spatial feature information of the image, and the deep layer of the network generally represents the complex semantic information. By combining image ξ^s with the domain adaptation image ξ^t, the shallow level features of the data image in the target domain are separated in the whole segmentation network, which mainly encodes the shallow texture features of data images. The deep level features of the data image of the source domain are separated in the whole segmentation network, which mainly encodes the semantic features of the data image. Combining the shallow level texture features in the target domain with the deep level semantic features in the source domain produces the final domain adaptation image. It is assumed that each convolutional layer l in a deep convolutional neural network has θ_l corresponding to the mapping response (i.e., θ_l channels), and the size of each channel is $H_l \times W_l$. Then, the characteristic response of each convolutional layer

l can be expressed as $\Theta_i^l \in R^{\theta_l \times H_l \times W_l}$ $(i = 0, 1)$, where 0 represents the target domain and 1 represents the source domain. The responses of different convolutional layers characterize the image content at different semantic levels. The shallow layer responds to the underlying features, and the deep layer responds to higher semantic features. In order to control the semantic content in the source image ξ^s better, different weights W are assigned to different layers to reflect the effect of each layer. The content in image ξ^s is preserved in the domain adaptation image ξ^s by minimizing the Euclidean distance of the function. The response objective function of the content is expressed as

$$D_{(s,t)} = \min_{\xi^t} \sum_l W^{(l,1)} L_{dis}\left(\Theta^{(l,0)}, \Theta^{(l,1)}\right) \tag{8}$$

The total objective function of the transitive transfer GAN is

$$L\left(\xi^s, \xi^t\right) = L_{sa} + L_D(\xi^s, \xi^t) + D_{(s,t)} \tag{9}$$

In other words,

$$L\left(\xi^s, \xi^t\right) = \min \sum_{m=1}^{M} L_{da}(\Gamma_m^s, \Gamma_m^t) + \min \frac{1}{\xi} \sum_{x \in \xi} L(f(\xi), y) + \min_{\xi^t} \sum_l W^{(l,1)} L_{dis}\left(\Theta^{(l,0)}, \Theta^{(l,1)}\right) \tag{10}$$

The segmentation loss function is the cross-entropy loss function in the segmentation network, and L_{seg} represents the cross-entropy loss function.

$$L_{seg} = -\sum_{h,w} \sum_{c \in C} T^{(h,w,c)} log\left[\boldsymbol{S}(X_n)^{(h,w,c)}\right] \tag{11}$$

The adversarial loss function is the loss function in the discriminator network, and L_{adv} is the adversarial loss function.

$$L_{adv} = -\sum_{h,w} log\left\{\boldsymbol{D}\left[\boldsymbol{S}(X_n)\right]^{(h,w,1)}\right\} \tag{12}$$

The total loss function is

$$L_o = L_{seg} + \lambda_1 L_{adv} \tag{13}$$

The standard function for network optimization is

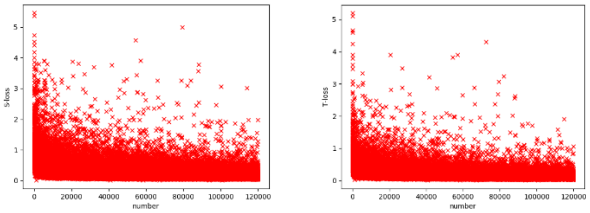
$$\max_D \min_G L_o L\left(\xi^s, \xi^t\right) \tag{14}$$

By maximizing the discrimination of the GAN and minimizing the function of the segmentation network, finally, the transfer from the source domain to the target domain is achieved, thus improving the generalization ability of the network.

IV. EXPERIMENTAL RESULTS

In this section, we perform a set of experiments to evaluate our proposed method and compare it with other state-of-the-art methods [38].

