Entropy Guided Adversarial Domain Adaptation for Aerial Image Semantic Segmentation

Aihua Zheng, Ming Wang, Chenglong Li, Jin Tang, and Bin Luo, *Senior Member, IEEE*

***Abstract*— Recent advances on aerial image semantic segmentation mainly employ the domain adaption to transfer knowledge from the source domain to the target domain. Despite the remarkable achievement, most methods focus on the global marginal distribution alignment to reduce the domain shift between source and target domains, leading to a wrong mapping of the well-aligned features. In this article, we propose an effective unsupervised domain adaptation approach, which relies on a novel entropy guided adversarial learning algorithm, for aerial image semantic segmentation. In specific, we perform local feature alignment between domains by learning a self-adaptive weight from the target prediction probability map to measure the interdomain discrepancy. To exploit the meaningful structure information among semantic regions, we propose to utilize the graph convolutions for long-range semantic reasoning. Comprehensive experimental results on the benchmark dataset of aerial image semantic segmentation and natural scenes demonstrate the superior performance of the proposed method compared to the state-of-the-art methods.**

***Index Terms*— Aerial image, graph convolutional network (GCN), information entropy, semantic segmentation, unsupervised domain adaptation (UDA).**

I. Introduction

SEMANTIC segmentation [1] is a fundamental computer vision task that assigns a predictive category for each pixel of an input image, which has wide practical applications such as auto-driving [2] and image editing [3].

With potential applications such as land-cover monitoring and planning [4], climate monitoring and forecasting [5], urban management [6], building extraction [7], recent efforts devoted to aerial image semantic segmentation for better understanding the remote scenarios with reliable inferential knowledge. However, the aerial images collected by airborne satellites or unmanned aerial vehicles normally contain a wealth of detailed land information with redundant information and noise, which bring great challenges for semantic segmentation.

Manuscript received August 8, 2021; accepted September 13, 2021. Date of publication September 29, 2021; date of current version January 26, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61976002, Grant 61976003, Grant 61860206004, and Grant U20B2068. *(Corresponding author: Chenglong Li.)*

Aihua Zheng and Chenglong Li are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei 230601, China (e-mail: ahzheng214@foxmail.com; lcl1314@foxmail.com).

Ming Wang, Jin Tang, and Bin Luo are with Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: wangming@stu.ahu.edu.cn; tangjin@ahu.edu.cn; ahu_lb@163.com).

Digital Object Identifier 10.1109/TGRS.2021.3113581

Recent achievements on natural image semantic segmentation, such as FCN [8], U-Net [9], SegNet [10], PSPNet [11], ResNet [12], and DeepLab [13] significantly boost the development of aerial image semantic segmentation [14]–[16] due to the powerful capability of feature learning and representation. Despite their excellent success in aerial image semantic segmentation, the key issue is these data-driven supervised methods rely on a large amount of annotations, which are extremely time-consuming and labor-intensive, especially for the extraordinary high-resolution aerial image. Although weakly supervised methods [17]–[21] and synthetic data [22], [23] can overcome this issue to some extent, they still require extra human efforts or have a large discrepancy in texture, spatial layout, color, and lighting conditions between synthetic and real images which is also called domain shift. Domain adaptation, as a particular case of transfer learning, has been widely explored to simulate the human visual system to learn from one or more related source domain data distribution and transfer the learned knowledge to the target domain [24], [25]. Unsupervised domain adaptation (UDA) bridges the gap between the labeled source and the unlabeled target domains mainly through maximum mean discrepancy (MMD) [26], [27], adversarial learning [28]–[34], or self-training [35]–[37]. Recent efforts evidence the promising achievement of adversarial learning-based DA in UDA semantic segmentation.

Adversarial DA uses the discriminator to improve the ability of the generator while learning domain-invariant features. Early efforts [30], [38]–[40] mainly focus on generative adversarial learning methods by employing the generative adversarial network (GAN) [41] structure, which generates simulated images to mimic the appearance of the target domain and retain the semantic invariant. Recent efforts [28], [29], [33], [34] adopt the subspace adversarial learning method to learn the domain-invariant features, which maps images from both domains to a common subspace and then uses a discriminator to reduce distribution discrepancy due to the similarity of the feature distribution in the feature space [29] or the spatial layout in the output space [33], and [32] uses the differences between the two classifiers to promote feature alignment. Others methods [31], [42]–[44] make efforts to integrate both generative and subspace adversarial learning methods. More recent works devote to introduce curriculum learning [45] and entropy minimization [46] to adversarial learning to reduce the domain shift. However, most existing adversarial learning-based UDA methods suffer from a major limitation: they focus on aligning the global marginal distribution of the features from both domains while ignoring

the local joint distribution shift that the discrepancy of regional feature similarity for multiple categories between domains, as so-called negative transfer [34].

Recent efforts [47]–[50] evidence the effectiveness of domain adaptation method on aerial image semantic segmentation. Benjdira *et al.* [47], [48] and Tasar *et al.* [49] translate images between domains to reduce style discrepancy and reuse labels as additional supervision based on the generative method. Fang *et al.* [50] apply feature generation and category-level subspace adversarial learning methods [34], via a geometry-consistent GAN embedded into a cotraining adversarial network to eliminate above negative transfer. It is difficult to train the discriminator to distinguish the transferred images or the generated high-dimensional features. At the same time, the above methods rarely take into account the domain-invariant feature alignment during the global adversarial learning, and the additional classifier increases the complexity of the model in the work of Fang *et al.* [50]. Furthermore, most of the above methods directly derive from natural images and ignore the specific characteristics of the aerial image, such as the existence of multiscale objects and the absence of regional semantic information in image preprocessing.

To solve these problems, we propose an entropy guided adversarial (EGA) learning method in the output-space to enhance the local feature alignment for UDA aerial image semantic segmentation and encode the long-range semantic dependencies between regions for compensating the semantic absence base on graph convolutional network (GCN). Information entropy, which indicates the amount of information and measures uncertainty, has been widely explored in semantic segmentation [37], [46], [51], [52]. As shown in Fig. 1(a), the model trained on the labeled source domain tends to produce over-confident prediction with low entropy. And the target domain predictions based on a source-only trained model usually represent low-entropy on the source-like (domain-invariant) regions (the category regions of "Build.") and high-entropy on the target-like (domain-variant) regions (the regions of "Imp. surf." and "Tree") due to the domain shift, as shown in Fig. 1(b) and (c). Furthermore, the information entropy map has a strong relationship with the error map, which is apparent in the labeled source domain. This relationship has also been studied in previous work [51], [52]. In particular, we have studied the entropy distributions of correct and incorrect predictions in the source and target domains in Fig. 1(c). The incorrect prediction usually corresponds to the high entropy, while the correct prediction corresponds to the low entropy. The traditional global feature alignment method ignores the local feature alignment, thus will cause prediction errors in domain-invariant regions, as shown in the third row of Fig. 1(d) (the category of "Build."). Therefore, regions should be discriminatively treated in adversarial learning. We argue that the entropy map can represent the domain shift, which in turn means we can use the entropy value to enhance the local feature alignment in adversarial learning. By the EGA learning, we encourage the network to preserve the domain-invariant features while enforcing the alignment of the domain-variant features.

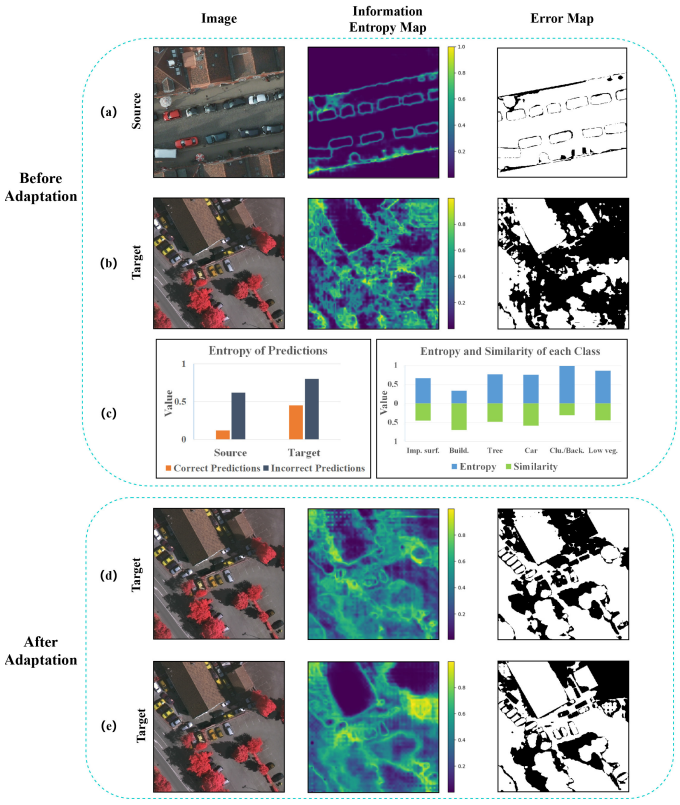


Fig. 1. Effectiveness of the proposed EGA domain adaptation for aerial image semantic segmentation. The first two rows show the results of source (a) and target (b) domain scene tested on the source-only trained model. In the third row (c), the left figure draws the relationship between the entropy and the predictions on both domains before adaptation, while the right one plots the relationship between the feature similarity between two domains for each class together with the entropy of the target domain. The last two rows show the results of the same target domain scene after traditional global feature alignment (d) and our method (e). Columns represent the images from both domains, the information entropy maps calculated on the prediction probability maps and the error maps, respectively. The error map is represented by a binary map to show the difference between the ground truth (GT) and the prediction map, where white and black pixels indicate the correct and incorrect predictions, respectively.

