Two-stage Domain Adaptation for Infrared Ship Target Segmentation

Ting Zhang, Haijian Shen, Sadaqat ur Rehman, Zhaoying Liu*, Yujian Li, Obaid ur Rehman

***Abstract*—Ship target segmentation in infrared scenes has always been a hot topic, since it is an important basis and prerequisite for infrared-guided weapons to reliably capture and recognize ship targets in the sea level background. However, given the small target and fuzzy boundary characteristics of infrared ship images, obtaining accurate pixel-level labels for them is hardly achievable, which brings difficulty to train segmentation networks. To improve the segmentation accuracy of infrared ship images, we propose a two-stage domain adaptation method for infrared ship target segmentation, where the segmentation model is trained using visible ship images with clear target boundaries. In this case, the source domain is the labeled visible ship images, while the target domain is the unlabeled infrared ship images. Specifically, in the first stage, we use an image style transfer network to convert the infrared ship images into those with visible light style, so that the visual disparity between the two domain images can be reduced. Next, the visible, infrared and converted infrared images are input into the Deeplab-v2 segmentation network for training, thereby obtaining the initial network weights. At this time, random attention modules are added separately to the low- and high-level spaces of Deeplab-v2, in order to improve its feature extraction capability. In the second stage, we mix the visible and infrared images through region mixing to acquire the mixed domain images, as well as their corresponding labels. Subsequently, Deeplab-v2 is further trained using the mixed domain images to attain better segmentation accuracy. Experimental results on both the home-made visible-infrared ship image dataset and the public infrared image dataset are superior to those existing mainstream methods, demonstrating its effectiveness.**

***Index Terms*—Domain adaptation, ship target segmentation, two-stage, style transfer, attention mechanism.**

I. Introduction

OCEANS play a vital role in the social progress and development of all countries in the world. Ships, as a major carrier of maritime transportation, play an important role in coastal safety monitoring [1]–[3]. Marine monitoring is inseparable from accurate and efficient segmentation of ship targets. Infrared thermal mapping technology works by detecting changes in the infrared radiation caused by differences in target temperature and radiation [4], [5]. It can work during daytime and at night. Therefore, precise infrared ship target segmentation is desired, and this is the motivation of our work.

However, low signal-to-noise ratio and fuzzy target boundaries make it difficult to accurately label infrared ship target images at the pixel level [6]–[8]. Meanwhile, visible ship target images have high signal-to-noise ratio and clear target boundaries, and those acquired in fine weather contain richer information, whose target structures have more distinct features and whose labeled images are easier to acquire. Hence, an effective method for addressing the difficulty of infrared image labeling is to apply the segmentation network trained with visible images to infrared images. Common practice is to use labeled visible images to train the segmentation network, which is then directly applied to the segmentation of infrared images. However, there exists a domain shift phenomenon due to the different distributions of visual features in the two domains [9]–[11], so that good segmentation accuracy of the extant segmentation network cannot be guaranteed on the infrared image datasets [12]. With the development of transfer learning, unsupervised domain adaptation methods have been adopted by researchers [13]–[15].

As shown in Figure 1, unsupervised domain adaptation aims to transfer knowledge from labeled source domain data to unlabeled target domain data [16], [17]. This method achieves domain alignment by learning the feature distribution between the source and target domains, thereby reducing domain shift [18]–[20]. Unsupervised domain adaptation has been widely applied in many fields, such as remote sensing image analysis [21]–[29], cross-modality medical image analysis [30]–[33], street scene images semantic segmentation [34]–[38], etc. Existing unsupervised domain adaptation methods can be roughly classified into two types: the input space-based methods and the output space-based methods [39].

With the input space-based methods, image processing or style transfer network is exploited to narrow the style difference between two domains before inputting images into the segmentation network [40]. For instance, Hoffman et al. proposed a domain adaptation approach based on cyclic consistency [34]. Hong et al. put forward a domain adaptation method based on conditional generative adversarial network [35]. Wu et al. proposed a single-stage unsupervised domain adaptation network for nighttime image segmentation with daytime images [36]. Yan et al. proposed a threshold-adaptive unsupervised domain adaptation model to dynamically optimize individual samples to obtain higher segmentation accuracy [37].

The output space-based domain adaptation methods aim to align the inter-domain data distributions at the feature and output layers of the segmentation network [41]. For instance, Hoffman et al. put forward an unsupervised domain adaptation

T. Zhang, H. Shen, Z. Liu are with the Faculty of Information Technology, Beijing University of Technology, 100124 Beijing, China (e-mail: zhangting@bjut.edu.cn, hj_shen0107@emails.bjut.edu.cn, zhaoying.liu@bjut.edu.cn). (Corresponding author: Zhaoying Liu)

S. Rehman is with the School of Sciences, Engineering and Environment, University of Salford, Manchester, UK (email: s.rehman15@salford.ac.uk)

Y. Li is with the School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin, China (e-mail: liyujian@bjut.edu.cn).

O. Rehman is with the Department of Electrical Engineering, Sarhad University of Sciences and IT, Pakistan (e-mail: obaid.ee@suit.edu.pk).

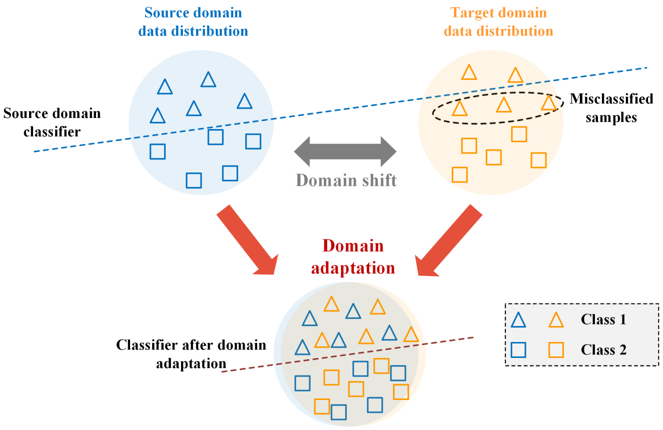
Source domain data distribution
Target domain data distribution
Source domain classifier
Misclassified samples
Domain shift
Domain adaptation
Classifier after domain adaptation
△ Class 1
□ Class 2

Fig. 1. The technology of unsupervised domain adaptation.

segmentation model based on fully convolutional network [38]. Tsai et al. proposed an unsupervised domain adaptation segmentation model based on adversarial learning [42]. Luo et al. proposed a category-level domain adaptation model [43], where the segmentation network was first used to extract image features. Then, the discriminator was used to align features. Vu et al. developed a domain adaptation method based on adversarial entropy minimization [44], where the entropy map was used to optimizing the segmentation network. Chen et al. proposed a domain adaptation approach based on the maximum squared loss, where the maximum squared loss was adopted to optimize the network [45]. Truong et al. put forward a domain adaptation method based on bijective maximum likelihood loss [46]. With this method, an initial segmentation model was obtained on the source domain through supervised training. Then, unsupervised training was performed on the target domain, and model training was accomplished using the BiMaL loss function.

The extant input space- and output space-based methods reduce the differences in image style either via image processing or style transfer networks, or perform adversarial learning on the output layer of segmentation networks with the utilization of discriminator network. They separately reduce the inter-domain data differences in varying spaces. Besides, these methods are all proposed for synthetic–real image datasets, while rarely for the cases in which the source domain comprises visible images and the target domain comprises infrared images. For the visible–infrared ship images, due to the large difference in their image style, the domain shift is large and the feature correlation is poor. When the foregoing methods are applied to the visible–infrared ship datasets, the segmentation accuracy decreases sharply. The main problems are three-folds: (1) Firstly, in case the image modalities of two domains are quite different, there lacks effective strategy for narrowing the style difference between the two domains; (2) Secondly, the segmentation network lacks ability to extract image features sufficiently, making the output space still not aligning enough; (3) Thirdly, the correlation information between images in two domains is ignored.

To solve the above problems, we present a two-stage domain adaptation method for infrared ship target segmentation (T-DANet). Initially, a style transfer network is designed to convert infrared ship images to those in visible image style, so that the inter-domain appearance differences in the input space can be reduced. Next, to reduce the inter-domain feature differences in the output space, a random attention module is constructed and added to the lower- and higher-level spaces of segmentation model respectively, enabling the segmentation network to extract richer features. Finally, to fuse the correlation information between images in different domains, a random inter-domain image splicing method is designed to further enhance the accuracy of segmentation network.

To summarize, the main contributions of our work are as follows:

- A two-stage domain adaptation method for infrared ship target segmentation (T-DANet) is proposed, as well as its corresponding training algorithm. It narrows the appearance differences between visible and infrared images with the style transfer network in the input space, and obtains the domain invariant features in the output space. By these two operations, it aims to improve the segmentation accuracy of infrared ship images;
- An image style transfer network is constructed. It reduces the appearance differences between visible and infrared ship images by converting the infrared images into visible ones;
- A random attention module is established. It enhances the feature extraction capability of segmentation network by enabling extraction of more image feature information.
- An inter-domain image splicing method is designed. By mixing the source domain images with the target domain image with source domain style, it can acquire the image correlation information between two domains, which helps improve the segmentation accuracy of target domain images;
- The proposed method attains better results than the existing mainstream methods on both the self-prepared infrared ship dataset VI-Ship and the public infrared dataset RGB_T. Further, the importance of the whole model and key submodules are also verified through substantial ablation experiments.

The rest of the paper is organized as follows. Section 2 briefly introduces the related work, Section 3 describes in detail the infrared ship segmentation method proposed in this paper, Section 4 gives the experimental results on the infrared ship dataset and the public dataset, as well as their related analysis, and the last section concludes the work of this paper.

II. Related Work

Our work is related to three major tasks: unsupervised domain adaptation, image style transfer and attention mechanism.

A. Unsupervised domain adaptation

Unsupervised domain adaptation is an effective way of solving the domain shift problem [47], It can be roughly classified into two types: the input space-based domain adaptation methods and the output space-based domain adaptation methods [39].