In order to reduce the detail information loss caused by the bilinear interpolation in the semantic segmentation networks and improve the accuracy of the segmentation boundaries, we propose a double-layered upsampling method. Different from the traditional one-time return to full resolution prediction image segmentation method, we use a deep deconvolution to gain a series of amplification filters, which are the convolution kernel of the segmentation network when amplifying low-resolution features, and enlarge the reduced feature map to the same resolution as the input image. Then, we combine this with the maximum pooling after saving the maximum weight of each filter location. The deconvolution depth is used as the characteristic of the figure in the pooling to get the characteristics of the figure which can be used to fill in the missing content. Finally, the result will be divided by the boundary of the network output informative label forecast figure. The experimental structure is shown in Fig.6. It can be seen that our method has better results in the boundary segmentation of the bottle, the girl's leather shoes, and the regressing animal, which indicates that the double-layered upsampling for smaller objects is more ideal than the bilinear interpolation upsampling segmentation network.



a) multi-low b) multi-high

FIGURE 5. The relationship between the loss and the steps of iterations based on different network models.

The experimental results after 20,000 iterations under full supervision training are shown in figure 10. In Figure 10, the first column is the original image, the second column is the Ground Truth, the third column is the baseline network, the fourth column is the DeepLab v2-adv network [43], and the fifth column is the proposed method. In Figure 10, the third column is the segmentation result without using the GAN, and the fourth and fifth columns are the segmentation results using the GAN. Figure 10 shows that using the GAN can improve the accuracy of segmentation boundaries. Especially, the proposed method performs better on the boundary of the bottle, the boundary of the little girl's leather shoes, and the boundary of the animal leg.

Table 4 compares the evaluation results of the proposed method with the other segmentation methods after 20,000 iterations under full supervision training. Through the comparison of the MIOU, the accuracy of the proposed method improves from 74.9% to 75.7% compared with the DeepLab v2-adv [43] method.

Table 5 is the evaluation value of the segmentation prediction of the dataset category after the baseline network. DeepLab v2-adv [43] method and the method in this paper are iteratively trained 20,000 times under full supervision.

FIGURE 6. The results of the segmentation prediction using a fully-supervised baseline network, a double-layered upsampling network and the proposed algorithm with 50% of the samples.

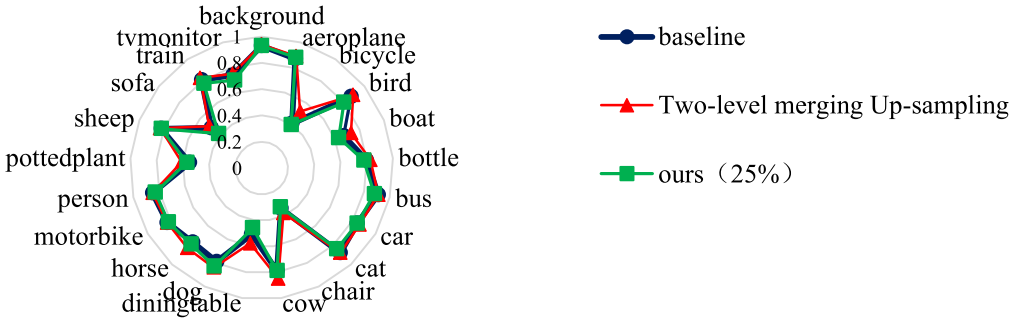


FIGURE 7. Comparison of the segmentation accuracy of the fully supervised and weakly supervised training segmentation networks in 21 categories.

It can be seen from the data in the table 6 that the proposed method data is significantly different from other algorithms.

In addition, we choose the weak supervised training network with the 50% labeled data set to compare it with the double-layered upsampling network and the baseline network