In addition, we note that the scene layout between the source and the target domain in aerial image presents a huge discrepancy comparing to the natural image, which brings more challenges to the segmentation task, as shown in Fig. 2(a) and (b). Due to the high resolution, aerial images are usually divided into multiblocks, which will result in missing region semantics, such as the "Build." cutting from the middle significantly destroy the local region structure. Therefore, the region's relationships are needed for long-range semantic dependence. As shown in Fig. 2(c), the feature response map after the long-range semantic reasoning (LSR) is enhanced by encoding more context information and refining object boundaries.

The main contributions of our work can be summarized as follows.

1) We propose a novel UDA network to handle the problem of local distribution shift for aerial image semantic segmentation.

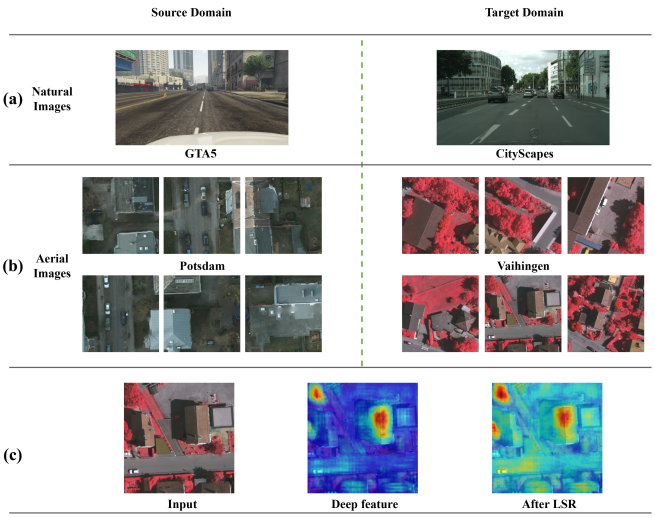


Fig. 2. Comparison of natural image and aerial image, and the feature response map with/without the LSR. The scene layout of the natural image in the source domain is more similar as in the target domain comparing with the aerial image by comparing (a) with (b). Furthermore, the divided blocks in the aerial image have lost/destroyed the local semantic information, as shown in (b). In (c), we show the deep feature response map with/without LSR.

2) An EGA domain adaption approach is proposed, which learns the adaptive weight of the predictive entropy map in the target domain to guide the adversarial learning for local feature alignment.
3) A LSR model is designed to explore local regional dependencies and compensate for the lack of semantic structure information.
4) Experiments on benchmark aerial segmentation dataset ISPRS and natural synthetic-to-real dataset demonstrate the effectiveness and robustness of our proposed method, which yields a new state-of-the-art performance.

II. Related Work

We briefly review methods of aerial image semantic segmentation and related works on UDA for semantic segmentation.

A. Aerial Image Semantic Segmentation

The powerful ability of deep models and the availability of large-scale aerial images with high visual quality provides an opportunity to enhance segmentation performance. Hence, many representative deep learning-based methods have been applied to aerial image semantic segmentation. Existing efforts mainly concern to modify the FCN framework for segmentation work [14]–[16], [51]–[58]. For example, Sherrah [53] incorporate the dilated convolution into FCN without extra down-sampling operation. To fuse high-level semantic information with low-level detail information, Wang *et al.* [51] propose a gate mechanism implemented by entropy maps to fuse feature maps from different levels. Furthermore, Panboonyuen *et al.* [57] employ a global convolutional network to capture complex features and produce weights for the feature in different layers by the channel attention. They also introduce domain-specific transfer learning that initializes the weights by using other aerial image to alleviate the scarcity of training dataset. Although the above methods achieve significant improvement, they mainly train segmentation networks through many annotated images, which are usually difficult to be collected in real scenarios. Meanwhile, they are based on the assumption that the training and testing sets fall into the same data distribution, which is also hard to be satisfied when the domain shift emerges.

Recent efforts [47]–[50] evidence the effectiveness of domain adaptation method for aerial image semantic segmentation. Benjdira *et al.* [47] proposed to generate images from the source domain to the target domain for reducing style discrepancy and reuse labels from the source domain. Tasar *et al.* [49] proposed a translating framework, ColorMapGAN framework to mimic the spectral distribution of the target image. Meanwhile, Benjdira *et al.* [48] proposed to use two GAN structures and a small set of labeled target images to improve the generalization ability of the model on the target domain. Fang *et al.* [50] proposed the category-sensitive domain adaptation via a geometry-consistent GAN embedded into a cotraining adversarial network. However, these generation-based methods are generally hard to train, and their segmentation performance is significantly restricted by the quality of the generated images or features.

B. UDA for Semantic Segmentation

Adversarial learning [41] is one of the most active techniques for UDA semantic segmentation due to its powerful ability in image or feature generation and the adversarial loss can be used to promote feature distribution alignment between domains. Adversarial learning can be considered into two parts, including generative adversarial learning and subspace adversarial learning methods.

The works [29], [32]–[34], [59] apply the subspace adversarial learning method to learn domain-invariant features. The map images from both domains to a common subspace (feature- or output-space) and then use discriminators to reduce domain discrepancies. Hoffman *et al.* [29] are the first to explore the adversarial learning based on feature subspace for UDA semantic segmentation, propose both global, and category-specific domain adaptation techniques. Tsai *et al.* [33] propose a multilevel adversarial network on different feature representation levels to realize output-space-based adversarial learning. However, such methods pursue global feature alignment and ignore the regional discrepancy between domains. Luo *et al.* [34] propose a category-level adversarial network using the cotraining mechanism and calculating the two distinct view-prediction local-aligned maps to guide the adversarial learning for keeping local semantic consistency. Zheng and Yang [59] propose an orthogonal method to exploit the intradomain knowledge and reduce the inherent uncertainty by minoring the discrepancy of two different classifiers. Saito *et al.* [32] minimize the difference from two classifiers to align feature distribution between domains. Some works [30], [31], [38]–[40], [42], [43] employ the generative adversarial learning to translate the image styles for decreasing the appearance difference in the

image-level and make feature invariant to style information. Bousmalis *et al.* [30], Li *et al.* [39], and Wu *et al.* [40] apply style transfer approaches to reduce appearance discrepancy.

Hoffman *et al.* [31] propose the cycle-consistent adversarial domain adaptation, based on the CycleGAN [38], to generate images and reuse corresponding labels by the constraint of the cycle-consistency and semantic losses. Li *et al.* [43] jointly consider subspace and generative adversarial learning methods, apply bidirectional learning for domain adaptation, and introduce a self-supervised learning method to alternatively learn the image generator and the segmentation model by perceptual loss. Vu *et al.* [46] propose two complementary losses, i.e., entropy loss and adversarial loss, based on information entropy to reduce the data distribution gap between domains. Zhang *et al.* [45] design curriculum-style learning by considering the consistency of the local and global label distributions for UDA semantic segmentation. Other works introduce the idea of self-training. Zou *et al.* [35], [36] and Saporta *et al.* [37] propose to generate high-confidence pseudo-labels on the target domain, and then use the pseudo-labels to train the target model.

III. Methods

In this section, we shall elaborate on the proposed LSR and EGA domain adaptation for aerial image semantic segmentation.

A. Problem Settings

Given the source aerial image set X_S with the paired pixel-level labels set Y_S, and the target unlabeled aerial image set X_T, the ultimate goal of unsupervised domain adaption in aerial image semantic segmentation is to learn a model G producing the pixel-level labels for the target aerial image. The traditional adversarial network employs a generative model G to transfer knowledge from the source domain to the target domain by minimizing a supervised segmentation loss in the source domain. The supervised segmentation loss defined as

$$\mathcal{L}_{\text{seg}}(G) = \mathbb{E}_{(x_s,y_s)\sim(X_S,Y_S)}[\ell(G(x_s), y_s)] \tag{1}$$

where $x_s \in X_S$, $y_s \in Y_S$, $\mathbb{E}[\cdot]$ represents the statistical exception and ℓ indicates the commonly employed cross-entropy loss.

Meanwhile, to generalize the model G to the target domain, adversarial-based methods encourage G to learn the domain-invariant features by confusing the domain discriminator D which tries to distinguish the outputs from the source or target domain. This process is achieved by minimizing an adversarial loss

$$\mathcal{L}_{\text{adv}}(G, D) = -\mathbb{E}_{x_s\sim X_S}\left[\log(D(G(x_s)))\right] - \mathbb{E}_{x_t\sim X_T}\left[\log(1 - D(G(x_t)))\right] \tag{2}$$

where $x_t \in X_T$, represents an unlabeled sample from target set.

As mentioned above, traditional adversarial learning methods [29], [31], [33], [42], [46] focus on the global feature alignment, which will cause the negative transfer. The direct effect is, some already aligned features may be destroyed by the global margin distribution alignment. Meanwhile, the aerial image has specific characteristics, such as the multiscale variations and the loss of structural continuity during data preprocessing. To relieve these problems, we proposed an EGA learning method to compute a self-adaptive weight that suppresses the interdomain shift during the adversarial learning by information entropy, and the LSR by GCN to explore the structure information.