The representative model of input space-based domain adaptation is CycADA [34]. It uses two style transfer networks in the input space to convert the target domain images into the source domain style and vice verse, respectively. CGAN [35] simultaneously inputs the source domain images and a conditional generative branch into the generator, and fuses the low- and high-level features of the generator to increase the feature diversity. DANNet [36] aligns the appearance distribution of two domains by utilizing an image reillumination network in the input space, thereby allowing more similar distribution of image intensities. Although it is effective for the domain adaptation of daytime–nighttime image datasets, in the case of visual–infrared image datasets, a preferable effect can hardly be attained due to the more complicated domain difference. TUFL [37] introduces a threshold-adaptive focal loss in the input space to optimize gradients. Through adaptive adjustment of threshold value in the loss function, the model can better adapt to the data distribution in the target domain, so that its performance is improved.

The representative model of output space-based domain adaptation is AdaptsegNet [42]. It proposes a multi-level adversarial network to separately perform adversarial learning on different level features of segmentation network, thereby reducing the domain shift. On this basis, CLAN [43] adds a category alignment branch to the discriminator of adversarial network for aligning the category-level distribution of source and target domains. ADVENT [44] adopts adversarial entropy minimization in the output space. On the one hand, minimization training is performed directly on the entropy map generated in the output space. On the other hand, the entropy map of predictions is sent into the discriminant network for generative adversarial training. MaxSquare [45] handles the class weight imbalance problem by using image-level weight factor in the output space. BiMaL [46] measures the efficiency of model learning by introducing an unaligned domain scoring into the output space. On the basis of minimizing the adversarial loss, it proposed a bias loss to map the network to a potential space, thereby improving the alignment of the two domains. Generally, the above output space-based domain adaptation methods lack operations for reducing the inter-domain image appearance differences, resulting in their low segmentation accuracy.

All of the aforementioned input or output space-based unsupervised domain adaptation approaches merely reduce domain shift in one space, limiting their performance. In this study, we design methods separately for reducing domain differences in the input and output spaces, and acquire the correlation information of images in two domains through inter-domain image mixing, with a view to further improving the segmentation accuracy.

B. Image Style transfer

Image style transfer refers to a process in which the content of one image (content image) is combined with the style of another image (style image) to produce a new image (generated image). This process often requires retention of the content image's semantic and structural information, as well as simultaneous capturing of the style image's texture and color information. As a common technique used in unsupervised adaptation semantic segmentation tasks, image style transfer converts the images of target domain (or source domain) into a style resembling that of the source domain (or target domain), thereby reducing the appearance difference between the two domain images.

Depending on whether a one-to-one correspondence between the images in source and target domains, the existing style transfer networks can be classified into paired transfer and unpaired transfer networks [48]. The paired style transfer networks refer to those whose training requires image pairs of the source and target domains. For example, pix2pix [49] converts the source domain images into the target domain ones via a conditional generative adversarial network. Neural Style Transfer [50] extracts the content and style features of images using a pretrained convolutional neural network (CNN). For each pair of input content and style images, the perception loss function is optimized by iteratively updating the pixel value of output images. On this basis, Fast Neural Style Transfer [51] iteratively updates the parameters of pretrained CNN by optimizing the perception loss function, so that the style transferred images can be quickly generated through a single forward propagation of content images with the trained network.

The unpaired style transfer networks refer to those whose training not requires image pairs. With this method, two generative adversarial networks are constructed first to achieve bidirectional conversion between the source and target domain images, and then different loss trainings are applied. For example, CycleGAN [48] proposes cyclic consistency loss to guarantee the reversibility of inter-domain mapping. On the basis of cyclic consistency loss, DualGAN [52] uses the Wasserstein distance as an adversarial loss. Later on, DiscoGAN [53] adds a reconstruction loss to ensure that there is little difference between the pre-conversion and converted images.

Although the above methods can achieve transfer of any image style, the quality of generated images is often not high, leading to the occurrence of image content loss. In this paper, we design an unpaired style transfer network. The designed network can generate high-quality infrared images of ships in visible style without the loss of ship targets.

C. Attention mechanism

An attention module, as a data processing method, allows a model to selectively focus on what is important while ignoring what is irrelevant during the processing of input data. According to different parts of the activated feature map, attention mechanisms can be roughly classified into the spatial and channel types [54].

Spatial attention refers to weighting every position of a feature map, which allows a model to focus more on the areas of value. For instance, Spatial Transformer Networks [55] enhance the robustness of geometric transformations like image scaling, rotation and translation through spatial transformation of input images. Non-local [56] obtains the

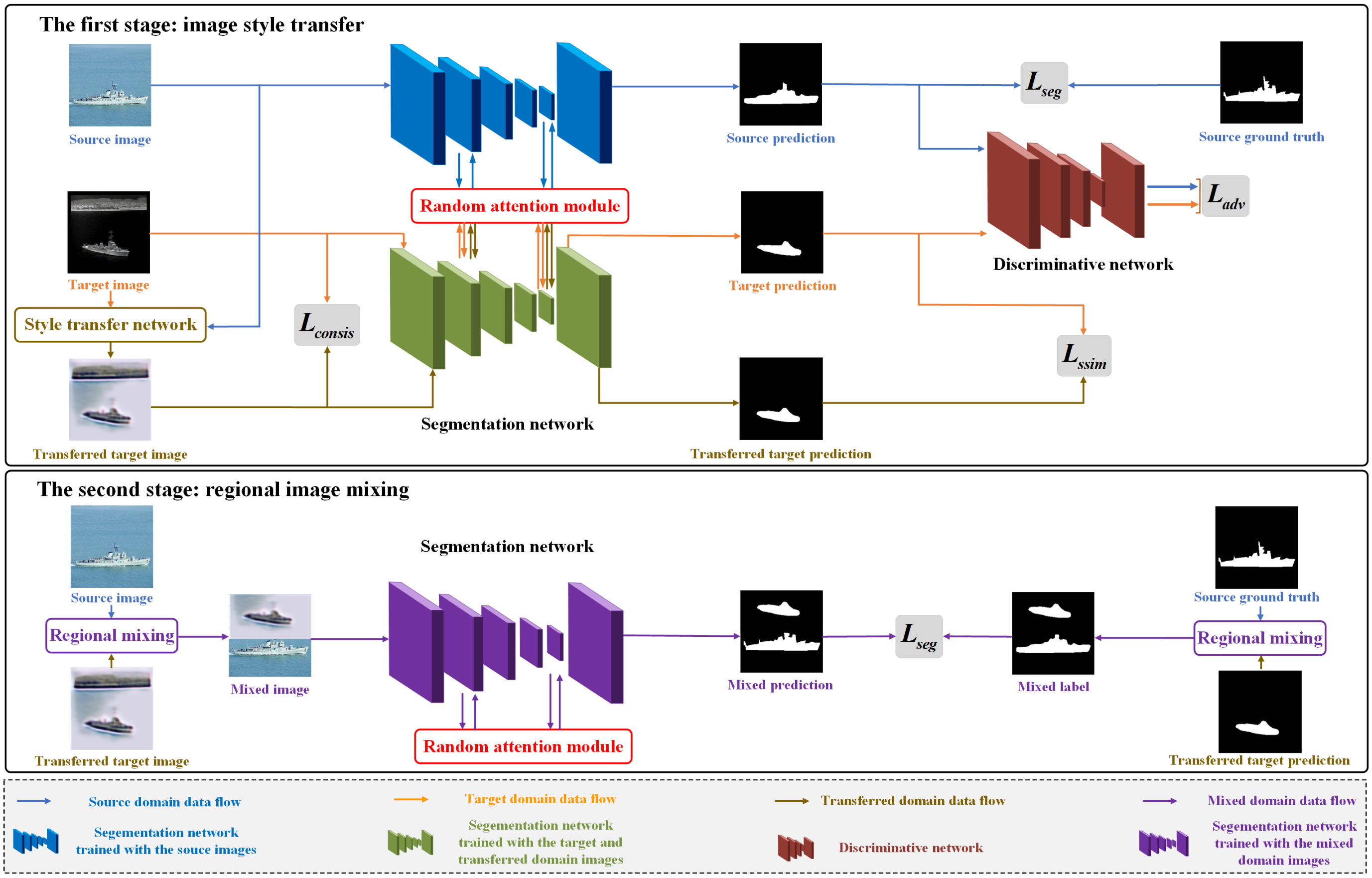

Fig. 2. The overall network structure. In the first stage, the infrared ship images are transferred to visible ones with an image style transfer network. The segmentation network is trained with the source images and the target images, separately. There are four loss functions in this stage: the adversarial loss, the consistency loss, the segmentation loss and the structural loss. In the second stage, the transferred images are mixed with visible ones to get the mixed domain images, which are used to continue to train the segmentation network. The cross-entropy loss is the loss function in this stage.

feature vector of corresponding position in the output feature map by weighting the feature vectors of all other positions. SimAM [57] is a parameter-free attention module, which first generates three-dimensional attention weights for each layer of feature map output by CNN, and then finds out the importance of each position in the feature map by optimizing an energy function. However, the above spatial attention mechanisms are computationally expensive, so that the model training becomes more time-consuming.

Channel attention refers to weighting each channel of a feature map, which allows a model to focus on more meaningful features. For instance, SENet [58] automatically learns the channel weights first, and then extracts features using weighted channels, thereby enhancing useful feature channels and suppressing useless feature channels, improving the network's representation capability. Through point convolution, MS-CAM [59] extracts the local and global channel attention weights first, then enhancing the expressiveness of features and suppressing the redundant information. ECA-Net [60] achieves local cross-channel interactions through one-dimensional convolution, improving the feature expressiveness. However, the above channel attention mechanisms are unable to capture the information in the spatial dimension, resulting in the limited improvement of the model feature extraction capability.

In this paper, we design a random attention module. It randomly selects a certain region of feature map for weighting, thus overcoming the heavy burden of global weighting operation. We add the random attention module separately in the low- and high-dimensional spaces of the segmentation network, enabling the network to fully extract the low- and high-level features of images, which in turn enhances the feature extraction capability of the network and improves the segmentation accuracy.