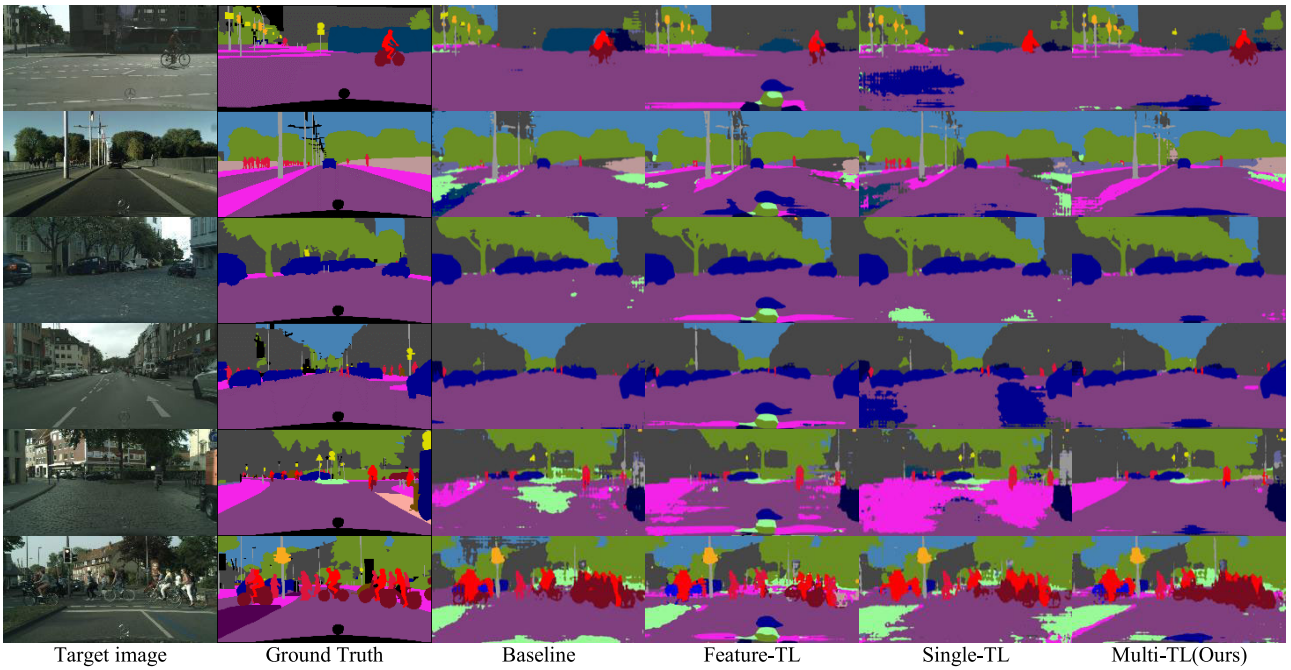


FIGURE 8. Comparison of the semantic segmentations of four network models.

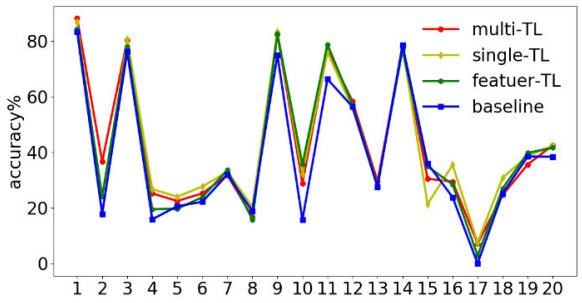


FIGURE 9. Results of the segmentation accuracy of the baseline segmentation network, the transfer GAN based on the shallow channel and deep channel and the multi-thread segmentation network using the proposed method, in which 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, and 20 represent road, sidewalk, building, wall, fence, pole, light, sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, bicycle, and MIOU, respectively.

with full supervision. The results show that when using weak supervision, the thinning of the segmentation image boundaries is not as accurate as that when using full supervision, but the accuracy of the intra-class segmentation prediction is higher than that of the baselines when using full supervision.

We use two challenging available datasets, "GTA5" [44], and "Cityscapes" [45], for double-layered upsampling semantic segmentation to quantitatively evaluate the proposed method in section B.

The cityscapes dataset contains 5000 images (2975 training images, 1525 test images and 500 validation images) with a resolution of 2048 × 1024. The GTA5 dataset consists of 24966 images with a resolution of 1914 × 1052. During testing, we conduct the evaluation on the Cityscapes validation set with 500 images that contain 19 categories.

The Stochastic Gradient Descent (SGD) with momentum is used for the double-layered upsampling semantic segmentation network of the baseline network in the experiment. The initial learning rate of the network is set to 2.5×10^{-4}, and polynomial attenuation of $n = 0.9$ is used to reduce the learning rate. When the maximum number of iterations is reached, the learning rate is attenuated to 0. The equation is shown as equation (3).

The transfer GAN based on the transitive domain adaptation proposed in this paper is trained. The transfer GAN based on multi-thread feature extraction and deep level and shallow level feature extraction are respectively trained, and the maximum number of iterations for each network is 120,000. The relationship between the loss and the number of iterations in the training process of the network model which is obtained by the transfer in the deep channel and shallow channel is shown in Fig. 5. Figs. 5 a) and b) show that the loss of the deep channel and shallow channel transfer network is basically stabilized at 1.0 or less in the training process of the transfer GAN based on multi-thread feature extraction when the number of iterations is about 60000. Furthermore, when the number of iterations of the network based on the deep channel is 80,000, the loss is basically stabilized below 0.5. The results show that the deep channel and shallow channel transfer can achieve good results, and the network model can converge.

