Unsupervised Domain Adaptation With Debiased Contrastive Learning and Support-Set Guided Pseudolabeling for Remote Sensing Images

Debojyoti Biswas, *Student Member, IEEE*, and Jelena Tešić, *Member, IEEE*

***Abstract*—The variability in different altitudes, geographical variances, and weather conditions across datasets degrade state-of-the-art (SOTA) deep neural network object detection performance. Unsupervised and semisupervised domain adaptations (DAs) are decent solutions to bridge the gap between two different distributions of datasets. The SOTA pseudolabeling process is susceptible to background noise, hindering the optimal performance in target datasets. The existing contrastive DA methods overlook the bias effect introduced from the false negative (FN) target samples, which mislead the complete learning process. This article proposes support-guided debiased contrastive learning for DA to properly label the unlabeled target dataset and remove the bias toward target detection. We introduce: 1) a support-set curated approach to generate high-quality pseudolabels from the target dataset proposals; 2) a reduced distribution gap across different datasets using domain alignment on local, global, and instance-aware features for remote sensing datasets; and 3) novel debiased contrastive loss function that makes the model more robust for the variable appearance of a particular class over images and domains. The proposed debiased contrastive learning pivots on class probabilities to address the challenge of FNs in the unsupervised framework. Our model outperforms the compared SOTA models with a minimum gain of +3.9%, +3.2%, +12.7%, and +2.1% of mean average precision for DIOR, DOTA, Visdrone, and UAVDT datasets, respectively.**

***Index Terms*—Debiased contrastive learning (DCL), object detection, remote sensing analytics, unmanned aerial vehicle (UAV) images, unsupervised domain adaptation (UDA).**

I. Introduction

REMOTE sensing images (RSIs) have numerous applications in surveillance and intelligence decision-making systems, such as agriculture, urban planning, rescue missions, and transportation systems. Research work has followed suit and demonstrated what automated analytics can uncover for the geographic mapping of resources [1], crop harvest analysis [2], emergency rescue [3], and terrestrial and naval traffic monitoring [4]. Automating aerial analytics requires localization and identification of objects in the frame. The challenge is that videos captured from high altitudes have a much higher content variability than videos captured with a person's phone.

Fig. 1. Visual difference between consumer [5] and RSIs [9].

Examples of low variability frames in consumer data and high variability in overhead structures of similar pixel size are illustrated in Fig. 1. We can see how much aerial imagery content covers large geographic areas and varies significantly within the same capture or drone flight region. We group the data variability along four dimensions w.r.t. object detection task, two related to video content capture variability and two related to the object in the video variability.

1) *Lighting conditions* significantly change the video footage captured even during one drone flight. The changes can be due to the time of day, season, weather, and cloud distribution. Fig. 2(a) shows the variations due to image capture time and lighting conditions, and the pixel intensity distribution varies.
2) *Variation in object size* is large in the same dataset due to different areas captured (e.g., urban versus rural). The objects in the frame can vary from under 0.01% to almost 70% of the entire frame. The variation is even higher between different datasets, as the footage is captured over multiple dates, terrains, and missions. Fig. 2(b) (left) contains well-defined objects, while Fig. 2(b) (right) contains lots of small (players and cars) densely packed objects.
3) *Geographical variance* of the terrestrial terrain captured in the imagery from such high altitude poses a critical challenge for object localization. Fig. 2(c) illustrates the example of the large geographical variance that can exist.
4) *Object distribution* variations in images make it challenging to separate nearly objects and eliminate overlapped objects while performing nonmax suppression (NMS).
5) *Object labeling* in aerial datasets is challenging as it is hard to distinguish correct labels among small and densely packed objects [11]. Today, only a few aerial datasets exist

Manuscript received 24 October 2023; revised 17 December 2023; accepted 27 December 2023. Date of publication 4 January 2024; date of current version 18 January 2024. This work was supported in part by Naval Air Systems Command Small Business Innovative Research under Grant N68335-18-C-0199 and in part by NVIDIA. *(Corresponding author: Debojyoti Biswas.)*

The authors are with the Department of Computer Science, Texas State University, San Marcos, TX 78666 USA (e-mail: debojyoti_biswas@txstate.edu; jtesic@txstate.edu).