B. Overview of the Proposed Model

As shown in Fig. 3, there are two main components in the proposed model, i.e., a generator G and a discriminator D. The generator G aims to generate feature representation and prediction map, which consists of a segmentation backbone (DeepLab-V2 [13] with ResNet-101 [12] in this article) and an embedding LSR model. The discriminator D is a convolutional neural network (CNN)-based binary classifier with a fully convolutional output. In our framework, G consists of three parts: 1) feature extractor E, that extracts CNN features F from both domains, 2) embedded LSR model, that explores the image structure information and enhances the local feature representation on F, and 3) classifier M, that produces the prediction probability map. The discriminator D works on the output-space to align the distributions of two domain predictions under the guidance of the information entropy map of the target domain.

For the source domain aerial image x_s, we first generate the deep feature map F_s by feature extractor E, followed by the LSR model to aggregate the node information through the information passing mechanism of the graph structure. Then, the aggregated node features are mapped to the original deep feature space and merged with deep features by summation operation, to enhance the information of the features without increasing the model parameters comparing to the concatenation. After that, we obtain the prediction map P_s of the combined features $\tilde{F}_s$ via the classifier M, followed by the supervised segmentation loss with the corresponding label y_s.

In the same manner, we can obtain the prediction map P_t for the target domain. In addition, based on the probability prediction map, we further calculate the entropy map as a self-adaptive weight Wg, which can guide the adversarial learning to perform feature alignment. Meanwhile, to pull the distribution of target prediction to the source domain, the discriminator D is introduced to distinguish the prediction map P_t from the source or target domain. Then we calculate the adversarial loss weighted by Wg to learn the diverse attention for each pixel.

C. LSR Model

Our proposed LSR model aims to explore the regional dependence relations in the aerial image by the information passing mechanism of the graph structure.

The GCN [60] operates on the instance graph to explore the data structure information. In this article, we propose to aggregate the meaningful structure information in aerial image segmentation via GCN due to its capability of extracting

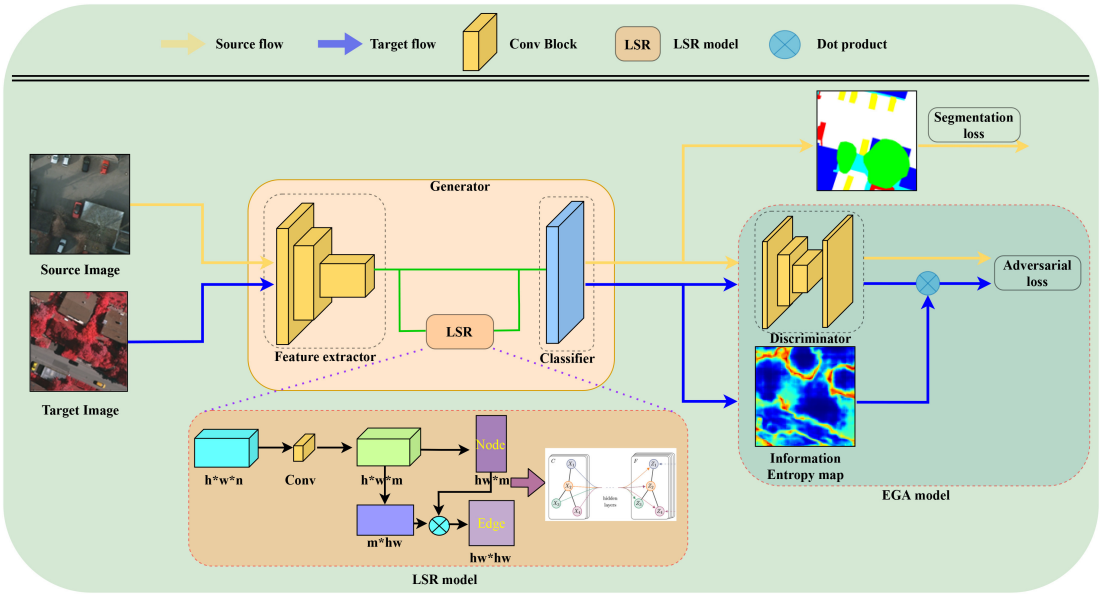


Fig. 3. Pipeline of our proposed EGA learning and the LSR model base on GCN. It contains two components, generator (G) and discriminator (D). The generator consists of a feature extractor (E), an embedded LSR model, and a classifier (M). For the source domain input, we use the prediction after the classifier to calculate a supervised segmentation loss. As for the target input, based on the output of the classifier, we first calculate a self-adaptive weight on the information entropy map. Meanwhile, to enforce the local distribution alignment between the source and target prediction, the discriminator calculates the adversarial loss weighted by the self-adaptive weight.

and aggregating node features by nodes relations from the graph structure. We directly construct the graph structure on the original CNN feature maps to preserve the spatial relationship. The feature map is one-eighth of the input image size, therefore the computational cost is not extremely large to handle the most existing cropped aerial or natural images. Given an undirected graph $g = (v, e)$, the nodes v indicate each pixel node of the deep feature map, the edges e represent the similarity between nodes, and $A \in \mathbb{R}^{v \times v}$ is an adjacency matrix. Unlike the standard convolution, which operates on a local regular-grid region, the graph can encode the long-range dependencies through the relation between nodes represented by the adjacency matrix. Therefore, it has a larger receptive field and can explore more context information. Formally, the graph convolution can be defined as

$$X^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X^{(l)}\Theta^{(l)}\right) \tag{3}$$

where $\tilde{A} = A + I$, I is an identity matrix, and $\tilde{D}$ is a degree matrix, represents the degree of each node, is used to normalize the new features of each node aggregated from the connected nodes. $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$, i and j, respectively, represent the ith row and the jth column of $\tilde{A}$. The value of degree matrix $\tilde{D}$ is used to calculate the symmetric normalized Laplacian matrix $\tilde{D}^{-(1/2)}\tilde{A}\tilde{D}^{-(1/2)}$ to measuring the relations of nodes and compute the new features of each node as the average of itself and connected nodes. $X^{(l)}$ is the vertex feature matrix in lth layer, σ is the nonlinear activation function (rectified linear unit (ReLU) used in our experiments), and $\Theta^{(l)}$ is a trainable weight matrix.

To apply the forward propagation rule in (3) on the CNN feature map to explore the long-range dependence and structure information. We directly perform graph reasoning on the original feature map. For an input deep CNN feature map F with the shape of $h \times w \times n$, where h, w, n represent the height, width, and channel number, respectively. To reduce parameters and computational costs, we use a 1×1 convolution reducing the channel into m. Therefore, the graph node input can be represented by transferring features after reduced dimension into $X \in \mathbb{R}^{hw \times m}$. As shown in Fig. 3, we calculate the adjacency matrix A by the dot-product distance of the dimension-reduced features. In this way, it will increase the computational cost in the LSR module when aggregating features. As we evaluated in the ablation study in Table V, it significantly boosts the performance by enhancing the contextual features without introducing significant computational cost. The similarity between position i and j is expressed as

$$A_{ij} = \phi(F)_i \phi(F)_j^T \tag{4}$$

where F is deep feature map extracted from CNN, $\phi(\cdot)$ is a 1×1 convolution, and $\phi(F) \in \mathbb{R}^{hw \times m}$ is a liner embedding followed by ReLU$(\cdot)$. The A can be calculated as follows:

$$A = \phi\big(F; W_\phi\big)\phi\big(F; W_\phi\big)^T \tag{5}$$

where W_ϕ is the learnable parameters for the linear transformations. Meanwhile, $\tilde{A} = A + I$ and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. We can formulate the graph model as

$$Z = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta\right) \tag{6}$$

where Θ is a trainable weight matrix, σ is the ReLU activation function, and Z is the output features.

After the LSR model, the output Z projects back into the original deep feature space and then fused with the original deep CNN feature map F. By this means, the refined feature $\tilde{F}$ encodes long-range semantic dependencies and aggregates the image structure information.

D. EGA Model

To relieve the negative transfer during domain adaptation, we propose an EGA model, which can learn a self-adaptive weight from the information entropy map to guide the adversarial process.

In this work, we employ the widely used Shannon entropy [61]. For a target prediction P_t with the shape of $C \times H \times W$, the entropy map Ent $\in [0,1]^{H\times W}$ can be calculated as follows:

$$\mathrm{Ent}^{(i,j)} = -\sum_{c=1}^{C} {P_t}^{(i,j,c)} \log {P_t}^{(i,j,c)} \tag{7}$$

where P_t is a prediction probability map in target domain flow, C, H, and W correspond to the class numbers, height, and width, respectively, and $i \in \{0,1,\ldots,H-1\}$, $j \in \{0,1,\ldots,W-1\}$. ${P_t}^{(i,j,c)}$ represents the predicted probability in the class c (not the real label) at the position (i, j).