The shallow layer of the network generally extracts the spatial feature information of the image, and the deep layer of the network generally displays abstract semantic information. The loss value can intuitively demonstrate the accuracy

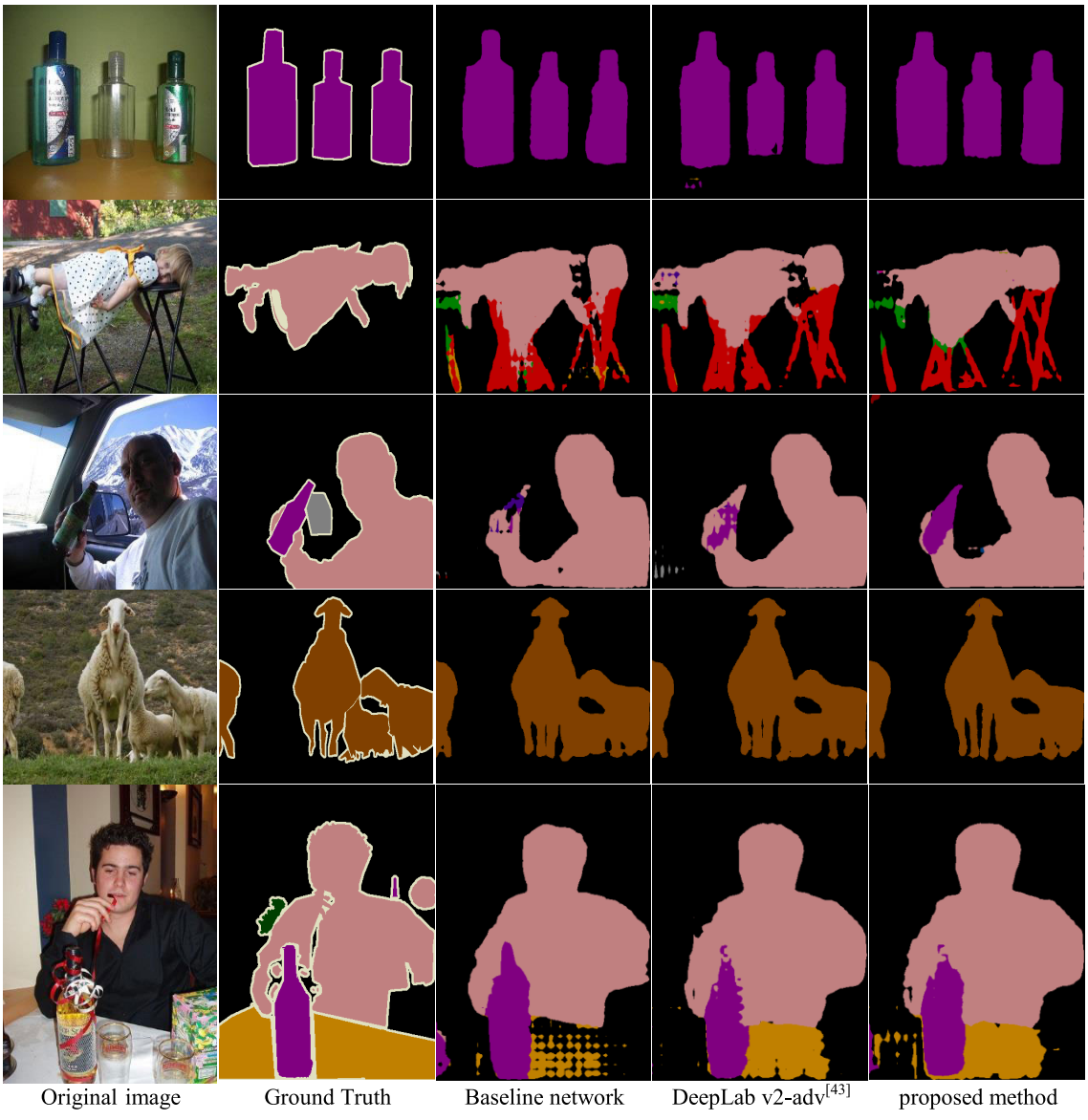


FIGURE 10. Comparison of the segmentation accuracy of the baseline network, the DeepLab v2-adv [43] method and the proposed method after 20,000 iterations under full supervision training.