Digital Object Identifier 10.1109/JSTARS.2024.3349541

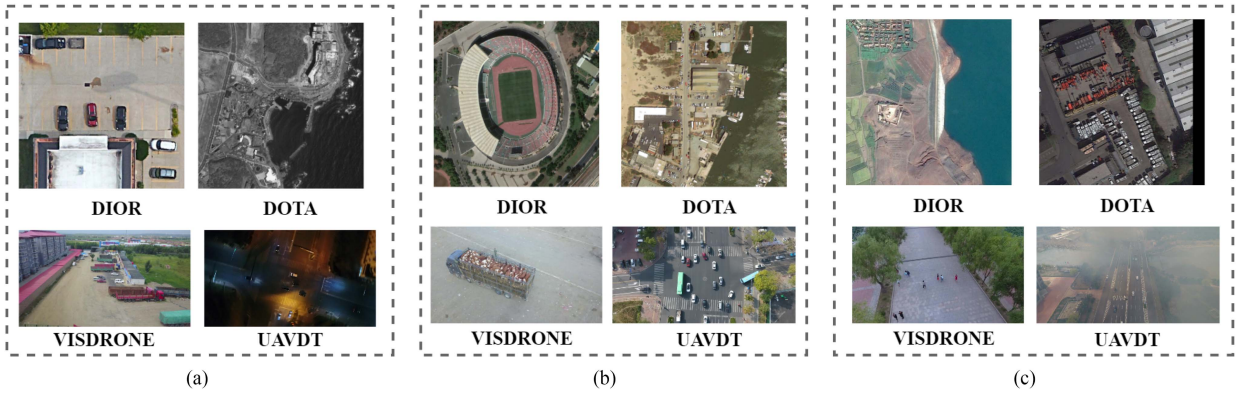

Fig. 2. High-variability remote sensing frames. (a) Lighting condition variations. (b) Variations in object shape and scale. (c) High variability due to geographical and weather changes.

that cover natural scenario object class diversity and a sufficient number of training examples.

A common technique to generalize a model is to train on one source dataset and fine-tune its application to another target dataset. However, such an approach is inefficient due to high domain shifts across datasets and the need for manual annotations of the target domain dataset. Therefore, *unsupervised domain adaptation (UDA)* methods offer a way to effectively transfer the knowledge gained from trained models on labeled source data to the unlabeled target data. UDA creates domain-invariant features using feature alignment techniques and reduces the domain gap between the different distributions of datasets. Based on this idea, the UDA methods have been widely used in the classification and segmentation tasks of RSIs [12], [14]. These techniques mainly focus on mitigating the disparity by leveraging semantic feature alignment between the source and target domains. Later, maximum mean discrepancy [16] was utilized to preserve the main statistical properties across domains by minimizing the distribution distance between the source and target domains.

Various domain adaption techniques have been proposed to improve the cross-domain classification and semantic segmentation tasks [12], [19], [31]. To our knowledge, few object detection benchmarks exist for RSIs. The dataset's highly dense and variable nature hinders the progress of pseudolabeling and optimal object detection performance of the RSIs. Xiong et al. [22] tackle the domain shift raised from the image and instance levels relying on the source-free feature alignment at the image and the instance level. On the other hand, Yan et al. [7] introduce a semantics-guided contrastive network to transfer semantic information for classes that have not been previously encountered. Chen et al. [10] present a cross-domain adaptation object detection network that is rotation invariant and relation aware. This network incorporates a relation-aware graph for aligning feature distributions and includes a rotation-invariant regularizer to handle variations in rotation. However, they still suffer from several limitations pointed out in [10]. Most UDA techniques require labeling the target datasets for instance-level domain adaptation (DA) and feature alignment. The existing pseudolabeling techniques are solely cluster based, not addressing the possible background noise being considered as foreground objects. Several deep learning clustering techniques [13], [18] have been devised for RGB and hyperspectral image (HSI) embedding classification tasks. These works [8], [13] use graph-based semisupervised learning techniques combined with tensor-based neural network embeddings for the problem of hyperspectral data classification. Moreover, spectral–spatial transformation was also introduced in [8] to learn superpixel-level spectral–spatial features from HSIs. The improved performance from deep-learning-based clustering methods comes with large computational overheads. However, we aim to use a faster technique without incurring more learnable parameters in the pipeline. Previous non-deep-learning methods use traditional k-means or one versus all for the target dataset pseudolabeling. In this work, we use an advanced clustering technique *K-means++* [6] for generating target labels due to its proven performance [15] in high-dimensional data. Then, the current contrastive learning approach follows the InfoNCE [20] loss function with a single positive instance. Two problems are involved with this technique.