As discussed above, information entropy can be used to measure the domain shift. Due to the shift between domains, the predictions of the regions with the large discrepancy on the target domain image may be under-confident, corresponding to high-entropy. While the predictions of similar regions are generally with low-entropy. Therefore, it is meaningful to use entropy weight to measure the discrepancies between domains via weighting the adversarial loss. The EGA loss can be defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{ega}}(G,D) = &-\mathbb{E}_{x_s\sim X_S}\left[\log(D(G(x_s)))\right] \\ &-\mathbb{E}_{x_t\sim X_T}\left[\mathrm{Wg}\times\log(1-D(G(x_t)))\right]\end{aligned} \tag{8}$$

where the weight map is a tensor with the size of $1 \times H \times W$, and $\mathrm{Wg}^{(i,j)} = (1/C) \times \mathrm{Ent}^{(i,j)}$. The domain adversarial loss map produced by D is upsampled to the size of the input image. An element-wise multiplication is operated between the weight map and domain adversarial loss map. After this operation, each location in the domain adversarial loss map has different attention, and the $\mathcal{L}_{\mathrm{ega}}(G, D)$ will focus on the low similarity regions and decrease the attention on the high similarity regions on the target sample, which can push local feature alignment between domains.

E. Training Objective

As the traditional adversarial network, our method is optimized by two loss functions, segmentation loss, and self-adaptive EGA loss.

1) Segmentation Loss: Given an aerial image x_s of shape $3 \times H \times W$ and correspond label map y_s of shape of $C \times H \times W$, C is the number of object classes in the domain.

The segmentation loss can be defined as

$$\mathcal{L}_{\mathrm{seg}}(G) = \sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} -{y_s}^{(i,j,c)} \log {P_s}^{(i,j,c)} \tag{9}$$

where ${P_s}^{(i,j,c)}$ represents the predicted probability belonging to class c at the location pixel (i, j). ${y_s}^{(i,j,c)}$ denotes the corresponding GT probability belonging to class c at the location (i, j). If pixel (i, j) belongs to class c, ${y_s}^{(i,j,c)} = 1$, else ${y_s}^{(i,j,c)} = 0$.

2) Self-Adaptive EGA Loss: In the traditional adversarial learning process, due to the negative shift in global feature alignment, we jointly consider local and global features alignment. In the DA process, we consider generating self-adaptive attention to different regions to promote local region feature alignment. By the EGA learning, we encourage the network to preserve the domain-invariant features while enforcing the alignment of the domain-variant features. And the self-adaptive weight Wg is calculated by the prediction probability of the target domain input. Given an aerial image x_t with the shape of $3 \times H \times W$, P_t represents the prediction after generator G. Thus, the self-adaptive weight map Wg can be calculated as follows:

$$\mathrm{Wg}^{(i,j)} = -\frac{1}{C}\sum_{c=1}^{C} {P_t}^{(i,j,c)} \log {P_t}^{(i,j,c)} \tag{10}$$

where dividing by C is to reduce the calculated information entropy value and prevent misalignment of features caused by high entropy. The local feature alignment loss formula is mentioned has mentioned in (8).

Furthermore, to balance local and global features alignment, we design λ_w and ε to promote the feature alignment. By considering both global and local feature alignment, our method can mitigate the domain shift and encourage the generator to learn more domain-invariant knowledge. The self-adaptive EGA loss can be formulated as

$$\begin{aligned}\mathcal{L}_{\mathrm{EGA}}(G,D) = &-\mathbb{E}_{x_s\sim X_S}\left[\log(D(G(x_s)))\right] \\ &-\mathbb{E}_{x_t\sim X_T}\left[(\lambda_w \mathrm{Wg}+\varepsilon)\log(1-D(G(x_t)))\right]\end{aligned} \tag{11}$$

where λ_w effects on the self-adaptive weight Wg, and the ε is a decimal to stabilize the training process.

With the above descriptions, the overall loss objective can be summarized as

$$\mathcal{L}(G,D) = \mathcal{L}_{\mathrm{seg}}(G) + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{EGA}}(G,D) \tag{12}$$

where λ_{adv} denotes the hyperparameters that control the influence of the $\mathcal{L}_{\mathrm{EGA}}(G, D)$, whose training objective is shown as follows:

$$G^*, D^* = \arg\min_G \max_D \mathcal{L}(G,D). \tag{13}$$

F. Analysis

According to the theory of domain adaptation proposed by Ben-David *et al.* [24], let $\mathcal{H}$ be the hypothesis class, S and T, respectively, express the source and the target domains, the theory defines the expected error on the target samples $\epsilon^T(h)$ as follows:

$$\forall h \in \mathcal{H}, \quad \epsilon^T(h) \le \epsilon^S(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(S,T) + \lambda \tag{14}$$

where $\epsilon^S(h)$ represents the expected error on the source domain which can be minimized easily by predictions of the generator G based on the source labels, $d_{\mathcal{H}\Delta\mathcal{H}}(S, T)$ measures the discrepancy distance between the source and the target distributions, and λ is the shared expected loss expected to be negligibly small.

Followed by Ben-David *et al.* [24], $d_{\mathcal{H}\Delta\mathcal{H}}(S,T)$ is minimized by the global domain adversarial learning. Recently, UDA semantic segmentation commonly use global margin alignment to narrow the discrepancy between domains [28]

$$\begin{aligned} & d_{\mathcal{H}\Delta\mathcal{H}}(S,T) \\ & = 2 \sup_{h\in\mathcal{H}\Delta\mathcal{H}} |P_{f\sim S}[h(f)=1] - P_{f\sim T}[h(f)=1]| \\ & \le 2 \sup_{h\in\mathcal{H}_d} |P_{f\sim S}[h(f)=1] - P_{f\sim T}[h(f)=1]| \\ & = 2 \sup_{h\in\mathcal{H}_d} |P_{f\sim S}[h(f)=0] + P_{f\sim T}[h(f)=1] - 1|. \quad (15) \end{aligned}$$

However, as we have mentioned, the global feature alignment chases the global marginal distribution without considering the local discrepancies between domains. As shown in Fig. 1(c), the higher similarity of the class features between domains, the smaller the entropy value. Furthermore, the lower similarity corresponds to the higher entropy. Therefore, in the feature alignment process, it is essential to discriminatively treat different classes. Thus, we propose to use the entropy value to reflect the discrepancies between domains, and adaptively learn the attention of distinctive regions to better promote the alignment of features between domains.

Our method consists of two parts: LSR and EGA learning. LSR aims to enhance the dependence of local regions through the message propagation mechanism and aggregate context information, which enhances domain-invariant feature extraction ability of the generator. Meanwhile, EGA extends the traditional adversarial loss with a self-adaptive attention $\mathrm{Wg}^{(i,j)} = -(1/C)\sum_{c=1}^{C} P_t^{(i,j,c)} \log P_t^{(i,j,c)}$. Wg indicates the amount of uncertain information extracted from the local regions, while the entropy output corresponding to the region discrepancy compared to the source domain. Therefore, larger Wg encourages the generator G to fool discriminator D and pays more attention to the local regions while adversarial learning. Smaller Wg indicates that the generator G has learned enough domain-invariant knowledge in this region, thus weakens the adversarial learning process to remain the local feature alignment.

IV. Experiments

To verify the effectiveness of the proposed method, we evaluate our method in comprehensive scenarios in both aerial and natural image segmentation datasets.

A. Setup

1) Datasets: We evaluate our method on a 2-D semantic segmentation benchmark dataset ISPRS. ISPRS is offered by the International Society for Photogrammetry and Remote Sensing 2-D Semantic Labeling Contest, which currently provides the best evaluation platform for aerial image semantic segmentation. It contains two subsets the Potsdam[1] and the Vaihingen.[2]

[1]http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[2]http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html

Potsdam subset contains 38 aerial images covering 3.42 km^2 area of the Potsdam city with a resolution of 5 cm. Those images are fixed with a size of 6000 × 6000 pixels in three channels: red, green, and blue. About 24 out of 38 images in Potsdam have pixel-annotation, which spans six categories: Building, Tree, Car, Impervious surfaces, Low vegetation, and Clutter/background.

Vaihingen subset contains 33 aerial images covering 1.38 km^2 area of the Vaihingen city with a resolution of 9 cm. The size of each image is approximately 2000 × 2000 pixels in three different channels: near-infrared, red, and green. About 16 out of 33 images provide pixel-annotation, with the same categories classification as in Potsdam.

Following the protocol in [47], [48], and [50], for Potsdam (source domain)-to-Vaihingen (target domain) task, we use 24 labeled images with the paired labels in Potsdam and 17 unlabeled Vaihingen images as the training data, while the remaining 16 Vaihingen images with the paired labels are as the testing data. On the contrary, for the Vaihingen-to-Potsdam task, we use 16 Vaihingen labeled images with the paired labels and 14 Potsdam unlabeled images as training data, while the remaining 24 Potsdam labeled images for testing. As preprocessing, we crop all the high-resolution images in both domains into 512 × 512 pixels by a sliding window with an overlap of 256 pixels.