change of the model during the training process. The lower the loss value is, the higher the model accuracy is and the better the performance is. We use cross entropy loss function in this paper. The T-SNE is a non-linear dimensionality reduction machine learning algorithm. It was proposed in 2008 and it is very suitable for the situation when decreasing the dimensionality from high dimensionality to 2 or 3 dimensions. It is not applicable to this article. The effect of the PAC is worse than T-SNE, so we do not use these two methods.

The trained segmentation network based on the transitive domain adaptation transfer adversarial method which is proposed in this paper is performed. The final network model is evaluated by using 500 verification images from the Cityscapes dataset, and the output of the semantic segmentation images is shown in Fig. 8. In Fig. 8, the first column is the image in the target domain, the second column is the Ground Truth, the third column is the image using the baseline segmentation network, the fourth column is the transfer GAN segmentation based on the shallow channel, the fifth column is the transfer GAN segmentation based on the deep channel, and the sixth column is the transfer GAN segmentation based on multi-thread feature extraction. The baseline segmentation network only trains the network model using the source domain dataset. Through the comparison of the semantic segmentations of the four network models in Fig. 8, the results show that the segmentation effect based on the transitive domain adaptation transfer adversarial method which is proposed in this paper is more accurate, and the boundary information of the object is also more accurate. Particularly, the semantic segmentation of larger

TABLE 2. Different proportional labeled datasets using the proposed method in this paper.

	12.5% labeled samples		25% labeled samples		50% labeled samples	
categories	DeepLab v2-adv[43]	our	DeepLab v2-adv[43]	our	DeepLab v2-adv[43]	ours
background	0.93	0.93	0.93	0.93	0.93	0.94
plane	0.83	0.84	0.87	0.86	0.88	0.88
bicycle	0.42	0.39	0.40	0.41	0.40	0.41
bird	0.82	0.84	0.81	0.87	0.84	0.87
ship	0.62	0.61	0.63	0.67	0.64	0.67
bottle	0.75	0.74	0.71	0.75	0.78	0.78
Bus	0.89	0.88	0.90	0.90	0.90	0.91
car	0.80	0.80	0.84	0.84	0.85	0.83
cat	0.84	0.86	0.86	0.87	0.89	0.88
chair	0.31	0.33	0.33	0.32	0.35	0.34
cow	0.65	0.73	0.75	0.76	0.84	0.77
table	0.46	0.48	0.49	0.50	0.46	0.55
dog	0.73	0.79	0.78	0.80	0.83	0.81
horse	0.72	0.66	0.72	0.74	0.79	0.77
motorcycle	0.75	0.71	0.79	0.81	0.82	0.81
human	0.82	0.80	0.83	0.83	0.83	0.84
potted	0.46	0.50	0.59	0.53	0.57	0.58
sheep	0.76	0.75	0.82	0.79	0.82	0.80
sofa	0.40	0.44	0.45	0.46	0.42	0.50
train	0.79	0.81	0.80	0.81	0.82	0.83
display	0.73	0.70	0.74	0.73	0.72	0.74
MIOU	0.68	0.69	0.71	0.72	0.73	0.74

volume categories in the target domain can be accurately obtained by using the pixel-level features of the same spatial domain for domain alignment. Furthermore, the smaller volume categories are easily segmented into categories that are close to the larger volumes by the transfer model.