1) The InfoNCE loss itself does not restrict the false negative (FN) image being selected as the negative case. For example, while performing local domain adaptation (LDA) and global domain adaptation (GDA), the negative cases are selected randomly, and an image similar to the query image (see Fig. 3) may be selected as a negative case.
2) The default InfoNCE loss works with only a positive example. However, it is essential to consider positive samples with variable appearance for a particular class over images and domains.

Therefore, instead of using the single example as the positive sample, we propose to use N numbers of positive samples for contrastive learning. Besides, we use the few-shot approach to remove the noise attracted by unwanted background object proposals. The previous work on debiased contrastive learning (DCL) [17] focuses only on balanced datasets. However, our experimental datasets are highly imbalanced; thus, this approach is invalid for our task. In summary, we propose the following research improvements:

1) a *novel framework* to address the high variability of RSIs for the object detection and labeling task in previously unseen datasets;

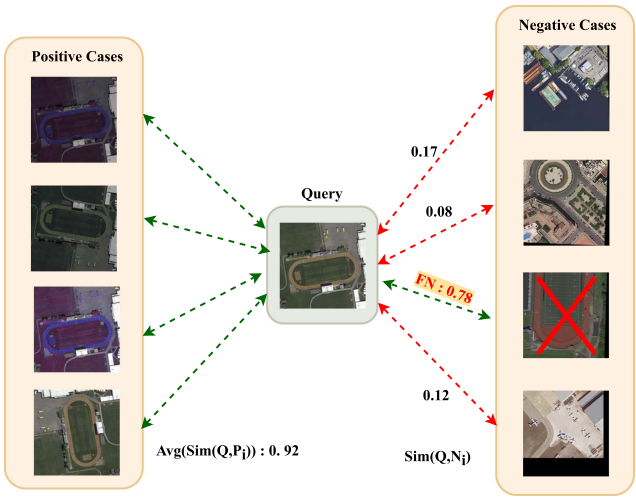

Fig. 3. Contrastive learning with multiple positive cases and FN filtering. Here, green connections denote higher similarity, and red connections denote lower similarity with the query case.

2) an *efficient pseudolabeling* process that depends on N-shot learning to remove the unwanted background noise from the target object proposals. The experiments show that curating target proposals significantly improve the target domain detection performance;
3) *DCL* for imbalanced remote sensing data, which is very important to produce domain-invariant features, but at the same time, we need to maintain class variance near the decision boundaries in the feature space. Also, we carefully filter out the *FN* examples that can disturb the learning process and result in poor performance;
4) *positive multisampling* of *N-variant* positive samples in DA [17].

The rest of this article is organized as follows. Section II summarizes related work, and Section III introduces the proposed unsupervised domain adaptation architecture with debiased contrastive learning (DCLDA) method describing the DCL approach and the different DA modules in the pipeline. In Section IV, the proposed framework is evaluated using the latest cross-domain detection benchmarks over two high-altitude and two low-altitude remote sensing datasets. Finally, Section V concludes this article.

II. Related Work

Full potential use of deep neural networks (DNNs) and machine learning has been crucial in solving recent consumer applications [23], [24]. Recent advantages in the field show that the object detection task can be successfully solved for the drone-captured Visdrone dataset [25] and the COCO consumer image benchmark dataset [26].

The key to the success of DNNs is the automatic feature extraction strategy, which is more efficient in extracting semantic details and local features. There have been numerous works to make object detection better and more efficient. The architecture of the object detection models can be divided into two branches: 1) one-stage detector and 2) two-stage detector. One-stage detectors [25], [27], [28] are by nature faster and lightweight due to less learnable parameters and FLOPS. For generating region proposals, one-stage detectors use different scale and aspect ratios of anchors. On the other hand, two-stage detectors use a separate module called region proposal network (RPN), which is responsible for generating strong region candidates for object detection.