2) Implementation Details: In the same manner as state-of-the-art methods, we employ the widely used DeepLab-V2 [13] framework with ResNet-101 [12] model pretrained on ImageNet [62] as the backbone. Then 1×1 convolution layer is used to reduce the channel m of the deep feature map from 2048 to 64. In our LSR model, we use a three-layer GCN network with the feature dimension as {64, 512, 1024} and insert the dropout layer between the GCN layer with a rate of 0.5, following a 1 × 1 convolution layer to recover the channel and match the size with the original deep feature map. Once obtaining the refined feature map, we combine the refined feature with the input feature by summation with a 1 × 1 convolution layer. Subsequently, Atrous Spatial Pyramid Pooling (ASPP) with dilated rates {6, 12, 18, 24} is employed to predict segmentation maps. Finally, an up-sampling layer with the softmax function output the prediction probability map with the size of the input image. Meanwhile, for the discriminator, we adopt the similar structure as in [63], which contains 5 convolution layers with kernel of 4 × 4, channel numbers of {64, 128, 256, 512, 1} and stride of 2. Each convolution layer follows a Leaky-ReLU parameterized by 0.2 except the last layer. Finally, the prediction of the discriminator is upsampled to the same size of the input for matching the size of the weight map. Empirically, followed by our baseline [33] and other state-of-the-art domain adaption works [46], [50], we use different optimizers to optimize generator and discriminator. We use stochastic gradient descent (SGD) [64] with the initial learning rate as 2.5×10^{-4} and a momentum of 0.9 as the optimizer for the generator. The discriminator optimizer employs Adaptive Moment Estimation (Adam) [65] with the initial learning rate 5×10^{-5} and $\beta_1 = 0.9, \beta_2 = 0.99$. Theoretically, SGD [64] is a mainstream algorithm to optimize the generator [28]. SGD can speed up the convergence with

TABLE I

COMPARISON RESULTS BASED ON THE DOMAIN ADAPTATION FROM POTSDAM-TO-VAIHINGEN. THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN, AND BLUE, RESPECTIVELY

Methods	adversarial learning	Reference	PA(%)	mF1(%)	Imp. surf.	Build.	Low veg.	Tree	Car	Clu./Back.	mIoU(%)
NoAdapt			46.26	36.19	26.86	41.67	12.56	44.48	16.70	1.76	24.01
CycleGAN [38]	generative	ICCV2017	54.44	43.38	34.91	52.75	20.65	42.28	21.45	5.54	29.60
Benjdira's [47]		Remote Sens.2019	52.50	43.09	31.87	52.36	23.67	41.76	21.70	4.48	29.31
CsDA [50]	generative & subspace	Remote Sens.2019	55.98	45.65	44.82	53.39	20.74	52.04	23.81	0.55	32.56
Baseline	subspace		61.76	46.94	54.05	60.82	21.78	56.67	15.56	0.40	34.87
AdaptSegNet [33]		CVPR2018	62.43	46.20	50.05	62.61	20.73	56.45	14.18	1.08	34.18
Advent [46]		CVPR2019	65.51	51.13	55.43	68.49	20.73	59.02	28.28	0.73	38.78
CLAN [34]		CVPR2019	64.91	48.76	57.30	59.19	24.94	59.10	17.28	0.84	36.44
MRNet [59]		IJCAI2020	65.31	50.33	54.11	75.39	16.16	54.99	29.39	0.81	38.47
Ours (concatenation)	subspace		65.50	52.90	55.43	65.42	28.85	53.11	27.25	5.25	39.22
Ours (summation)	subspace		67.65	53.56	57.11	70.24	28.73	56.55	23.65	5.84	40.35

a strong generalized optimal solution while avoiding local minimums. Adam [65] can speed up the convergence of the discriminator, thereby promoting the feature extraction of the generator for the domain-invariant. We set both optimizers a weight decay of 5×10^{-4}. Meanwhile, the learning rate decays by a poly learning rate policy. The initial learning rate is multiplied by $(1 - \text{iter}/(\text{max\_iter}))^{\text{power}}$ with power of 0.9. In all experiments, the maximum iterations are set to 100 k, and the batch size is 1. In our best model, the hyperparameters λ_{adv}, λ_w, and ϵ are set to 0.001, 5, and 0.6, respectively, on Potsdam-to-Vaihingen domain adaptation task. Our experiments are implemented in a PyTorch environment with a single NVIDIA Tesla V100-16G GPU.

Followed by the state-of-the-art methods [47], [48], we use four metrics to evaluate the performance of our method, pixel accuracy (PA), mean F1 score (mF1), intersection over union (IoU), and mean IoU (mIoU).

B. Quantitative Comparison

1) Potsdam-to-Vaihingen: Table I reports the experimental results of our method comparing with the advanced methods on Potsdam to Vaihingen datasets. Due to the domain shift between the two datasets, the source-only method NoAdapt presents stumbling performance. Although the generative methods improve the results by a large margin, they are hard to preserve the semantic information during the generation, which leads to limited performance. Generally speaking, the domain adaptation methods significantly boost the performance in overall metrics (mIoU, PA, and mF1) and individual categories, revealing their ability to transfer knowledge from the labeled source domain to the unlabeled target domain. By establishing the EGA learning for enhancing local feature alignment and the LSR for the structure information, our method is significantly superior to all the compared methods, which verifies the effectiveness of our method for learning domain-invariant features. Specifically, first, by comparing all domain adaptation methods with NoAdapt, the segmentation metrics achieve a large degree of improvement. Meanwhile, we note that the IoU of "Car" declines to some extent after adaptation, which endures the side effect of pursuing global feature alignment. Second, the generative adversarial learning methods have slight improvement due to the semantic information retain hardly for the translated image. CsDA [50] achieves better segmentation performance benefit from the extra feature- and output-space adversarial learning, which considering the spatial contextual similarity. Third, the result of our method together with the Baseline, AdaptSegNet [33], Advent [46], CLAN [34], and MRNet [59] present the superiority of output-space feature alignment. Our method in "Build." and "Imp. surf." make significant improvement due to the consideration of the regional dependencies and structure information enhancement. As an entropy-guided local feature alignment method, our method can even effectively segment the "Tree" and "Low veg." which present high interclass similarity. At last, although MRNet [59] and Advent [46] achieve promising mIoU, they are still overshadowed than our method due to their limited capability in some challenging categories such as "Low veg." and "Clu./Back.," which evidences the robustness of our method on challenging scenarios for aerial image semantic segmentation. In addition, summation operation outperforms concatenation operation in our model, which verifies the summation can enhance the information of the features without increasing the model parameters. We use the summation in the following experiments if not specified.

2) Vaihingen-to-Potsdam: To further verify the effectiveness of our method, we evaluate our method by exchanging the source domain and target domain data as the Vaihingen-to-Potsdam. As shown in Table II, first of all, our method consistently beats the state-of-the-art methods as in the Potsdam-to-Vaihingen task as shown in Table I. Note that, the images in Vaihingen dataset cover a narrow area with a lower resolution with relatively rough annotation, while Potsdam includes images from a larger area with much higher resolution and more refined annotation. This brings more challenges in the Vaihingen-to-Potsdam domain adaptation task.

3) Synthetic-to-Real: To verify the generalization ability of our method, we conduct experiments of domain adaption from two synthetic datasets GTA5 [22] and Synthia [23] to the real dataset Cityscapes [66], as shown in Table III (GTA5-to-Cityscapes) and Table IV (Synthia-to-Cityscapes). First, our method achieves promising performance against the state-of-the-art domain adaptation methods on natural images, which verifies the effectiveness of the proposed method. Second, our method focuses on local feature alignment and LSR thus can effectively reduce the impact of negative transfer, such as the infrequent classes "fence," "pole," "bike" in GTA5-to-Cityscapes, and "sign" in Synthia-to-Cityscapes. Third, the improvement on the natural dataset is not as significant

TABLE II

COMPARISON RESULTS BASED ON THE DOMAIN ADAPTATION FROM VAIHINGEN-TO-POTSDAM. THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN, AND BLUE, RESPECTIVELY

Methods	adversarial learning	Reference	PA(%)	mF1(%)	Imp. surf.	Build.	Low veg.	Tree	Car	Clu./Back.	mIoU(%)
NoAdapt			44.33	37.20	35.78	37.68	12.90	14.60	33.33	9.55	23.97
CycleGAN [38]	generative	ICCV2017	51.69	44.94	36.08	50.14	23.84	31.24	25.65	12.95	29.99
Benjdira's [47]		Remote Sens.2019	52.26	47.84	36.69	42.73	30.64	33.41	34.56	14.38	32.07
CsDA [50]	generative & subspace	Remote Sens.2019	45.25	40.36	32.33	37.25	22.40	26.60	43.29	0.01	27.00
Baseline	subspace		55.13	46.15	42.77	48.75	35.42	28.96	36.70	0.26	32.14
AdaptSegNet [33]		CVPR2018	57.02	47.95	46.58	52.68	32.16	32.95	39.42	0.19	34.00
Advent [46]		CVPR2019	60.03	50.20	49.80	54.85	40.19	26.94	46.71	0.21	36.45
CLAN [34]		CVPR2019	59.33	48.54	47.96	49.95	42.25	26.50	41.00	0.17	34.64
MRNet [59]		IJCAI2020	58.25	50.21	48.56	54.34	36.40	26.20	54.52	0.02	36.67
Ours	subspace		63.01	51.01	47.47	63.19	42.57	30.93	39.42	0.36	37.32

TABLE III

COMPARISON RESULTS BASED ON THE DOMAIN ADAPTATION FROM GTA5-TO-CITYSCAPES. THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN, AND BLUE, RESPECTIVELY

Methods	Reference	road	side.	buil.	wall	fence	pole	light	sign	vege.	terr.	sky	pers.	rider	car	truck	bus	train	motor	bike	mIoU
NoAdapt		75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
Baseline		86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
AdaptSegNet [33]	CVPR2018	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
Advent [46]	CVPR2019	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
CLAN [34]	CVPR2019	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
MRNet [59]	IJCAI2020	89.1	23.9	82.2	19.5	20.1	33.5	42.2	39.1	85.3	33.7	76.4	60.2	33.7	86.0	36.1	43.3	5.9	22.8	30.8	45.5
Ours		88.4	34.3	81.2	27.9	22.4	32.9	39.7	35.3	82.3	27.3	73.5	61.8	31.3	83.5	32.2	38.5	1.8	30.7	38.2	45.4