The comparative evaluations of the proposed method and the method in ref. [38] are shown in Table 3 and Fig. 9. In the multi-thread transfer GAN, the basic texture features of the data are extracted by the shallow convolution layer, and the complex semantic features of the data are extracted by the deep level convolutional layer. Then, using the feature structure alignment method for the spatial domain, the features of the shallow layer in the target domain are combined with the semantic features of the deep layer in the source domain, and finally transitive domain adaptation transfer learning is achieved. In the transfer learning of the deep channel, a domain adaptation method combining the semantic feature of the deep level in a source domain with the underlying feature of the shallow level in the target domain is compared with the single-level [38] based on the DeepLab v2 baseline network. Table 3 shows that the segmentation accuracy of the transfer learning method proposed in this paper is more accurate. We compare the two methods in the shallow channel which add spatial feature domain distribution alignment in a discriminator using a baseline network and a domain adaptation method in the pixel-level output space based on the DeepLab v2 baseline network [38]. Table 3 demonstrates that the adaptation effect of the proposed method in this

TABLE 3. Results of the segmentation accuracy using the proposed method and ref. [38].

	road	sidewalk	building	wall	fence	pole	light	sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	MIOU
multi(our)	**88.23**	**36.70**	80.69	25.74	23.02	26.24	32.90	18.74	82.56	29.92	76.69	58.45	**29.61**	77.89	30.57	29.43	7.23	25.74	36.86	**43.01**
multi[38]	86.50	36.00	79.90	23.40	23.30	23.90	35.20	14.80	**83.40**	33.30	75.60	**58.50**	27.60	73.70	32.50	35.40	3.90	30.10	28.10	42.37
single(our)	86.94	17.85	**80.90**	**26.81**	**24.05**	**27.72**	33.01	20.42	83.38	32.29	76.10	55.97	27.95	78.14	21.38	35.51	7.61	**30.94**	**38.72**	42.41
single[38]	86.50	25.90	79.80	22.10	20.00	23.60	33.10	**21.80**	81.80	25.90	75.90	57.30	26.20	76.30	29.80	32.10	7.20	29.50	32.50	41.43
feature(our)	84.34	2.42	78.04	19.57	19.73	23.89	**33.71**	15.85	82.65	**35.73**	**78.86**	57.24	28.18	76.87	34.96	28.48	2.88	27.07	39.85	41.69
feature[38]	83.70	27.60	75.50	20.30	19.90	27.40	28.30	27.40	79.00	28.40	70.10	55.10	20.20	72.90	22.50	**35.70**	**8.30**	20.60	23.00	39.25
baseline(our)	83.46	17.88	76.26	15.93	20.64	22.23	31.99	18.93	75.01	15.77	66.38	56.45	27.55	**78.60**	**36.00**	23.88	0.00	25.30	38.53	38.46
baseline[38]	75.80	16.80	77.20	12.50	21.00	25.50	30.10	20.10	81.30	24.60	70.30	53.80	26.40	49.90	17.20	25.90	6.50	25.30	36.00	36.64

TABLE 4. MIOU evaluation of different algorithms after 20000 iterations.

method	MIOU/%
SmallFov-light[21]	72.00
DeepLab v2-adv baseline network[43]	73.60
DeepLab v2-adv[43]	74.90
Proposed method	75.70

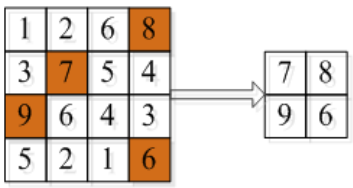


FIGURE 11. Max-pooling process.

paper is better than that of the discriminator of the baseline network [38].

Table 1 and table 2 adopting Pascal VOC2012 dataset [46], mainly use 21 categories of image segmentation data of Pascal VOC2012 dataset. The Pascal VOC2012 dataset is used to select labeled datasets of different proportions, and the weakly supervised learning segmentation network proposed in this paper is compared to ensure the consistency of the experimental dataset. This experiment evaluates the output of the network model on the standard verification set of 1449 images. In the training process, a random crop size of 321 × 321 is used.

The datasets used for network training in the table 3 are GTA5 dataset, Cityscapes dataset and SYNTHIA dataset. The Cityscapes dataset mainly uses data from the leftImg8bit folder and the gtFine folder. Each folder of the leftImg8bit folder and the gtFine folder contains three subfolders, namely train, val, and test, with a total of 5000 labeled images. Including 2975 training images, 500 verification images and 1525 test images, each image has a resolution of 2048 × 1024, which contains 50 city scenes with different scenes, different backgrounds, and different seasons, a total of 19 categories. The GTA5 dataset is a synthetic online game dataset, which contains 24,966 images from the game Grand Theft Auto and the label map of each image. The resolution of each image is 1914 × 1052, and there are 19 categories in total. The SYNTHIA dataset is similar to the GTA5 dataset. In this paper, the SYNTHIA-RAND-CITYSCAPES dataset for urban landscapes is selected, which contains 9400 labeled data. The resolution of each image is 1024 × 760, and there are 13 categories in total. First, the GTA5 dataset is used to train the fully supervised segmentation model of the unsupervised domain adaptive segmentation network proposed in this paper, and then combine it with the Cityscapes dataset for unsupervised semantic segmentation evaluation and