III. Method

The architecture of the proposed T-DANet model is displayed in Figure 2. As is clear, the first stage involves the image style transfer, during which the infrared ship images are converted into visible style images first with the utilization of style transfer network. Then, the style transfer network and segmentation network are trained using the converted images together with labeled visible light images, thereby obtaining the initial weights of the segmentation network. During the second stage, the image regional mixing stage, the visible light images are first subjected to regional mixing with the infrared ship images that are converted into visible style. Then, using the mixed domain images, the segmentation network is trained in a fully-supervised way to ultimately obtain the network weights. After the two stages, we verify the performance of segmentation network on the test set. The rest of this section introduces the details of each component and the loss function trained the network.

A. Style transfer network

To reduce the appearance difference between images in the two domains, a style transfer network is designed. Extracting the features of infrared ship images is difficult, since they have low signal-to-noise ratio and contain much noise. Contrastively, visible light images have rich details and clear target boundaries. Hence, we use a style transfer network to convert infrared ship images into visible light images. Figure 3 displays the structure of the network. As is clear, the network first extracts the style information of visible ship images, followed by extraction of the content information of infrared ship images. Finally, it combines and upsamples them to obtain new infrared ship images in visible style.

Concretely, at the first step, the ship images from source and target domains are obtained for separately extracting the style and content information. The input channels for the two domain images are all 3 in quantity and 256×256 in size. For the source domain visible images, their style information is obtained sequentially via three successive convolutional layers, a global average pooling layer and two fully-connected layers. For the infrared images in target domain, their content information is obtained via three successive convolutional layers and four successive residual blocks [61]. Then, the style information is input into the multilayer perception to obtain a set of adaptive instance normalization (AdaIN) [62] parameters. The content information is then processed by residual blocks with AdaIN layers. The AdaIN layer can be seen as a style normalization that converts the features of an arbitrary style image into the same distribution, thus enabling arbitrary style transfer [62]. The formula for the AdaIN layer is Eq. (1). Finally, a new feature map is generated by fusing the content information of infrared ship images with the style information of visible ship images, which is upsampled to the size of input feature map to ultimately obtain the target domain images in visible style. Finally, a new feature map is generated by fusing the content information of infrared ship images with the style information of visible ship images, which is upsampled to the size of input feature map to ultimately obtain the target domain images in visible style.

$$\mathrm{AdaIN}(\boldsymbol{f}_{con}, \boldsymbol{f}_{sty}) = \sigma(\boldsymbol{f}_{sty}) \left(\frac{\boldsymbol{f}_{con} - \mu(\boldsymbol{f}_{con})}{\sigma(\boldsymbol{f}_{con})} \right) + \mu(\boldsymbol{f}_{sty}) \tag{1}$$

where $\boldsymbol{f}_{con}$ is the feature of content image, $\boldsymbol{f}_{sty}$ is the feature of style image, μ and σ denote the mean and variance, respectively.

B. Segmentation network

After obtaining the style transferred images, we input them and the visible images into the segmentation network to acquire the predicted output images in two domains. Figure 4 depicts the structure of the segmentation network.

As is clear, for input images with the size of 256×256×3, their features are firstly extracted using five convolutional layers. They are the first 5 convolutional layers of ResNet101 [61]. Secondly, the extracted features are input to an atrous spatial pyramid pooling module [63] to capture image contextual information at multiple scales. The pooling sampling rates are {6,12,18,24}, respectively, and four feature maps with size of 4×4×1024 are obtained. Finally, the four feature maps are added up along the channel dimension and the prediction maps are obtained by upsampling, size 256×256×2.

To enhance the feature extraction capability for small target regions and capture details of target boundaries of the segmentation network, a random attention module branch is added separately to its second and fifth layer. The random attention module can help the segmentation network to focus on small local regions of interest, and by enhancing these local features, the performance of small targets can be improved. This module can also help the segmentation network focus and capture some subtle boundary features that are easy to be filtered, which is also helpful for sharpening the fuzzy boundary. Figure 5 depicts the structure of the random attention module.

According to Figure 5, the specific process of this module is as follows: for the input feature map of current convolutional layer, a certain region is randomly selected by taking a random number between 1/2 and 3/4 of the size of the current feature map. Then, the feature map is input into the global pooling module to reduce the map size. Next, two convolutional layers with a convolutional kernel size of 1×1 are connected. After upsampling operation, pixel-by-pixel multiplication is accomplished with the original feature map to generate a new feature map. Thereafter, the new and original feature maps are added pixel-by-pixel to ultimately obtain the enhanced feature map. The enhancement process of the i-th (either 2 or 5) convolutional layer in segmentation network can be expressed as:

$$Output\,(i) = \boldsymbol{x} + Up\left(Conv_{1\times 1}\left(Conv_{1\times 1}\left(Gp\left(\boldsymbol{x}\right)\right)\right)\right)\boldsymbol{x} \tag{2}$$

where $\boldsymbol{x}$ represents the input feature map, $Gp(\cdot)$ refers to the global pooling operation. $Conv_{1\times 1}(\cdot)$ denotes the 1×1 convolutional layer in random attention module and $Up(\cdot)$ refers to the upsampling operation.

C. Discriminative network

The primary role of discriminative network is to distinguish the source domain segmentation results from the target domain segmentation results. Figure 6 displays the structure diagram of the discriminative network.

As shown in Figure 6, the discriminative network is a five-layer convolutional network with the convolutional kernel sizes of all 4×4, the step sizes of all 2 and the channel numbers of {64,128,256,512,1}, respectively. Then, upsampling operation is performed to restore the output feature map to the input size and return the discrimination result, 1 is for the source domain, and 0 is for target domain.

D. Two-stage training with mixed domain images

This subsection describes the connection between the two stages. The two stages are trained in sequence. In the first

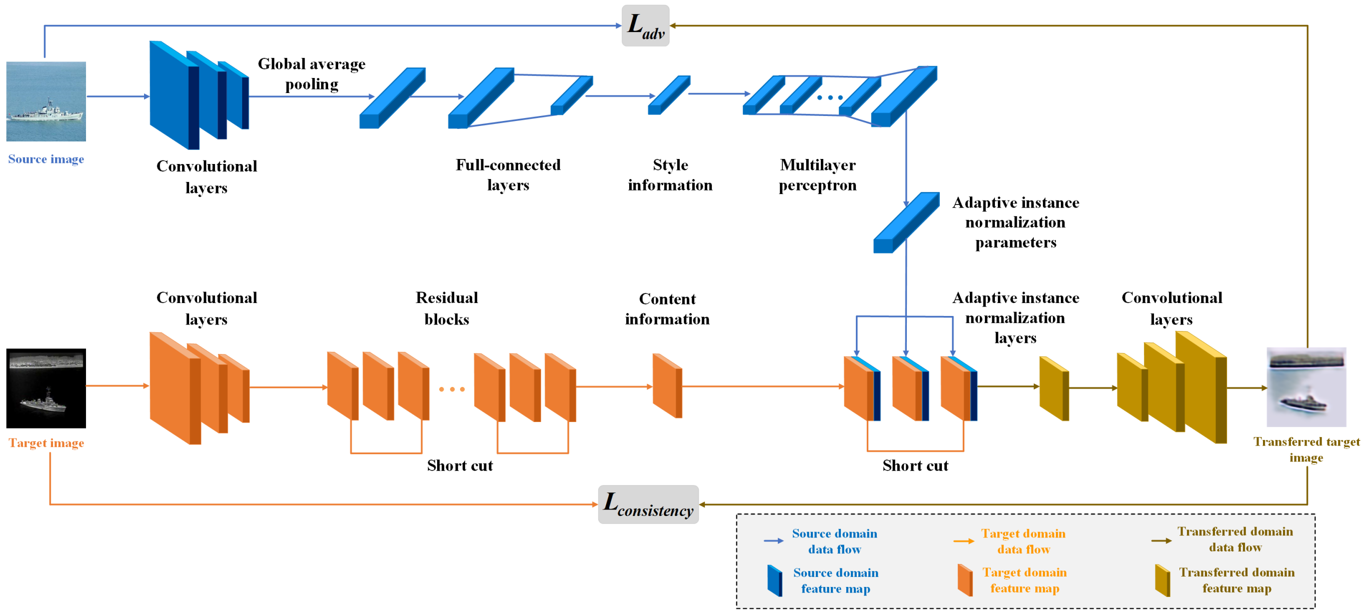

Fig. 3. The style transfer network structure

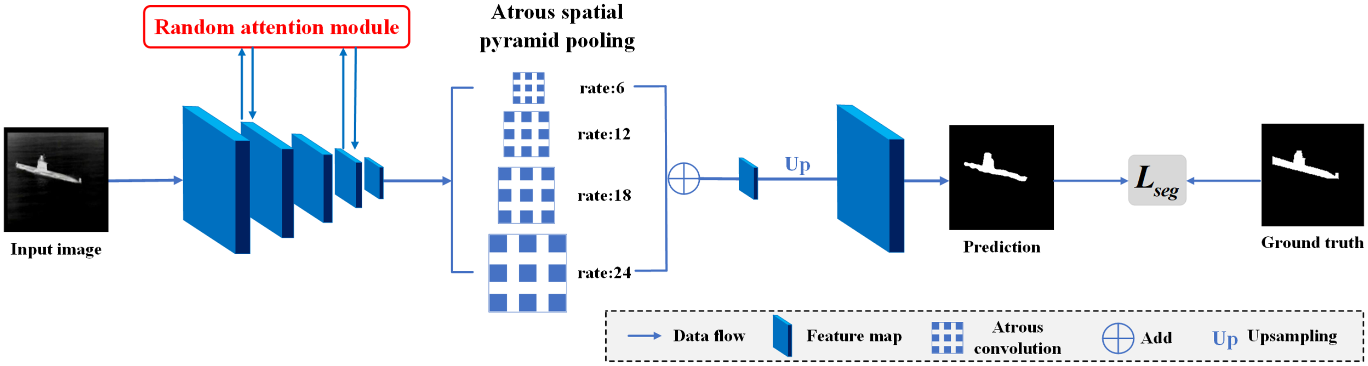

Fig. 4. The structure of the segmentation network

stage, the input of the segmentation network comprises three parts: visible ship images, original infrared ship images and infrared ship images in visible style. Meanwhile, the output comprises their corresponding prediction results. Theoretically, the predictions by segmentation network are identical for two images with the same target but different styles. Thus, the outputs of infrared images with both styles are used to constrain the consistency in the target domain, and to enhance the stability of segmentation network.