TABLE IV

COMPARISON RESULTS BASED ON THE DOMAIN ADAPTATION FROM SYNTHIA-TO-CITYSCAPES. THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN, AND BLUE, RESPECTIVELY

Methods	Reference	road	side.	buil.	light	sign	vege.	sky	pers.	rider	car	bus	motor	bike	mIoU
NoAdapt		55.6	23.8	74.6	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	38.6
Baseline		79.2	37.2	78.8	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.3	45.9
AdaptSegNet [33]	CVPR2018	84.3	42.7	77.5	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	46.7
Advent [46]	CVPR2019	85.6	42.2	79.7	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	48.0
CLAN [34]	CVPR2019	81.3	37.0	80.1	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	47.8
MRNet [59]	IJCAI2020	82.0	36.5	80.4	18.0	13.4	81.1	80.8	61.3	21.7	84.4	32.4	14.8	45.7	50.2
Ours		84.5	39.0	79.0	18.5	20.6	77.6	82.7	58.2	21.8	77.3	30.8	12.1	51.4	50.3

as on the aerial image dataset. The main reason is, compared with natural images, the large-scale objects in aerial images tend to lose more semantic information during cropping. Meanwhile, there are larger discrepancies between classes in true orthophoto of the aerial image. Our method is more suitable for aerial image segmentation through LSR to enhance feature representation and information entropy to promote local feature alignment.

C. Qualitative Comparison

Fig. 4 visualizes some qualitative segmentation examples obtained by our method compared to the state-of-the-art competitors on Potsdam-to-Vaihingen. Due to the domain shift between the Potsdam and the Vaihingen, the predictions of NoAdapt [Fig. 4(a)] introduce much noise and lose the object boundary and structure information. The segmentation results of the generative adversarial learning methods, CylceGAN [38] and the Benjdira's [47] as shown in Fig. 4(b) and (c) are also unsatisfactory for the limited generated images by global consistency constraint. CsDA [50] as shown in Fig. 4(d), achieves a significant improvement by introducing extra feature generation and output-space adversarial learning that considers the spatial distribution information. However, it still suffers from unclear object boundaries such as the "Build." category and confuses misclassification such as "Low veg." and "Tree" categories. The Baseline, AdaptSegNet [33], and Advent [46] pursue the global margin alignment, which results in much misclassification such as "Tree" and "Low veg." Furthermore, the "Build." boundaries are still rough and involving many holes, as shown in Fig. 4(e)–(g). CLAN [34] is able to generate more precise and accurate predictions by considering the category-level data distribution and local semantic consistency, as shown in Fig. 4(h). However, due to the large appearance discrepancy, such as the left and the right "Build." comparing to the middle "Build." as shown in the second row, it suffers the challenge of intradomain discrepancy for aerial image semantic segmentation. MRNet [59] presents refined segmentation results by considering the intradomain adaptation. However, it fails to isolated individuals from each other, such as "Build."

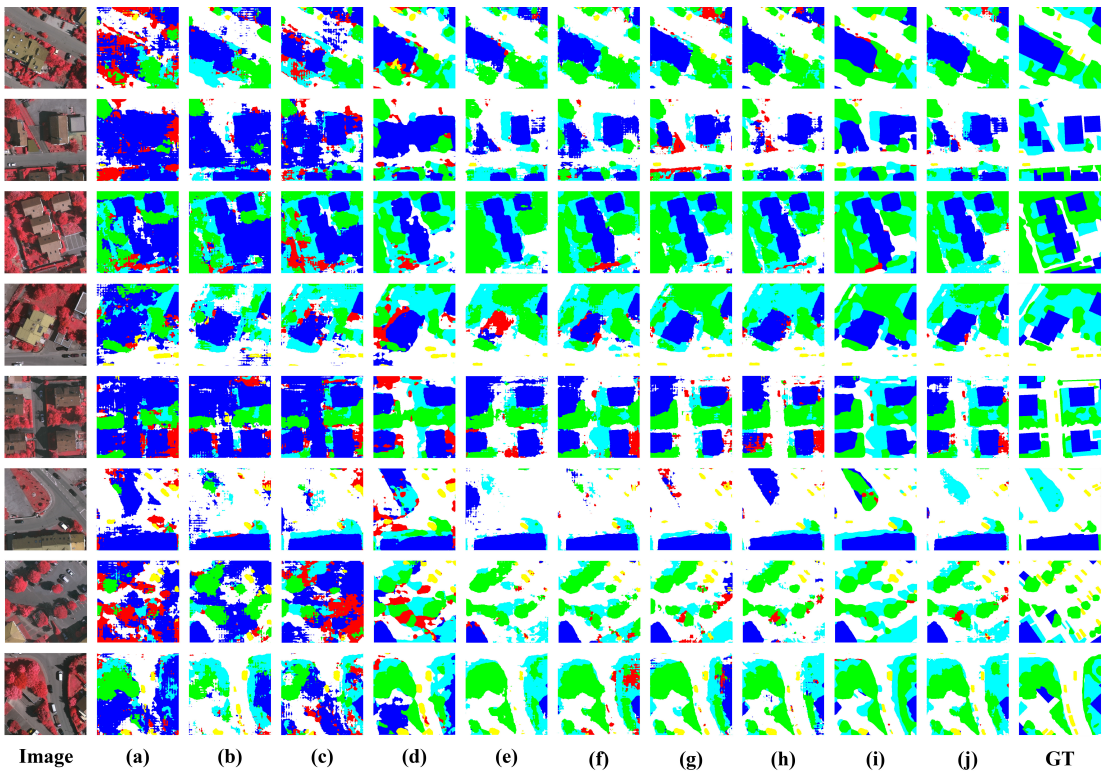


Fig. 4. Visual Results of our proposed method compared with other state-of-the-art methods on Potsdam-to-Vaihingen. The blue, green, yellow, cyan, white, and red separately represent the categories of "Build.," "Tree," "Car," "Low veg.," "Imp. surf.," and "Clu./Back." (a) NoAdapt, (b) CycleGAN, (c) Benjdira's, (d) CsDA, (e) Baseline, (f) AdaptSegNet, (g) Advent, (h) CLAN, (i) MRNet, (j) Ours.

TABLE V
ABLATION STUDIES ON THE POTSDAM-TO-VAIHINGEN DOMAIN ADAPTATION

	Method	PA(%)	mF1(%)	Imp. surf.	Build.	Low veg.	Tree	Car	Clu./Back.	mIoU(%)	PAD	Cost(s/iteration)
B1	Baseline	61.76	46.94	54.05	60.82	21.78	56.67	15.56	0.40	34.87	0.90	0.41
B2	B1 + LSR	65.21	52.72	53.81	66.63	28.32	**58.84**	23.37	5.42	39.40	-	0.54
B3	B1 + EGA	64.32	50.92	54.40	66.14	24.98	57.04	**24.45**	1.81	38.14	0.72	0.42
B4	B1 + LSR+ EGA	**67.65**	**53.56**	**57.11**	**70.24**	**28.73**	56.55	23.65	**5.84**	**40.35**	-	0.55

in the second and fourth rows in Fig. 4(i). Furthermore, it presents obvious misclassification, such as the "Low veg." misclassified as "Tree" as shown in the sixth row. Our proposed method with the LSR model explores the long-range dependencies of the input image for structure information and self-adaptive adversarial learning for local feature alignment. Thus it produces more precise predictions and preserves better object boundary information, as shown in Fig. 4(j). Especially, we achieve much accurate prediction on small objects such as "Car" than other methods by the local feature alignment, as shown in the fourth and eighth rows. By comparing the third, fourth, and fifth rows of Fig. 4(j) with other state-of-the-art methods, our method can predict a more precise "Build." boundary. Even though some "Build." boundary is still rough in our method, as shown in the second row of in Fig. 4(j), our method can effectively separate them from adjacent objects. Furthermore, due to the sensor variation between domains, the "Low veg." and "Tree" result in high similarity in appearance, as shown in the last two rows. This brings a big challenge for the state-of-the-art methods while can be successfully segmented by our method benefit from the proposed EGA learning. The key reason is we consider the discrepancies of local regions between domains and the weight of information entropy to guide the local feature alignment.