TABLE 5. Segmentation accuracy of baseline network, DeepLab v2-adv [43] method and proposed method after 20,000 iterations.

categories	baseline network	DeepLab v2-adv[43]	proposed method
background	0.93	0.94	0.94
plane	0.87	0.89	0.89
bicycle	0.41	0.41	0.41
bird	0.87	0.87	0.87
ship	0.67	0.67	0.69
bottle	0.80	0.81	0.81
bus	0.91	0.91	0.92
car	0.84	0.85	0.87
cat	0.88	0.88	0.89
chair	0.34	0.36	0.36
cow	0.79	0.83	0.83
table	0.51	0.53	0.56
dog	0.79	0.82	0.82
horse	0.77	0.80	0.81
motorcycle	0.83	0.83	0.83
human	0.85	0.85	0.85
potted	0.55	0.59	0.59
sheep	0.82	0.83	0.83
sofa	0.49	0.49	0.52
train	0.81	0.83	0.85
display	0.74	0.74	0.75
MIOU	0.74	0.75	0.76

TABLE 6. Comparison of proposed method in table 5 with other algorithms.

Algorithms.	Baseline network	DeepLab v2-adv[43]
Probability	$4.94*10^{-6}$	$1.89*10^{-3}$

verification. In this paper, 500 verification maps and 19 categories of semantic labels are used to verify and evaluate the semantic segmentation method of direct push domain adaptation in the experiment.

Table 4 and table 5 use the Pascal VOC2012 dataset. In this paper, we mainly use the image segmentation data of the PascalVOC2012 dataset, which contains 20 foreground object classes and 1 background class. The PascalVOC2012 dataset mainly includes three types of tasks: classification, detection and segmentation. The main research content of this paper is semantic segmentation, so the selected dataset is Pascal VOC2012 segmentation task dataset. The Pascal VOC2012 segmentation task dataset is used to test the effect of fully supervised segmentation on the two-level merged upsampling segmentation network proposed in this chapter. The segmentation task dataset contains 1464 training sets, 1449 verification sets and 1456 test sets, and pixel-level labeled images are used for training, verification and testing. In this paper, the network model is evaluated on the standard verification set of 1449 images. In the training process, the size of 321 × 321 was randomly scaled and cropped.

V. CONCLUSION

In this paper, we develop a segmentation structure and share many similarities between the source and target domains. Our method combines transitive domain adaptation, transfer learning and a deep GAN in a novel way. This method

improves the generalization ability of the semantic segmentation and reduces the number of manually labeled samples in an unsupervised way. We construct a multi-level GAN to train the shallow layer and deep layer of the segmentation network. To enhance the adaptive learning of the model, the feature of the shallow layer in the target domain is combined with the semantic feature of the deep layer in the source domain and the spatial structure of the pixel-level features in both domains is aligned by every thread of the GAN. The experimental results demonstrate that the proposed method is more accurate compared to the state-of-the-art algorithms.

REFERENCES

[1] K. J. Gao, S. Y. Sun, G. S. Yao, and H. T. Zhao, "Semantic segmentation of night vision images for unmanned vehicls based on deep learning," *J. Appl. Opt.*, vol. 38, no. 3, pp. 421–428, 2017.

[2] M. Ruihao, Z. Feng, W. Qingxiao, L. Rongrong, and W. Jingyang, "Dense stereo matching based on image segmentation," *J. Opt.*, vol. 39, no. 3, 2019, Art. no. 0315001.

[3] X. Huang, Z. G. Ling, and X. X. Li, "Discriminative deep feature learning method by fusing linear discriminant analysis for image recognition," *J. Image Graph.*, vol. 23, no. 4, pp. 0510–0518, 2018.