In the second stage, the segmentation network is further trained by exploiting the spatial layout similarity between the two domain images. The segmentation network trained in the first stage can provide pseudo labels for the shifted target domain images in the second stage. Through regional mixing technique, the source domain images are mixed with the target domain images in source domain style. Their labels are subjected to the same operation to generate new mixed domain images and labels. The new image–label pairs are used to train the segmentation network, thereby improving the segmentation accuracy of target domain images. Figure 7 describes the image mixing process.

The motivation of randomly cropping half of the image is to fuse the spatial layout information between the two domains while preserving sufficient target information. There are three steps to select the subregions. Specifically, for a given infrared image, firstly, we randomly generate the starting point of the cropping rectangle; secondly, we determine one of the horizontal lines as the starting point's row; thirdly, we compute 1/2 of the height of image, and crop out the half subregions. The reason we select half of the image is that we can cover the majority of the targets. If we use smaller ratios, for example, 1/4, it may lose the target. Otherwise, it may lead to an increase in computational expense.

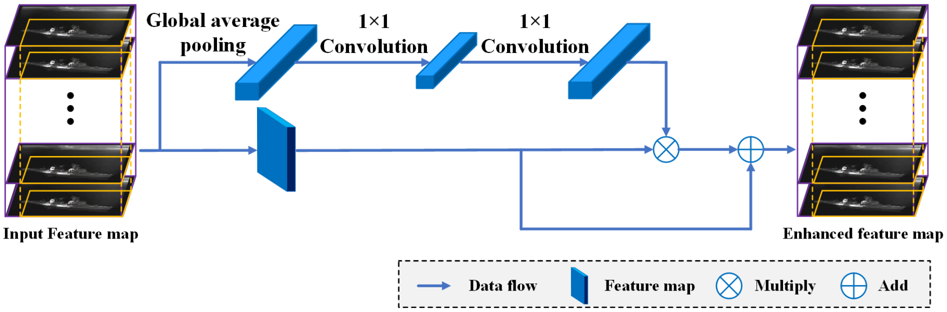

Fig. 5. The structure of random attention module

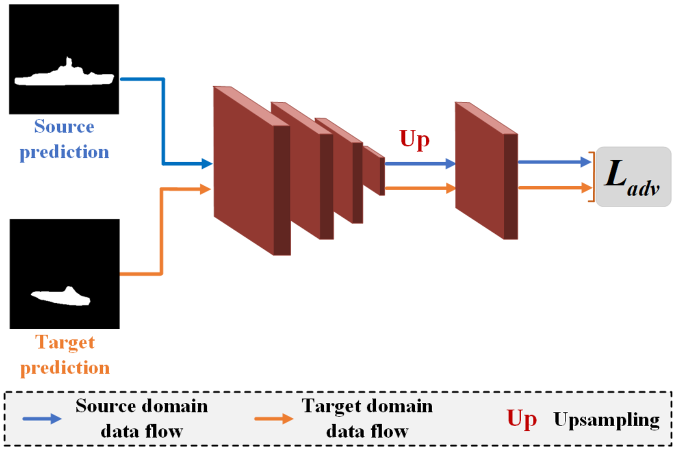

Fig. 6. The structure of the discriminative network

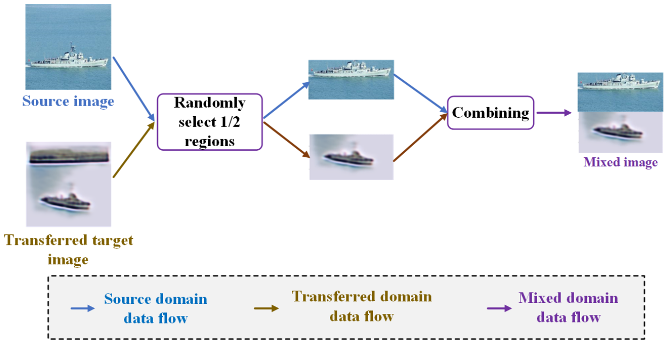

Fig. 7. The mixed mode structure

E. Loss function

This subsection describes the loss functions used to train the T-DANet. In the first stage, the original infrared ship images are transferred to the infrared images in visible style first by using a style transfer network. For this network, an adversarial loss and a consistency loss are used. The adversarial loss makes it difficult for the discriminative network to tell whether its input is from visible or infrared images. The specific formula for adversarial loss function is as follows:

$$
\begin{aligned}
L_{GAN}\left(D\right) = &-\sum_{l=1}^{N}\sum_{h,w}\log\left(D\left(\boldsymbol{x}_s^l\right)\right) \\
&-\sum_{l=1}^{N}\sum_{h,w}\log\left(1-D\left(G\left(\boldsymbol{x}_t^l\right)\right)\right)
\end{aligned} \tag{3}
$$

where $\boldsymbol{x}_s^l$ represents the visible ship images in source domain, and $\boldsymbol{x}_t^l$ represents the l-infrared ship images in target domain, N is the number of the train samples. G stands for the network generated during the recombination of content and style information, and D stands for the discriminative network.

The consistency loss function ensures that the content information of infrared ship images before and after transfer are as consistent as possible. Its specific formula is as follows:

$$
L_{consis}(D) = \sum_{l=1}^{N}\sum_{h,w}\left\|G(\boldsymbol{x}_t^l) - \boldsymbol{x}_t^l\right\|_1 \tag{4}
$$

where $\boldsymbol{x}_t^l$ represents the infrared ship images in target domain, and N is the number of training samples.

For the segmentation network, only the visible ship images have truth labels. Visible ship images have, on the one hand, fully supervised cross-entropy loss with their own labels and, on the other hand, adversarial loss with infrared ship images. Besides, the prediction results of infrared ship images in visible style are theoretically consistent with those of original infrared ship images. Hence, we use the structural similarity loss to constrain the predictions of these two target domain images.

The cross-entropy loss is defined as:

$$
L_{seg}\left(F_s\right) = -\sum_{l=1}^{N}\sum_{h,w}\sum_{c\in C}\boldsymbol{Y}_s^{l(h,w,c)}\log(\boldsymbol{P}_s^{l(h,w,c)}) \tag{5}
$$

where $\boldsymbol{P}_s^l$ represents the predicted image generated by the l-th image in source domain via the segmentation network, $\boldsymbol{Y}_s^l$ refers to the truth label corresponding to the l-th image, and N is the number of training samples. h , w and c denote length, width and class, respectively. F_s stands for the segmentation network in the source domain.

For the adversarial loss, in order to improve the segmentation accuracy of hard-to-classify pixel points, information entropy is introduced. The smaller the information entropy value of an image is, the higher credibility it has; and vice verse. Therefore, we give larger weights to the pixels with higher entropy values of predicted images in the target domain. The entropy value of a target domain prediction image is defined as:

$$E_t = -\sum_{h,w}\sum_{c\in C} \boldsymbol{P}_t^{(h,w,c)} log(\boldsymbol{P}_t^{(h,w,c)}) \tag{6}$$

where $\boldsymbol{P}_t$ represents the predicted output image generated by the image in target domain via the segmentation network. h , w and c denote length, width and class, respectively.

At this time, the adversarial loss function is formulated as:

$$L_{adv}\left(D\right) = -\sum_{l=1}^{N} E_t \sum_{h,w} log(D(\boldsymbol{P}_t^l)^{(h,w,1)}) \tag{7}$$

where $\boldsymbol{P}_t^l$ represents the predicted output image generated by the l-th image in target domain via the segmentation network, N is the number of training samples, h and w denote length and width, respectively.

The loss function of the segmentation network is as follows:

$$L_{sn}\left(F, D\right) = L_{seg}\left(F_s\right) + \lambda_{adv} L_{adv}\left(D\right) \tag{8}$$

where λ_{adv} is used to equalize the two loss functions.

The structural similarity loss is described as:

$$L_{ssim} = \sum_{l=1}^{N}\sum_{h,w} \left\| 1 - (SSIM(\boldsymbol{P}_{t'}^l, \boldsymbol{P}_t^l)) \right\|_1 \tag{9}$$

where $\boldsymbol{P}_{t'}^l$ and $\boldsymbol{P}_t^l$ respectively denote the predictive outputs of infrared images after and before transfer, and N is the number of training samples. Structural Similarity Index (SSIM) [64] refers to the structural similarity between two images, which is specifically expressed as:

$$SSIM = \frac{\left(2\mu_{\boldsymbol{x}}\mu_{\boldsymbol{x}} + c_1\right)\left(2\theta_{\boldsymbol{xy}} + c_2\right)}{\left({\mu_{\boldsymbol{x}}}^2 + {\mu_{\boldsymbol{x}}}^2 + c_1\right)\left({\theta_{\boldsymbol{x}}}^2 + {\theta_{\boldsymbol{x}}}^2 + c_2\right)} \tag{10}$$

where $\mu_{\boldsymbol{x}}$ and $\mu_{\boldsymbol{y}}$ denote the average pixel values of images $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, $\theta_{\boldsymbol{x}}$ and $\theta_{\boldsymbol{y}}$ represent the variances of pixel values for images $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and c_1, c_2 are different constant values used for maintaining stability.

The overall loss function in the first stage is formulated as follows:

$$\begin{aligned} L_{stage1} = \alpha_1 L_{GAN}\left(D\right) + \alpha_2 L_{consis}\left(D\right) \\ + \alpha_3 L_{sn}\left(F_s, D\right) + \alpha_4 L_{ssim} \end{aligned} \tag{11}$$

where α_1, α_2, α_3 and α_4 denote the weight coefficients between various measured loss functions, whose values are 0.1, 1, 0.1 and 0.1 by experience, respectively.