D. Component Analysis

1) Ablation Study: To further verify the contribution of the components in our model, we conduct the ablation study on several variants of our method on Potsdam-to-Vaihingen domain adaptation task, as reported in Table V. B1 is our Baseline, which takes a global feature alignment base on the DeepLab-V2 framework. B2 embeds GCN-based LSR on Baseline B1 to explore the region correlation and the structure information of the input. B3 combines the information entropy of the target domain input to guide the adversarial process, called EGA on Baseline B1. B4 is our final model, which combines both LSR and EGA to the Baseline. By the information passing mechanism, the embedded LSR model can capture the structure information in the input image, which purges the prediction map, especially on large-scale objects. Therefore, B2 achieves significant improvement comparing with B1, as shown in Table V. By arguing the entropy map

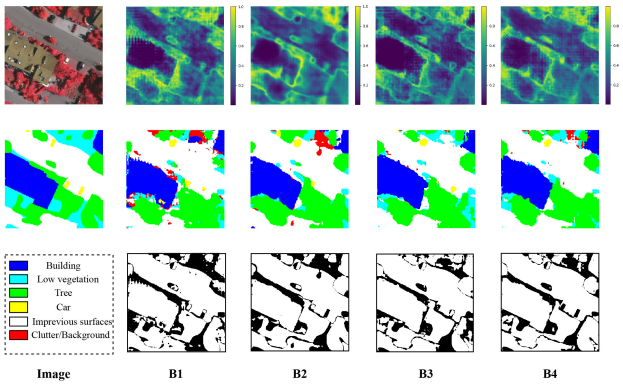


Fig. 5. Visual Results of our proposed method on Potsdam-to-Vaihingen. B0: NoAdapt, B1: Our baseline, B2: B1+ our proposed LSR, B3: B1+ our proposed EGA, B4: B1+ LSR + EGA. The top row represents the input target image and the predicted probability maps of different variants. The middle row is the GT and prediction results, while the bottom row shows the prediction error map comparing to the GT.

representing the interdomain shift, the EGA model can refine the segmentation results by the local aligned method, especially on the local regions and the small objects. Compared B1 with B3 in Table V, it demonstrates the EGA model can effectively improve the segmentation accuracy. By integrating both LSR and EGA, our final model B4 further improves the segmentation results, which verifies the contribution of the two components. In addition, we evaluate the computational cost of each module on the baseline during the training as shown in Table V. Both LSR and EGA modules significantly increase the segmentation accuracy without introducing distinct processing costs.

Fig. 5 further visualizes the segmentation results of the variants. Comparing the results of B2 with B1, the "Build." presents more refined boundaries. However, the two categories "Tree" and "Low veg." around "Build." tend to be one category due to the high interclass similarity. The prediction error map in B3 is significantly better than B1. Since EGA concerns the interdomain discrepancy while simultaneously preserves the well-aligned features, B3 can pay different attention to different regions by information map for the domain-variant categories "Tree" and "Low veg.," which leads to significantly better segmentation than B1 and B2. And, our EGA can preserve well-aligned features and accelerate the unaligned features as a local feature alignment approach. Therefore, B4 achieves a more accurate segmentation performance by combining both LSR and EGA modules, which reinforce each other during the learning.

2) Evaluation on LSR: To evaluate the contribution of the proposed LSR via capturing the long-range semantic dependencies, we further visualize the feature response maps of the DA network on Potsdam-to-Vaihingen task as shown in Fig. 6. By comparing the feature response maps with/without LSR, the feature maps with LSR have high responses by considering the relations and aggregative context information between different regions. Specifically, the small objects have high responses which will not be discarded during the feature extraction, such as the "Tree" in the second and the sixth columns, and the "Car" in the fourth column. The big objects present more clear boundary responses after LSR, shown as

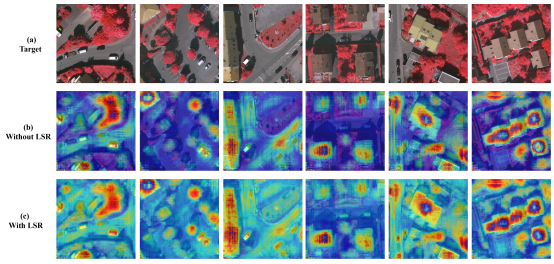


Fig. 6. Visual representation of the feature response maps with/without our proposed LSR on Potsdam-to-Vaihingen. (a) is the original target images, (b) represents the feature response maps without fusion of the LSR feature, and (c) represents the feature response maps after fusion.

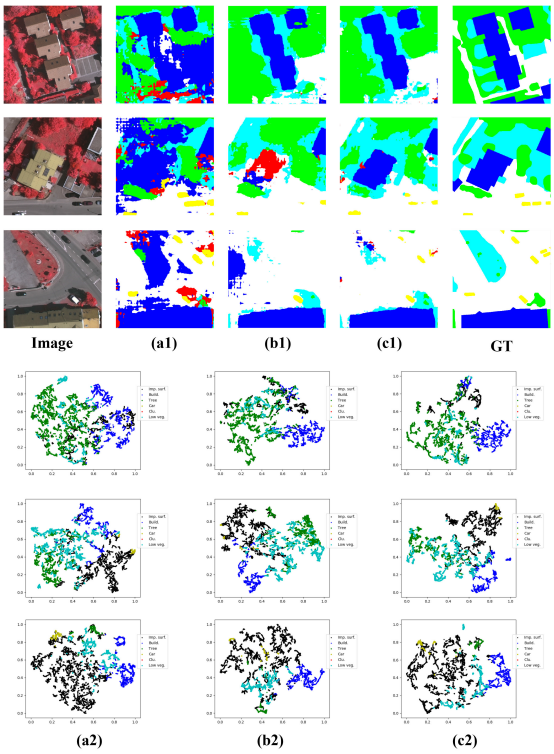


Fig. 7. Comparison of segmentation results from the source-only method NoAdapt (a1), Baseline (b1) and Baseline with EGA (c1) on Potsdam-to-Vaihingen. And the h-dimensional features of (a1)–(c1) are mapped into a 2-D space with t-SNE [67] in (a2)–(c2). Each color represent a class that has drawn to the upper right of distribution maps. The comparison of feature distribution confirms that our method can enhance the feature alignment.

the "Build." in the fourth and the fifth columns, and the "Imp. surf." in the third column.

3) Evaluation on EGA: Fig. 7 visualizes the segmentation results and the corresponding feature distribution maps by t-SNE [67] of the source-only method NoAdapt, comparing with our Baseline and the Baseline with EGA. From the segmentation results in Fig. 7(a1) and the feature distribution Fig. 7(a2) we can see, NoAdapt can capture the domain-invariant features, but generates misclassification in the domain-variant regions. By considering the global alignment, the Baseline can promote the generator to learn domain-invariant features. However, due to the lackness of local alignment, it tends to map domain invariant regions into other categories, such as the category of "Build." in the first and second rows and

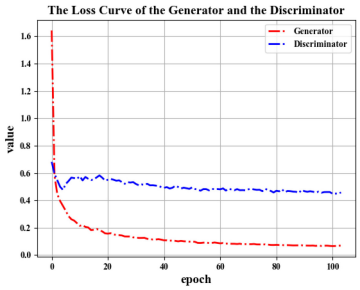


Fig. 8. Converge curve of the generator G and the discriminator D in our proposed EGA method on Potsdam-to-Vaihingen.

the car category in the second and third rows. By introducing the proposed EGA to the Baseline to balance the local and the global feature alignment, Baseline with EGA as shown in Fig. 7(c1) can well promote the alignment in the domain-variant regions. Furthermore, by comparing the feature distribution in Fig. 7(a2)–(c2), EGA can reduce category confusion while maintaining the original aligned categories by considering the local feature alignment approach, such as the "Build." in the first and third rows and the "Car" in the second and third rows.

In addition, we use Proxy $\mathcal{A}$-distance (PAD) to measure the domain distance followed by Ben-David *et al.* [24]. The PAD is calculated based on the generalization error ϵ by a classifier to discriminate source and target examples as

$$\hat{d}_{\mathcal{A}} = 2(1 - 2\epsilon). \tag{16}$$

Herein, we keep the discriminator D as the classifier in our experiments. As reported in Table V, EGA effectively reduces PAD, which indicates that considering the discrepancy in local regions can better promote feature alignment and reduce the domain distance.

E. Other Analysis

1) Convergence Analysis: In addition, we use the loss curve of the generator G and the discriminator D to indicate the convergence performance, as shown in Fig. 8. We can see the loss of the D drops rapidly at the beginning, which shows that D can effectively distinguish the shift between domains. In the initial training, the G has extracted a lot of uncertain information on the target domain due to the domain shift, our proposed EGA method will increase the weight of the uncertainty region in adversarial learning, and the loss of D will increase. As G learns more domain-invariant knowledge, the uncertainty information on the target domain decreases, and the reduced weight weakens the adversarial loss of the domain-invariant region to maintain local feature alignment. Therefore, the loss of D declines and gradually converges to a stable value. As the ability of the discriminator gradually weakens, the loss cure of G gradually smooths and converges to 0.

2) Parameter Analysis: To evaluate the impact of the hyperparameters in our proposed method, we perform sensitivity study on Potsdam-to-Vaihingen domain adaptation task about the parameters of the self-adaptive EGA loss (ε, λ_w) in (11), the weight of adversarial loss (λ_{adv}) in (12),

TABLE VI
PARAMETER ANALYSIS ON POTSDAM-TO-VAIHINGEN

Parameters	Values	mIoU(%)	Parameters	Values	mIoU(%)
ε	0.3	38.10	λ_w	1	38.95
	0.4	38.16		5	40.35
	0.5	38.61		10	39.69
	0.6	40.35		15	39.43
	0.7	39.53		20	37.78
	0.8	38.95		25	37.40
λ_{adv}	0.00025	38.18	m	16	38.55
	0.0005	38.73		32	39.26
	0.001	40.35		64	40.35
	0.002	39.45		128	38.95
	0.004	38.43		256	38.68

the reduced dimension parameter (m), and the results are reported in Table VI. We design λ_w and ϵ to balance local and global features alignment. Specifically, λ_w controls the self-adaptive weight for adversarial loss, if λ_w is large then the effect of the entropy weight dominates and the model is strongly biased toward a few classes. ϵ represents the influence of traditional global adversarial loss during training, which is limited to the range between (0, 1), while larger ϵ will bias to the global feature alignment. Following prior works [33], [46], [50], we set λ_{adv} around 0.001. The parameter m is used to reduce the number of channels in the feature maps to balance the performance and computation resources. We vary each of the four parameters while fixing the other three. From Table VI we can see, our method is not sensitive to the parameters. We empirically set $\{\varepsilon, \lambda_w, \lambda_{\text{adv}}, m\} = \{0.6, 5, 0.001, 64\}$ for the best performance.