[4] A. Zhe, X. Xiping, Y. Jinhua, Q. Yang, and L. Yang, "Design of augmented reality head-up display system based on image semantic segmentation," *Acta Optica Sinica*, vol. 38, no. 7, 2018, Art. no. 0710004.

[5] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[6] T. M. Quan, D. G. Hildebrand, and W. K. Jeong, "FusionNet: A deep fully residual convolutional neural network for image segmentation in connectomics," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 1–10.

[7] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, vol. 1, Oct. 2003, pp. 10–17.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst*, vol. 60, no. 2, 2012, pp. 1097–1105.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[13] R. Girshick, "Fast R-CNN," in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[16] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 1529–1537.

[17] X. Ren, L. Bo, and D. Fox, "RGB-(D) scene labeling: Features and algorithms," in *Proc. Comput. Vis. Pattern Recognit.*, 2012, pp. 2759–2766.

[18] R. Socher, C. C. Y. Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *Proc. 28th Int. Mach. Learn.*, 2011, pp. 129–136.

[19] F. Liu, G. Lin, and C. Shen, "CRF learning with CNN features for image segmentation," *Pattern Recognit.*, vol. 48, no. 10, pp. 2983–2992, 2015.

[20] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[21] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 1–12.

[22] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 1–14.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1–10.

[27] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1–10.

[28] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3992–4000.

[29] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," in *Proc. NIPS Workshop Adversarial Training*, Barcelona, Spain, Dec. 2016, pp. 1–12.

[30] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian boundary potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 109–117.

[31] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6810–6818.

[32] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4500–4509.

[33] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," *CoRR*, vol. abs/1612.02649, pp. 1–9, Dec. 2016.

[34] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.

[35] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2020–2030.

[36] H. Huang, Q. Huang, and P. Krähenbühl, "Domain transfer through deep activation matching," in *Proc. ECCV*, 2018, pp. 1–16.

[37] Y. Zou, Z. Yu, B. V. K. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. ECCV*, 2018, pp. 1–17.

[38] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472–7481.

[39] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 6810–6818.

[40] Y. Chen, W. Li, and L. Van Gool, "ROAD: Reality oriented adaptation for semantic segmentation of urban scenes," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 1–10.

[41] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 1451–1460.

[42] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–10.

[43] W. C. Hung, Y. H. Tsai, Y. T. Liou, Y. Y. Lin, and M. H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–17.

[44] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. ECCV*, 2016, pp. 102–118.

[45] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. CVPR*, 2016, pp. 3213–3223.
[46] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

SHUHUAN WEN was born in Heilongjiang, China, in July 1972. She received the Ph.D. degree in control theory and control engineering from Yanshan University, Qinhuangdao, China, in 2005. She was a Visiting Scholar with Ottawa University, Carleton University, and Simon Fraser University, Canada, from 2011 to 2013. She is currently a Professor of automatic control with the Department of Electric Engineering, Yanshan University. She has coauthored one book and more than 40 articles. Her research interests include humanoid-robot control, force/motion control of a parallel robot, fuzzy control, and 3-D object recognition and reconstruction.

WENBO TIAN received the B.S. degree in automation from Yanshan University, Qinhuangdao, China, in 2017, where he is currently pursuing the M.S. degree with the Department of Electric Engineering. His main research interests include computer vision and SLAM.

HONG ZHANG (Fellow, IEEE) received the B.Sc. degree in electrical and computer engineering from Northeastern University, Boston, MA, USA, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA. He is currently a Professor with the Department of Computing Science, University of Alberta. His current research interests include robotics, computer vision, and image processing. He is a Fellow of the Canadian Academy of Engineering.

SHAOKANG FAN received the B.S. degree in automation from the Shandong University of Science and Technology. He is currently pursuing the M.S. degree with the Department of Electric Engineering, Yanshan University.

NANNAN ZHOU was born in Shandong, China, in December 1991. She received the bachelor's degree from the Department of Mechanical and Electrical Engineering, Dezhou University, in 2016. She has coauthored one journal article. Her research interests include humanoid–robot control and multirobot cooperation.

XIONGFEI LI was born in Hebei, China, in November 1994. He received the bachelor's degree from the Department of Mechanical and Electrical Engineering, Shaoxing University, in 2017. His main research interests include computer vision and SLAM.