A. Object Detection in RSIs

Shi et al. [29] propose an anchor-free-based detector called centerness-aware network, which captures the symmetrical shape of objects in remote sensing videos. Biswas and Tešić [30] suggest a strong custom backbone and an image difficulty scoring technique to help detect small and complex objects. Wu et al. [32] use the local and global contrast information to effectively detect small bright and dark objects from infrared images. The authors embed a small-sized U-Net into a larger U-Net backbone, which allows the multilevel and multiscale representation learning of objects. Zhang et al. [33] find that context-based feature extraction is more effective for detecting complex objects and scenes in the overhead imagery. The global context-weaving network incorporates a global context aggregation module and a feature refinement module [34], and transformer-based convolutional neural network encoders are used for better feature extraction [35]. Qingyun et al. perform extensive image augmentation to increase the number of samples in the minor classes. Zhu et al. [25] modify darknet53 backbone with Cross Stage Partial DenseNet and add a transformer head in the detection layer, which gains state-of-the-art (SOTA) results of overhead drone images. Overall, the overhead video frame images require special care in anchor design for one-stage detectors, and a good RPN should be chosen in two-stage sensors to capture every small object from different levels of features.

B. Unsupervised Domain Adaptation

Training data for RSIs can differ significantly from the source domain to the target domain regarding geographical, illumination, and visual characteristics. Besides RGB images, hyperspectral RSIs also suffer from variable illumination, environmental changes, and instrumental noise conditions. Hong et al. [21] handle these issues as a dictionary learning problem, where the spectral variability dictionary and estimation of the abundance maps are learned simultaneously. For a labeled source dataset and an unlabeled target dataset, UDA methods generalize the model by aligning source and target [36]. Chen et al. [37] adjust the decision boundary biased toward the target data source domain and add adversarial training in conjunction with image-to-image translation techniques. Xiong et al. [22] rely on the source-free feature alignment at the image and the instance to tackle the domain shift raised from the image and instance levels.

On the other hand, Mattolin et al. [38] implement the confidence-based mixing (ConfMix) of source and target domain images, where the confidence of an instance proposal is

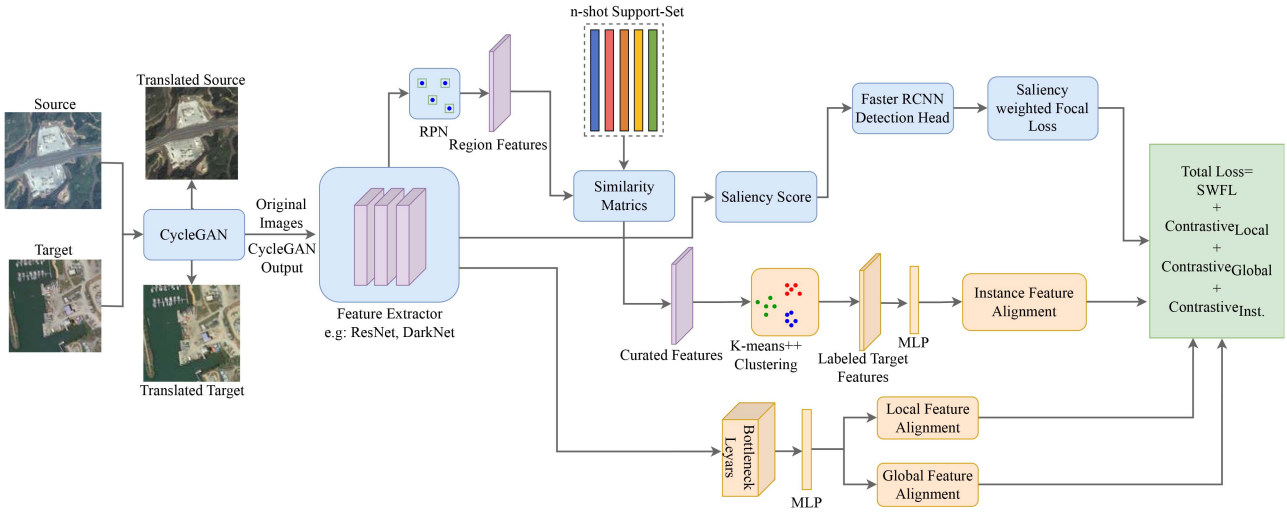

Fig. 4. UDA architecture with debiased contrastive learning.