V. CONCLUSION

In this article, we propose a GCN-based LSR and EGA learning network for UDA of aerial image semantic segmentation. LSR encodes the dependencies between regions and learns the structure information by the information passing mechanism of the graph structure. EGA uses the information entropy to guide the adversarial learning between the two domains, encouraging the network to retain the domain-invariant features and promote the alignment of the domain-variant features. Experiments on the two benchmark datasets demonstrate the robustness and effectiveness of our proposed method, which yields competitive performance compared with other state-of-the-art methods. In the future, we will introduce the self-supervised learning method in our framework to provide additional supervised information for further performance boost.

REFERENCES

[1] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*. [Online]. Available: http://arxiv.org/abs/1704.06857

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[3] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang, "Deep image harmonization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3789–3797.

[4] L. Matikainen and K. Karila, "Segment-based land cover mapping of a suburban area—Comparison of high-resolution remotely sensed datasets using classification trees and test field points," *Remote Sens.*, vol. 3, no. 8, pp. 1777–1804, Aug. 2011.

[5] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bull. Amer. Meteorol. Soc.*, vol. 93, no. 12, pp. 1879–1900, Dec. 2012.

[6] Y. Tang and L. Zhang, "Urban change analysis with multi-sensor multispectral imagery," *Remote Sens.*, vol. 9, p. 252, Mar. 2017.

[7] S. Xu *et al.*, "Automatic building rooftop extraction from aerial images via hierarchical RGB-D priors," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7369–7387, Dec. 2018.

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

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[10] 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.

[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2017, pp. 2881–2890.

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

[13] 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. 2017.

[14] K. Chen, K. Fu, M. Yan, X. Gao, X. Sun, and X. Wei, "Semantic segmentation of aerial images with shuffling convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 173–177, Feb. 2018.

[15] G. Chen, X. Zhang, Q. Wang, F. Dai, Y. Gong, and K. Zhu, "Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1633–1644, May 2018.

[16] C. Persello and A. Stein, "Deep fully convolutional networks for the detection of informal settlements in VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2325–2329, Dec. 2017.

[17] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5038–5047.

[18] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1742–1750.

[19] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1713–1721.

[20] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1635–1643.

[21] C. Niu, J. Zhang, Q. Wang, and J. Liang, "Weakly supervised semantic segmentation for joint key local structure localization and classification of aurora image," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7133–7146, Dec. 2018.

[22] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 102–118.

[23] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3234–3243.

[24] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 151–175, May 2010.

[25] W. Mei and D. Weihong, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Jul. 2018.

[26] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 97–105.

[27] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.

[28] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 1180–1189.

[29] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," 2016, *arXiv:1612.02649*. [Online]. Available: http://arxiv.org/abs/1612.02649

[30] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3722–3731.

[31] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[32] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3723–3732.

[33] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[34] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2019, pp. 2507–2516.

[35] 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. Eur. Conf. Comput. Vis.*, 2018, pp. 289–305.

[36] Y. Zou, Z. Yu, X. Liu, B. V. K. V. Kumar, and J. Wang, "Confidence regularized self-training," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5982–5991.

[37] A. Saporta, T.-H. Vu, M. Cord, and P. Pérez, "ESL: Entropy-guided self-supervised learning for domain adaptation in semantic segmentation," 2020, *arXiv:2006.08658*. [Online]. Available: http://arxiv.org/abs/2006.08658

[38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2242–2251.

[39] P. Li, X. Liang, D. Jia, and E. P. Xing, "Semantic-aware grad-GAN for virtual-to-real urban scene adaption," 2018, *arXiv:1801.01726*. [Online]. Available: http://arxiv.org/abs/1801.01726

[40] Z. Wu *et al.*, "DCAN: Dual channel-wise alignment networks for unsupervised scene adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 518–534.

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

[42] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3752–3761.

[43] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6936–6945.

[44] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "CrDoCo: Pixel-level domain transfer with cross-domain consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2019, pp. 1791–1800.

[45] 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.

[46] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2517–2526.

[47] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sens.*, vol. 11, no. 11, p. 1369, 2019.

[48] B. Benjdira, A. Ammar, A. Koubaa, and K. Ouni, "Data-efficient domain adaptation for semantic segmentation of aerial imagery using generative adversarial networks," *Appl. Sci.*, vol. 10, no. 3, p. 1092, Feb. 2020.

[49] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7178–7193, Oct. 2020.

[50] B. Fang, R. Kou, L. Pan, and P. Chen, "Category-sensitive domain adaptation for land cover mapping in aerial scenes," *Remote Sens.*, vol. 11, no. 22, p. 2631, Nov. 2019.

[51] H. Wang, Y. Wang, Q. Zhang, S. Xiang, and C. Pan, "Gated convolutional neural network for semantic segmentation in high-resolution images," *Remote Sens.*, vol. 9, no. 5, p. 446, 2017.

[52] S. Guo, Q. Jin, H. Wang, X. Wang, Y. Wang, and S. Xiang, "Learnable gated convolutional neural network for semantic segmentation in remote-sensing images," *Remote Sens.*, vol. 11, no. 16, p. 1922, Aug. 2019.

[53] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*. [Online]. Available: http://arxiv.org/abs/1606.02585

[54] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.
[55] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 474–478, Mar. 2018.
[56] S. Liu, W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li, "ERN: Edge loss reinforced semantic segmentation network for remote sensing images," *Remote Sens.*, vol. 10, no. 9, p. 1339, Aug. 2018.
[57] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning," *Remote Sens.*, vol. 11, no. 1, p. 83, 2019.
[58] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.
[59] Z. Zheng and Y. Yang, "Unsupervised scene adaptation with memory regularization *in vivo*," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1076–1082.
[60] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8950–8959.
[61] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.
[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[63] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434
[64] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. Int. Conf. Comput. Statist.*, 2010, pp. 177–186.
[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[66] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[67] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

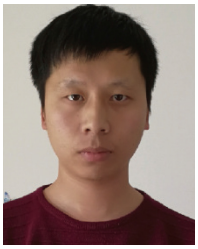
Aihua Zheng received the B.Eng. degree in computer science and technology from Anhui University, Hefei, China, in 2006.

She visited University of Stirling, Stirling, U.K., and Texas State University, San Marcos, TX, USA, during June to September in 2013 and during September 2019 to August 2020, respectively. She is currently an Associate Professor and Ph.D. Supervisor with the School of Artificial Intelligence, Anhui University. Her main research interests include vision-based artificial intelligence and pattern recognition. As the first author or corresponding author, she has published more than 40 academic papers including top conferences papers in the Association for the Advancement of Artificial Intelligence (AAAI) and International Joint Conference on Artificial Intelligence (IJCAI), and authoritative journals in IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS (TSMCS), *Pattern Recognition* (PR), *Pattern Recognition Letters* (PRL), *Neurocomputing* (NeuCom), *Cognitive Computation* (CogCom), *Neural Computing and Applications* (NCA), etc. Her main research interests include vision-based artificial intelligence and pattern recognition. Especially on person/vehicle re-identification, audio visual computing, and multimodal intelligence.

Dr. Zheng is a member of China Computer Federation (CCF) and China Society of Image and Graphics (CSIG). She is also serving as reviewers for representative conferences and journals, including AAAI, IJCAI, IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON MULTIMEDIA (TMM), IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS), PR, etc. She has been awarded the best paper in SERA 2017 and the best student paper in the workshop in International Conference on Multimedia and Expo (ICME) 2019.

Ming Wang received the B.Eng. degree in computer science and technology from Anhui University of Technology, Maanshan, China, in 2019. He is currently pursuing the M.Eng. degree in computer science and technology from Anhui University, Hefei, China.

His research interests include computer vision and optical remote sense image semantic segmentation.



Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively.

From 2014 to 2015, he worked as a Visiting Student with the School of Artificial Intelligence, Sun Yat-sen University, Guangzhou, China. He was a Post-Doctoral Research Fellow at the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. He is currently an Associate Professor and Ph.D. Supervisor at the School of Artificial Intelligence, Anhui University. His research interests include computer vision and deep learning.

Dr. Li was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively.

He is currently a Professor and Ph.D. Supervisor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision, pattern recognition, and machine learning.

Bin Luo (Senior Member, IEEE) received the B.Eng. degree in electronics and the M.Eng. degree in computer science from Anhui University, Hefei, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, York, U.K., in 2002.

He is currently a Professor with Anhui University. He has authored over 200 articles in journals, edited books, and refereed conferences. His research interests include random graph-based pattern recognition, image and graph matching, graph spectral analysis, and video analysis.

Dr. Luo is also the Chair of the IEEE Hefei Subsection. He has served as a Peer Reviewer for international academic journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition* (PR), *Pattern Recognition Letters* (PRL), the *International Journal of Pattern Recognition and Artificial Intelligence*, *Knowledge and Information Systems*, and *Neurocomputing* (NeuCom).