In the second stage, the segmentation network of the first stage is trained in the fully-supervised way using the acquired mixed domain images and corresponding label images. During training, the cross-entropy loss function is used, which is specifically formulated as:

$$L_{stage2} = -\sum_{l=1}^{N}\sum_{h,w}\sum_{c\in C} \boldsymbol{Y}_{mixed}^{l(h,w,c)} \log\left(\boldsymbol{P}_{mixed}^{l(h,w,c)}\right) \tag{12}$$

where $\boldsymbol{Y}_{mixed}^l$ represents the label corresponding to the mixed domain image, $\boldsymbol{P}_{mixed}^l$ refers to the mixed domain image, N is the number of training samples. h , w and c denote length, width and class, respectively.

Algorithm 1 Learning algorithm of T-DANet

```
Input: source dataset $I_S$, target dataset $I_T$, number of iterations for the first stage $K1$, number of iterations for the second stage $K2$;
Output: network weights;
  [The first stage of training]
  $i = 0$
  while $i <= K1$ do
    for $j$ to $range(len(I_S))$ do
      Input the images from the two domains $I_S$ and $I_T$ into the style transfer network, and obtain new infrared ship images with visible style;
      Input the original images from the two domains and the transferred images into the segmentation network, and obtain their respective prediction results;
      Input the outputs of the two domains into the discriminative network to optimize the adversarial loss;
      Calculate the structural similarity loss between the transferred infrared image and the original infrared image;
      Update the weight parameters of the segmentation network and discriminative network;
    end for
  end while
  [The second stage of training]
  $i = 0$
  while $i <= K2$ do
    for $j$ to $range(len(I_S))$ do
      Generate mixed images by randomly stitching images from the two domains;
      Input the mixed images into the segmentation network to obtain the prediction results;
      Calculate the cross-entropy loss between the prediction result and the ground truth;
      Update the weight parameters of the pre-trained segmentation network;
    end for
  end while
```

F. Learning algorithm

The specific training algorithm of the proposed T-DANet is shown in Algorithm 1. In the first stage, the style transfer network is used to narrow the appearance difference between two domains. Then, the obtained three input images are sent into the segmentation network for generative adversarial training. In the second stage, fully-supervised training is performed using the mixed images and corresponding labels, and the segmentation network is constantly optimized to improve its accuracy.

IV. Experimental results

In this section, we sequentially introduces the datasets, experimental details, evaluation indices and different experimental results of the proposed model, as well as giving the corresponding analysis. All the experiments were realized with the deep learning framework Pytorch [65], the operating

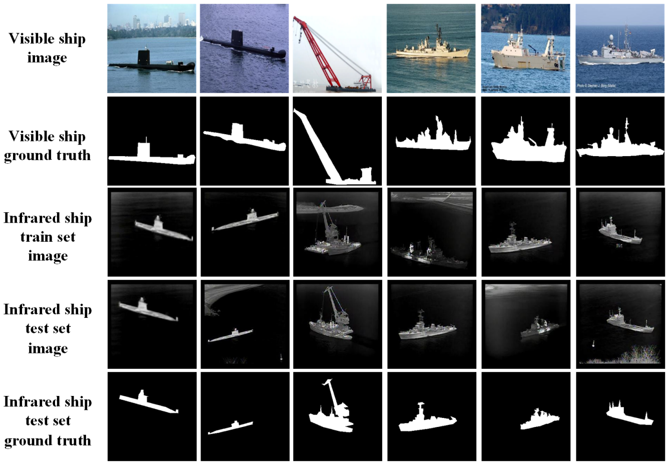
Visible ship image
Visible ship ground truth
Infrared ship train set image
Infrared ship test set image
Infrared ship test set ground truth

Fig. 8. Examples of the VI-Ship dataset images

system is Windows 10, and the GPU model is NVIDIA Tesla K40c.

A. Datasets

To verify the segmentation accuracy of the unsupervised domain adaptation technique on the Visible-infrared Ship (VI-Ship) image dataset, a visible–infrared ship image dataset is created in this study, as shown in Figure 8, where the visible images are collected from the Internet, and the infrared images are obtained by actual photographing. The visible images are taken as the source domain, whereas the infrared images as the target domain. The source domain of the visible–infrared ship dataset contains 1,000 visible images and their labels, all of which are used for training. Meanwhile, the target domain contains 196 infrared ship images.

Prior to the network training, we carry out horizontal flipping, translation, scaling and cropping operations separately on the source and target domain images, to enhance the diversity of images. Ultimately, 5,000 visible ship images and 960 infrared ship images are acquired. The infrared images are divided into training and test sets, and Labelme [66] is utilized to manually label the test images. In the end, the training set consists of 576 images without labeled images, while the test set consists of 384 images, with corresponding label images. All images are 256×256 in size. Figure 8 displays some examples of the images. As is clear, the first and second lines in the figure represent the visible ship images in source domain and corresponding labels, respectively. The third line are the infrared ship training set images in target domain without labels, while the fourth and fifth lines respectively denotes the infrared ship test set images in target domain and corresponding labels.

B. Implementation details

(a) Basic parameters configuration. In the first stage, the style transfer network and segmentation network are trained from scratch. For the style transfer network, we trained it first to get transferred images. Specifically, we use the visible ship images as the source domain and the infrared ship images as the target domain to construct the training set, and train the parameters of the transfer network. The training objectives are to minimize an adversarial loss and a consistency loss, so as to achieve accurate style transfer from the infrared ship target domain to the visible style domain. Its total training rounds is 600, and its training process is optimized via the Adam optimizer [67]. The momentum parameters are set to 0.9 and 0.999, respectively. The initial learning rate of the network is set to 10^{-4}, and the learning rate drops by half every 100 rounds. For the segmentation network, we use DeepLab-v2 [63] as the framework with ResNet-101 [61] as the backbone. The stochastic gradient descent is adopted as the optimization method [68]. The initial learning rate of the network is set to 1.5×10^{-4}. Due to the instability of adversarial learning, its total training epochs is set to 1.5×10^{4}, and its weight is recorded every 1,000 rounds.

In the second stage, we further fine-tune the segmentation network with the mixed domain images based on the first stage. We set the learning rate to be 1.5×10^{-4}, and use the stochastic gradient descent as the optimization method. With this hyperparameter configuration, it is enough to train the segmentation network with 500 epochs.

(b) Selecting α_1, α_2, α_3 and α_4 in loss function of Stage 1. When training the first stage, there are four hyperparameters needing to be determined beforehand in Eq. (11). We select them with three steps. Firstly, we set all of them to be 1 to get the initial segmentation result. Secondly, we set them be 0.5 and 0.1 in sequence, and identify their impact on the segmentation performance. According to the experimental results listed in Table I, we find all of them improved the experimental results. Moreover, α_1 and α_2 have a greater impact on the results than α_3 and α_4. Therefore, we set both α_3 and α_4 be 0.1. Thirdly, we fix α_3 and α_4 be 0.1, and use grid search to select α_1 and α_2 . For these two hyperparameters, we use grid search from 0.01:0.01:0.1 to get proper values. Finally, we select these four hyperparameters be 0.1,1,0.1,0.1, respectively.

TABLE I
THE SEGMENTATION RESULTS OF T-DANET WITH DIFFERENT HYPERPARAMETERS OF α_1, α_2, α_3 AND α_4

α_1	α_2	α_3	α_4	mIoU(%)	Δ mIoU(%)
1	1	1	1	52.63	-
0.5	1	1	1	53.44	**+0.81**
1	0.5	1	1	51.69	**-0.94**
1	1	0.5	1	52.80	+0.17
1	1	1	0.5	52.84	+0.21
0.1	1	1	1	54.12	**+1.49**
1	0.1	1	1	50.94	**-1.69**
1	1	0.1	1	52.96	+0.33
1	1	1	0.1	53.07	+0.44

(c) Selecting of λ_{adv}. In Eq. (8), the parameter λ_{adv} stands for the weight of adversarial loss. It was selected from the candidate set {0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1}. We select the proper value according to their segmentation accuracies in Table II. According to Table II, we finally choose 0.01 as the value of λ_{adv}.

TABLE II
THE SEGMENTATION RESULTS OF DIFFERENT λ_{adv}.

λ_{adv}	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
mIoU(%)	**55.40**	55.21	55.17	55.19	54.93	54.89	54.71	54.63	54.22	54.07

C. Evaluation indices

The evaluation indices are the mean intersection over union (mIoU) [69], the number of parameters of model (Params) and the floating point operations of model (FLOPs). Among them, mIoU represents the ratio of intersection to union between the predicted and actual regions, which is specifically formulated as follows:

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{13}$$

where p_{ii} denotes the number of correct classifications, p_{ij} denotes the number of predictions as category j while actual category is i, p_{ji} denotes the number of predictions as category i while actual category is j, and $k+1$ represents the total number of categories in the dataset.

Params refers to the parameter quantity of the entire network, whose unit is M (Million). FLOPs indicates the number of floating point operations performed during model inference, whose unit is GFLOPs (giga/billion floating point operations).

D. Comparison with the state-of-the-art methods

In this subsection, the proposed method is comparatively analyzed with the existing domain adaptation segmentation methods on the VI-Ship dataset. Table III details the relevant experimental results, where the bold numbers represent the best results. Figure 9 displays the visualization results. Specifically, Figure 10 shows the advantages of our method on ship images with small targets and fuzzy boundaries compared to other methods.