calculated based on the objectness score and the bounding box uncertainty score of each instance proposal from the image. A novel SemantIc-complete Graph MAtching (SIGMA) [39] framework was proposed for the domain adaptation task, which completes mismatched semantics and reformulates the adaptation with graph matching. Primarily, the graph-embedded semantic completion module can address mismatched semantics by producing hallucination graph nodes within the absent categories. However, the above methods do not handle the imbalanced dataset problem and high-domain gap scenarios available in RSIs.

C. Contrastive Learning for DA

It is hard to discriminate object classes in high-variable RSIs. Contrastive learning is a technique that is a good fit as it contrasts samples against each other to learn commonalities and differences between respective object classes. Wu et al. [40] propose a probabilistic model to analyze the influence of the negative sampling ratio on training sample informativeness. Yan et al. [7] propose a semantics-guided contrastive network to transfer semantic information for classes not seen before. Bai et al. [41] propose a strategy called RefosNet to a representation focus shift network, which adds the rotation transformations to contrastive learning methods to improve the robustness of representation. Li et al. [42] use contrastive learning on overhead imagery for the semantic segmentation task. Biswas and Tešić [43] perform contrastive learning for object detection on the image-level feature alignment. However, these works do not address the noise introduced in the pseudolabeling process. Also, the mentioned contrastive learning approaches are unsuited for highly imbalanced datasets, where debiasing is required to reduce FN samples.

III. Methodology

The baseline detection architecture is built on [43], as illustrated in Fig. 4, which uses a better backbone and the saliency-weighted custom focal loss function for improved performance. The saliency information from each image is used to calculate the difficulty score of each image. Based on this saliency/difficulty score, the loss function assigns more penalties on difficult images and less on easy images.

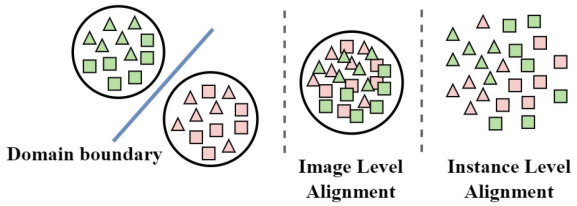

Fig. 5. Contrastive learning alignments: different colors represent different domains, and shapes represent different categories.

Contrastive learning evaluates pair-to-pair relationships by measuring the similarities between different sample pairs, such as query–positive or query–negative. Here, the query is the subject feature, whereas positive samples are augmented features similar to the subject, and negative examples are randomly selected features dissimilar to the subject feature. Performing only image-level contrastive DA is a vital feature alignment strategy that ensures that local and global features from the source and target datasets are domain invariant by overlapping two distributions. However, it comes with the sacrifice of instance-level discriminability, as illustrated in Fig. 5 (middle). Hence, our goal is simultaneously aligning the image and instance levels, as shown in Fig. 5 (right).

A. Unsupervised Domain Adaptation

In this article, we perform UDA at local, global, and instance levels. The goal is to generate domain-invariant features at different levels of image features and perform better in unseen/target datasets. We also prove the performance gain from our proposed debiased contrastive loss in the learning phase. We denote the source as S and the target dataset as T. The CycleGAN network produces synthesized images (see input images in Fig. 4) from

source to target and vice versa. The synthesized images from source to target are denoted as S', where the object formation is the same as the source image, but the pixel color emulates the target dataset. On the other hand, T' denotes target-to-source conversion, where object formations are the target and pixel color follows the source domain. The DA with contrastive learning is performed bidirectional between (S, T') and (T, S') for better transferability and to minimize the domain discrepancies between the two datasets. Considering (S, T') and (T, S') as the source and target domain pairs, respectively, we take local features from the earlier stage of the backbone representing pixel-level and texture information and global features from the later part of the backbone, which means a more abstract version of the object. The authors performed only local–global (LG) domain adaptation in the baseline paper [43]. However, we take it further to instance-level transformation with pseudolabeling in the target dataset.