TABLE III
EXPERIMENTAL RESULTS OF DIFFERENT METHODS ON THE VI-SHIP DATASET

	Methods	mIoU(%)	Params(M)	FLOPs(GFLOPs)
Input-based	CycADA [34]	38.44	**20.36**	**24.90**
	CGAN [35]	51.03	25.84	29.82
	TUFL [37]	51.85	45.27	47.79
	DANNet [36]	52.02	44.95	47.74
Output-based	MaxSquare [45]	34.35	40.85	47.75
	FCN WId [38]	48.93	48.18	24.92
	AdaptsegNet [42]	49.51	40.85	47.70
	ADVENT [44]	51.34	40.85	47.71
	BiMaL [46]	51.62	41.38	47.71
	CLAN [43]	52.11	43.06	47.65
Ours	**T-DANet(Stage1)**	55.40	-	-
	T-DANet(Stage2)	**56.69**	53.47	47.85

It can be concluded from Table III and Figure 9 that:

1) The proposed T-DANet (Stage1) and T-DANet (Stage2) attain mIoU of 55.40% and 56.69%, respectively, with the latter showing a 1.29% increase compared to the former. Suggestively, the two-stage mixed domain training is highly effective and practical. There are two reasons: firstly, in the second stage, the input images blend the spatial layout information of both the visible ship images and transferred infrared ship images, thus increasing the inter-domain information correlation between the two datasets; Secondly, by generating the pseudo labels, the infrared ship images for full-supervised training can improve the segmentation accuracy.
2) Compared to the results of input space-based methods, the results of our T-DANet (Stage1) improved by 3.38–16.96%. This indicates that the style transfer network and random attention module are feasible for the T-DANet.
3) Compared to the results of output space-based methods, the results of our T-DANet(Stage1) increased by 3.29–21.05%. This demonstrates that narrowing the domain difference in both the input and output spaces is helpful for improving the segmentation results of infrared ship images.
4) As exhibited in Figure 10, our proposed T-DANet demonstrates more robust segmentation on small and obscured ship targets compared to other methods. For example, CycADA completely fails to identify the tiny ship in the second image. Our approach succeeds owing to the tailored random attention mechanism focusing on small areas. CLAN produces blurred boundaries for the ship in the third image. In contrast, our model yields clearer contours, benefiting from the feature enhancement of attention modules. The comparison verifies the superiority of T-DANet on handling small and blurry infrared targets, which validates the effectiveness of our model design.
5) Regarding the model parameters, the proposed T-DANet (Stage2) has the largest number of parameters, since it uses a style transfer network to convert infrared ship images into those in visible style, leading to the increase of network parameters. Despite the sacrifice of some temporal and spatial efficiencies, the proposed method attains the best segmentation accuracy based on the mIoU.
6) Regarding the computational complexity, the proposed T-DANet (Stage2) achieves comparable FLOPs to other state-of-the-art methods built on similar backbone networks, which is 47.85 GFLOPs. This is because the core segmentation network dominates the computation, while the additional modules like attention bring minor increases. Though there is a sacrifice of some efficiency due to the style transfer network, our model retains equivalence in computation to mainstream approaches. Meanwhile, our method attains superior accuracy, validate by the highest mIoU score.

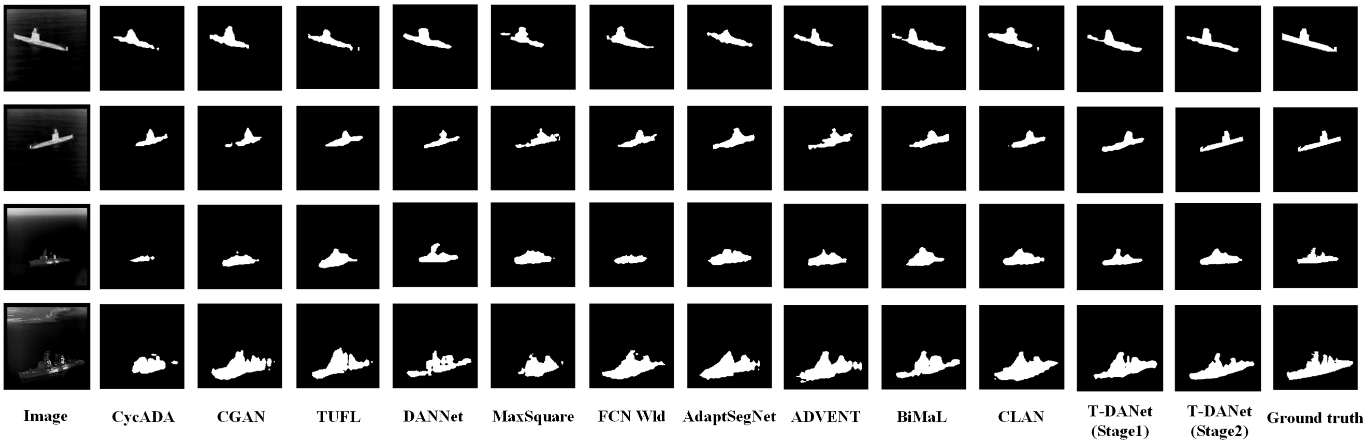

Fig. 9. The experimental results of different methods on the VI-ship dataset

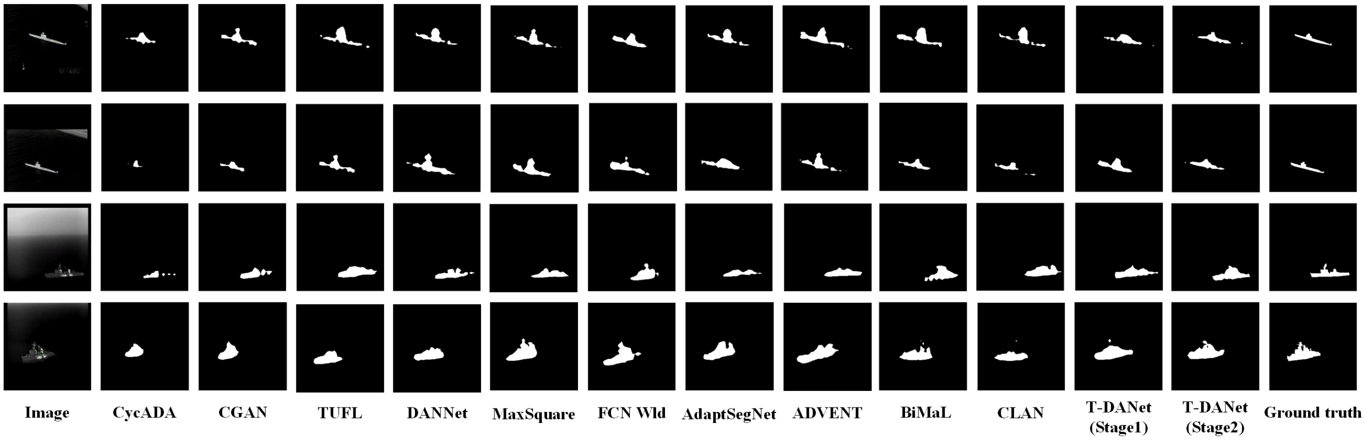

Fig. 10. The experimental results of different methods on the VI-ship dataset small targets

7) In summary, the proposed T-DANet (Stage1) and T-DANet (Stage2) attain higher mIoU, and the images predicted with them have more distinct targets and richer detailed information, suggesting the effectiveness of the proposed method.

E. Effects of style transfer network on segmentation results

This subsection compares the experimental preprocessing results of two domain images by the proposed T-DANet (Stage1) versus the T-DANet with no style transfer , pix2pix [49], Neural Style Transfer (NST) [50], Fast Neural Style Transfer (Fast NST) [51], CycleGAN [48], DualGAN [52] and DiscoGAN [53], respectively. The corresponding models are denoted as T-DANet-NoST, T-DANet-pix2pix, T-DANet-NST, T-DANet-Fast NST, T-DANet-CycleGAN, T-DANet-DualGAN, T-DANet-DiscoGAN, respectively. Table IV details the experimental results, whereas Figure 11 shows the visualized experimental results.

It can be seen from Table IV and Figure 11 that:

TABLE IV
THE INFLUENCE OF THE STYLE TRANSFER ON SEGMENTATION RESULTS

Methods	mIoU(%)
T-DANet-NoST	50.72
T-DANet-pix2pix	53.17
T-DANet-NST	53.42
T-DANet-Fast NST	53.39
T-DANet-CycleGAN	54.06
T-DANet-DualGAN	54.31
T-DANet-DiscoGAN	54.35
T-DANet(Stage1)	**55.40**

1) Compared with the mIoU (50.72%) of T-DANet-NoST model, the proposed T-DANet (Stage1) attains the mIoU of 55.40%, with an increase of 4.68%. This indicates that it is helpful for improving the segmentation accuracy by narrowing the appearance difference between images in two domains.
2) Compared with the results of our T-DANet with different style transfer networks, the segmentation performance of our T-DANet (Stage1) improved by 1.05–2.23%. It demonstrates that our style transfer network is more effective in narrowing the domain gap through photo-realistic image synthesis than others. The reason maybe that our network decomposes the images into domain-independent content information and domain-dependent stylistic information, which enables effective separation of semantic and visual features in images, thus improv-

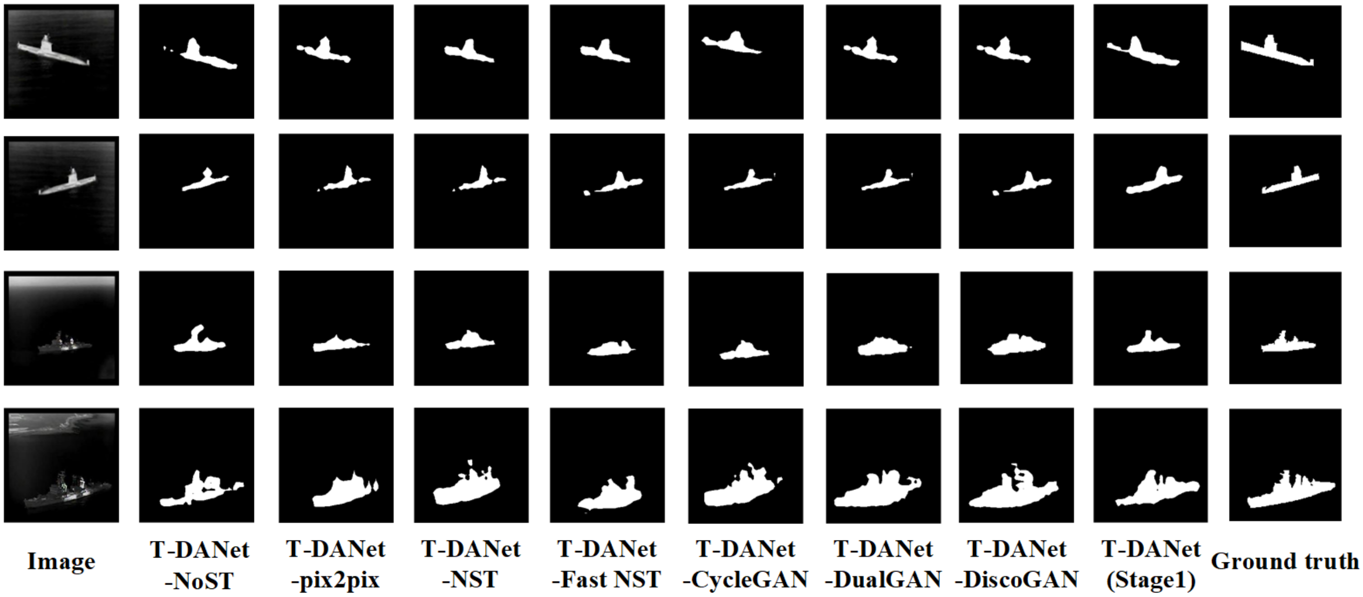

Fig. 11. The influence of the style transfer on segmentation results

TABLE V
THE INFLUENCE OF THE RANDOM ATTENTION MODULE ON SEGMENTATION RESULTS

Methods	mIoU(%)
T-DANet-NoRAM	54.23
T-DANet(Stage1)	**55.40**

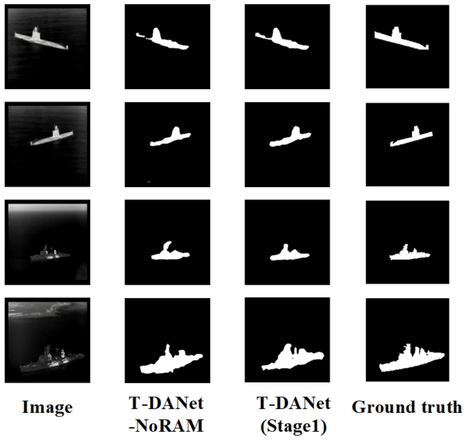

Fig. 12. The influence of RAM on segmentation results

ing the quality and flexibility of the transfer.

3) As is clear from the visualized results, the ship targets preprocessed with T-DANet (Stage1) are more complete than others. While for other segmentation results, they have either lost some details and boundaries of the ships, or introduced some noises.
4) In summary, the proposed T-DANet (Stage1) uses a style transfer network to preprocess images in two domains, it can generate well-transferred images with visible style. By this operation, it can not only narrow the appearance difference between the two domains, but also are helpful for achieving better segmentation results.

F. Effects of random attention module on segmentation results

In this subsection, experimental results of the proposed T-DANet (Stage1) are compared with those of T-DANet with no random attention module (T-DANet-NoRAM) in the segmentation network. Table V details the experimental results, whereas Figure 12 gives the visualized experimental results.

It can be got from Table V and Figure 12 that:

1) When no random attention is added, the T-DANet-NoRAM attains a mean crossover ratio of 54.23%; when random attention is added, the proposed T-DANet (Stage1) attains a mean crossover ratio of 55.40%, showing an increase of 1.17%. Suggestively, random attention is helpful for improving the image feature capturability of segmentation network, thereby elevating the segmentation accuracy.
2) As is clear from the visualized results, the ship targets segmented by T-DANet (Stage1) have clearer contours, indicating that random attention is conducive to capturing the low-level information of ship targets, leading to clearer segmentation boundaries.
3) In summary, adding RAM module to the segmentation network can enhance the feature expressiveness of image information and help the network extract more detail feature information during the training process, thereby improving the segmentation accuracy of the network.

G. Comparisons with other methods on public dataset RGB_T

To verify the generalizability of the proposed method, this subsection compares the target segmentation accuracy of the proposed T-DANet (Stage2) with other methods on the public dataset RGB_T [70]. The dataset contains a total of 1,569

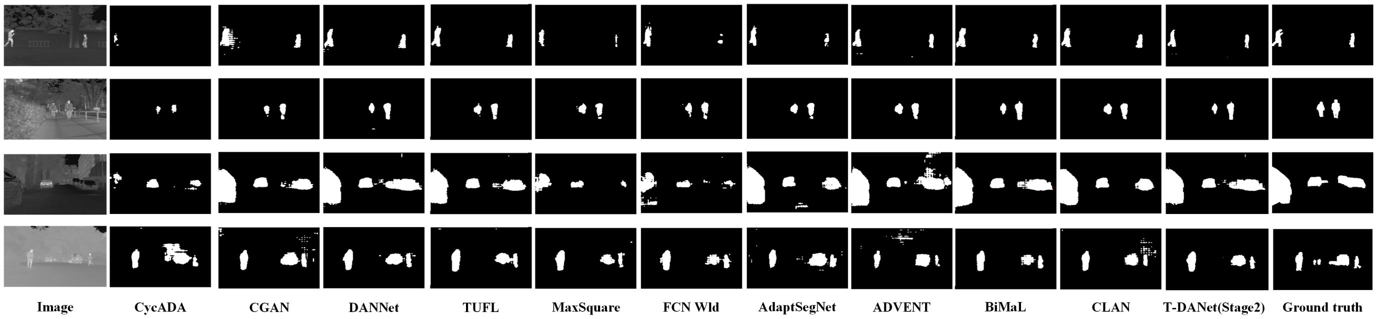

Fig. 13. The segmentation results of different methods on the RGB_T dataset

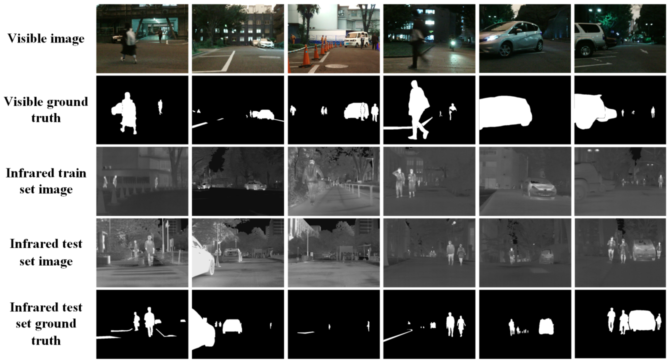

Fig. 14. Examples of the RGB_T dataset images

TABLE VI
EXPERIMENTAL RESULTS OF DIFFERENT METHODS ON THE RGB_T DATASET

	Methods	mIoU(%)
Input-based	CycADA [34]	32.48
	CGAN [35]	39.91
	TUFL [37]	41.72
	DANNet [36]	41.36
Output-based	MaxSquare [45]	35.71
	FCN WId [38]	37.69
	AdaptsegNet [42]	38.32
	ADVENT [44]	40.75
	BiMaL [46]	40.98
	CLAN [43]	40.14
Ours	**T-DANet(Stage2)**	**43.70**

pairs of RGB and infrared scene images, all of which are 480 × 640 in size, with the RGB images as the source domain and the infrared images as the target domain during the training period. The images of the RGB_T dataset section are shown in Figure 14. The first and second rows of the figure show the source domain visible images and the corresponding labels, respectively, the third row are the target domain infrared training set images without labels, and the fourth and fifth rows display the target domain infrared test set images and the corresponding labels, respectively.

Table VI details the comparison results, whereas Figure 13 reveals the visualized results. It can be concluded from them that:

1) The proposed T-DANet (Stage2) attains the mIoU of 43.70% on the public dataset RGB_T, which is 0.85–11.22% higher than that of other methods. Clearly, our method achieves greater segmentation results than other methods on the public dataset as well.
2) As is also clear from the visualized results, the proposed T-DANet (Stage2) exhibits relatively intact target segmentation regions on the infrared images. Its segmentation results are more complete than those of other methods, without excessive segmentations, suggesting its good performance.
3) In summary, the proposed method all along exhibits high segmentation accuracy on the public dataset RGB_T, indicating its preferable segmentation accuracy and generalizability.

V. CONCLUSION

In this paper, we propose a two-stage domain adaptation method for infrared ship target segmentation. It trains the segmentation network in two stages to further improve the segmentation accuracy of infrared ship images. Specifically, an image style transfer network is designed first to reduce the appearance difference between visible and infrared images. Then, a random attention module is constructed to enhance the feature extraction capability of the segmentation network. Finally, a random inter-domain image splicing method is developed to acquire the image correlation information between two domains. The effectiveness and generalizability of the proposed method are validated on the self-made VI-Ship dataset, as well as on the public dataset RGB_T.

Nonetheless, the proposed method still has the limitation of an incomplete segmentation on the infrared ship images with small targets. In the future, we will focus on the segmentation technique for small target ship images, and design appropriate modules to make the network pay attention to small target ships, with a view to segment more accurately.

REFERENCES

[1] H. Luo, K. Wu, Z. Guo, L. Gu, and L. M. Ni, "Ship detection with wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 7, pp. 1336–1343, Jul 2011.

[2] L. Bo, X. Xiaoyang, W. Xingxing, and T. Wenting, "Ship detection and classification from optical remote sensing images: A survey," *Chinese Journal of Aeronautics*, vol. 34, no. 3, pp. 145–163, Mar 2021.

[3] M. J. Er, Y. Zhang, J. Chen, and W. Gao, "Ship detection with deep learning: a survey," *Artificial Intelligence Review*, pp. 1–41, Mar 2023.

[4] D. Jin and X. Bai, "Distribution information based intuitionistic fuzzy clustering for infrared ship segmentation," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 8, pp. 1557–1571, Aug 2019.

[5] K. Tanaka, N. Ikeya, T. Takatani, H. Kubo, T. Funatomi, V. Ravi, A. Kadambi, and Y. Mukaigawa, "Time-resolved far infrared light transport decomposition for thermal photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2075–2085, Jun 2019.

[6] Y. Zheng, X. Lv, L. Qian, and X. Liu, "An optimal bp neural network track prediction method based on a ga–aco hybrid algorithm," *Journal of Marine Science and Engineering*, vol. 10, no. 10, p. 1399, 2022.

[7] Y. Zheng, P. Liu, L. Qian, S. Qin, X. Liu, Y. Ma, and G. Cheng, "Recognition and depth estimation of ships based on binocular stereo vision," *Journal of Marine Science and Engineering*, vol. 10, no. 8, p. 1153, 2022.

[8] Y. Yao, F. Shu, Z. Li, X. Cheng, and L. Wu, "Secure transmission scheme based on joint radar and communication in mobile vehicular networks," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[9] L. Liu, S. Zhang, L. Zhang, G. Pan, and J. Yu, "Multi-uuv maneuvering counter-game for dynamic target scenario based on fractional-order recurrent neural network," *IEEE Transactions on Cybernetics*, 2022.

[10] Y. Liu, G. Li, and L. Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[11] H. Zhang, G. Luo, J. Li, and F.-Y. Wang, "C2fda: Coarse-to-fine domain adaptation for traffic object detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 12 633–12 647, 2021.

[12] Z. Jin, B. Liu, Q. Chu, and N. Yu, "Isnet: Integrate image-level and semantic-level context for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Oct 2021, pp. 7189–7198.

[13] L. Zhang and X. Gao, "Transfer adaptation learning: A decade survey," *IEEE Transactions on Neural Networks and Learning Systems*, Jun 2022.

[14] Y. Zhang, "A survey of unsupervised domain adaptation for visual recognition," *arXiv preprint arXiv:2112.06745*, Dec 2021.

[15] X. Liu, C. Yoo, F. Xing, H. Oh, G. El Fakhri, J.-W. Kang, J. Woo *et al.*, "Deep unsupervised domain adaptation: A review of recent advances and perspectives," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, Aug 2022.

[16] K. Saito and K. Saenko, "Ovanet: One-vs-all network for universal domain adaptation," in *Proceedings of the ieee/cvf international conference on computer vision*, Oct 2021, pp. 9000–9009.

[17] Y. Zheng, L. Li, L. Qian, B. Cheng, W. Hou, and Y. Zhuang, "Sine-ssa-bp ship trajectory prediction based on chaotic mapping improved sparrow search algorithm," *Sensors*, vol. 23, no. 2, p. 704, 2023.

[18] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, May 2016.

[19] J. Yang, J. Liu, N. Xu, and J. Huang, "Tvt: Transferable vision transformer for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Jan 2023, pp. 520–530.

[20] S. M. Ahmed, D. S. Raychaudhuri, S. Paul, S. Oymak, and A. K. Roy-Chowdhury, "Unsupervised multi-source domain adaptation without access to source data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Jun 2021, pp. 10 103–10 112.

[21] J. Peng, Y. Huang, W. Sun, N. Chen, Y. Ning, and Q. Du, "Domain adaptation in remote sensing image classification: A survey," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9842–9859, Nov 2022.

[22] L. Qian, Y. Zheng, L. Li, Y. Ma, C. Zhou, and D. Zhang, "A new method of inland water ship trajectory prediction based on long short-term memory network optimized by genetic algorithm," *Applied Sciences*, vol. 12, no. 8, p. 4073, 2022.

[23] X. Zhang, S. Wen, L. Yan, J. Feng, and Y. Xia, "A hybrid-convolution spatial–temporal recurrent network for traffic flow prediction," *The Computer Journal*, p. bxac171, 2022.

[24] Y. Cheng, S. Lan, X. Fan, T. Tjahjadi, S. Jin, and L. Cao, "A dual-branch weakly supervised learning based network for accurate mapping of woody vegetation from remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 124, p. 103499, 2023.

[25] M. Yang, Y. Wang, Y. Liang, and C. Wang, "A new approach to system design optimization of underwater gliders," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 5, pp. 3494–3505, 2022.

[26] F. Song, Y. Liu, D. Shen, L. Li, and J. Tan, "Learning control for motion coordination in wafer scanners: Toward gain adaptation," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 12, pp. 13 428–13 438, 2022.

[27] X. Liang, Z. Huang, S. Yang, and L. Qiu, "Device-free motion & trajectory detection via rfid," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 17, no. 4, pp. 1–27, 2018.

[28] L. Liu, Y. Wu, Y. Wang, J. Wu, and S. Fu, "Exploration of environmentally friendly marine power technology-ammonia/diesel stratified injection," *Journal of Cleaner Production*, vol. 380, p. 135014, 2022.

[29] B. Wang, Y. Zhang, and W. Zhang, "A composite adaptive fault-tolerant attitude control for a quadrotor uav with multiple uncertainties," *Journal of Systems Science and Complexity*, vol. 35, no. 1, pp. 81–104, 2022.

[30] S. u. Rehman, S. Tu, Z. Shah, J. Ahmad, M. Waqas, O. u. Rehman, A. Kouba, and Q. H. Abbasi, "Deep learning models for intelligent healthcare: implementation and challenges," in *Artificial Intelligence and Security: 7th International Conference, ICAIS 2021, Dublin, Ireland, July 19–23, 2021, Proceedings, Part I 7*. Springer, 2021, pp. 214–225.

[31] S. u. Rehman, Y. Huang, S. Tu, B. Ahmad *et al.*, "Learning a semantic space for modeling images, tags and feelings in cross-media search." PAKDD, 2019.

[32] J. Latif, S. Tu, C. Xiao, S. Ur Rehman, A. Imran, and Y. Latif, "Odgnet: a deep learning model for automated optic disc localization and glaucoma classification using fundus images," *SN Applied Sciences*, vol. 4, no. 4, p. 98, 2022.

[33] S. u. Rehman, S. Tu, Y. Huang, G. Liu *et al.*, "Csfl: A novel unsupervised convolution neural network approach for visual pattern classification," *Ai Communications*, vol. 30, no. 5, pp. 311–324, 2017.

[34] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. Pmlr, Jul 2018, pp. 1989–1998.

[35] W. Hong, Z. Wang, M. Yang, and J. Yuan, "Conditional generative adversarial network for structured domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Jun 2018, pp. 1335–1344.

[36] X. Wu, Z. Wu, H. Guo, L. Ju, and S. Wang, "Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2021, pp. 15 769–15 778.

[37] W. Yan, Y. Qian, C. Wang, and M. Yang, "Threshold-adaptive unsupervised focal loss for domain adaptation of semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, Jan 2022.

[38] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, Dec 2016.

[39] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 766–785, Mar 2019.

[40] M. Toldo, A. Maracani, U. Michieli, and P. Zanuttigh, "Unsupervised domain adaptation in semantic segmentation: a review," *Technologies*, vol. 8, no. 2, p. 35, Jun 2020.

[41] Z. Ullah, M. I. Mohmand, S. u. Rehman, M. Zubair, M. Driss, W. Boulila, R. Sheikh, I. Alwawi *et al.*, "Emotion recognition from occluded facial images using deep ensemble model." *Computers, materials and continua*, vol. 73, no. 3, 2022.

[42] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, Jun 2018, pp. 7472–7481.

[43] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2019, pp. 2507–2516.

[44] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2019, pp. 2517–2526.

[45] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Oct 2019, pp. 2090–2099.

[46] T.-D. Truong, C. N. Duong, N. Le, S. L. Phung, C. Rainwater, and K. Luu, "Bimal: Bijective maximum likelihood approach to domain adaptation in semantic scene segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Oct 2021, pp. 8548–8557.

[47] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia *et al.*, "A review of single-source deep unsupervised visual domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 473–493, Feb 2020.

[48] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, Oct 2017, pp. 2223–2232.

[49] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Jul 2017, pp. 1125–1134.

[50] L. A. Gatys, A. S. Ecker, and M. Bethge, “A neural algorithm of artistic style,” *arXiv preprint arXiv:1508.06576*, Aug 2015.

[51] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, Oct 2016, pp. 694–711.

[52] Z. Yi, H. Zhang, P. Tan, and M. Gong, “Dualgan: Unsupervised dual learning for image-to-image translation,” in *Proceedings of the IEEE international conference on computer vision*, Oct 2017, pp. 2849–2857.

[53] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, “Learning to discover cross-domain relations with generative adversarial networks,” in *International conference on machine learning*. PMLR, Aug 2017, pp. 1857–1865.

[54] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, “Attention mechanisms in computer vision: A survey,” *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, Mar 2022.

[55] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, “Spatial transformer networks,” *Advances in neural information processing systems*, vol. 28, Dec 2015.

[56] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Jun 2018, pp. 7794–7803.

[57] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, “Simam: A simple, parameter-free attention module for convolutional neural networks,” in *International conference on machine learning*. PMLR, Jul 2021, pp. 11 863–11 874.

[58] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Jun 2018, pp. 7132–7141.

[59] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, “Attentional feature fusion,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Jan 2021, pp. 3560–3569.

[60] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “Eca-net: Efficient channel attention for deep convolutional neural networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Jun 2020, pp. 11 534–11 542.

[61] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Jun 2016, pp. 770–778.

[62] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proceedings of the IEEE international conference on computer vision*, Oct 2017, pp. 1501–1510.

[63] 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 transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, Apr 2017.

[64] A. Hore and D. Ziou, “Image quality metrics: Psnr vs. ssim,” in *2010 20th international conference on pattern recognition*. IEEE, Aug 2010, pp. 2366–2369.

[65] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, Dec 2019.

[66] A. Torralba, B. C. Russell, and J. Yuen, “Labelme: Online image annotation and applications,” *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1467–1484, Aug 2010.

[67] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, Dec 2014.

[68] S. Ruder, “An overview of gradient descent optimization algorithms,” *arXiv preprint arXiv:1609.04747*, Sep 2016.

[69] S. K. Behera, A. K. Rath, and P. K. Sethy, “Fruits yield estimation using faster r-cnn with miou,” *Multimedia Tools and Applications*, vol. 80, pp. 19 043–19 056, Mar 2021.

[70] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, “Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Sep 2017, pp. 5108–5115.