B. Support-Set Guided Pseudolabeling

Ground truth (GT) exists for the source dataset region proposals. GT is used to separate positive and negative samples in contrastive learning. We do not have any GT for the target dataset, so we must generate labels for the target proposals to guide contrastive learning. To perform pseudolabeling, the target domain instance feature vectors in a mini-batch are collected from the RPN module (see Fig. 4).

Early-stage target feature vectors are prone to background noise and mistake many background scenes as foreground objects. Therefore, we introduce a support-set guided curation step in the process that reduces the number of false positives from target object proposals. First, we take R samples from each of the C classes and create an *R-shot* support set to guide the labeling process. Here, the dimension of the *R-shot* support set is $R \prod C$. Then, we match all the features in a mini-batch with the support set using cosine similarity metrics. Next, we keep features that match any support samples passing some defined threshold. As features are less useful during early epochs, we restrict the number of unlabeled features for labeling to minimize computation time and the target instance contrastive loss. After every defined step size, we progressively increase the number of features by some factors for the pseudolabeling task. The curated features are then used for target pseudolabeling through a clustering method.

The K-means++ is an improved version of the original K-means clustering algorithm that aims to select better initial centroids in high dimensions and reduces the chance of the algorithm getting stuck to local optima compared to K-means [6]. Thus, we use K-means++ to generate pseudolabels through clustering from deep features. The clustering performance of the K-means++, as shown in Fig. 6, and the value of K for clustering is selected empirically. The selection process of K is described later in Section IV-F and Table I.

C. Debiased Contrastive Learning

Contrastive learning is a process of matching different distributions based on query (Q) and key (K) embeddings [44],

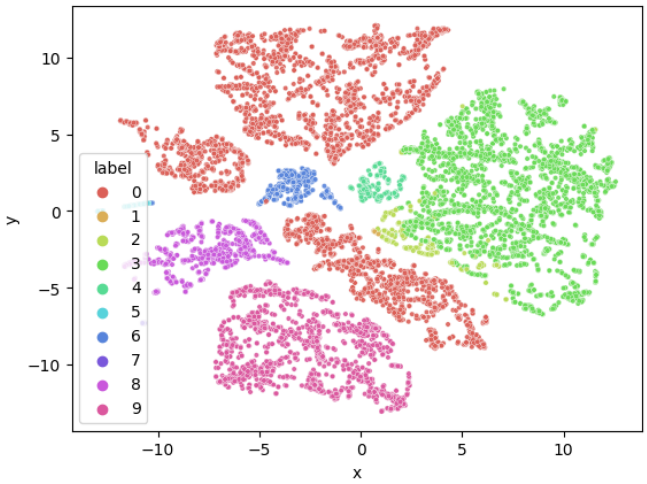

Fig. 6. Clustering visualization for pseudolabeling in 12 000 features over ten classes of the DOTA dataset.

TABLE I
TARGET DETECTION PERFORMANCE (MAP) WITH/WITHOUT AGGREGATED PSEUDO LABELING, CLUSTERING TIME, AND THE NUMBER OF CLUSTERS

Method	Cluster # DOTA	Cluster # UAVDT	Total Time(s)	DOTA (mAP)	UAVDT (mAP)
Without target Labeling	–	–	–	43.1	35.7
K-means++	2	1	0.3	48.4	39.1
K-means++	5	2	1.10	**50.6**	**41.5**
K-means++	10	3	2.94	44.0	37.2

The clustering time is given for a mini-batch of 4000 features from both the target datasets.

[46]. The value of the contrastive loss function is lower when there are high similarities between the query (Q) and positive key (K^+) pair and low similarities between the Query(Q) and negative keys (K^-) pairs. Contrastive learning performs domain alignment by keeping similar points closer and different points distant, as illustrated in Fig. 5. The most used formula for contrastive learning is outlined as follows:

$$\mathrm{CL} = -\log \frac{\exp(\mathrm{sim}(Q, K^+)/\tau)}{\sum_{i=1}^{N} \exp(\mathrm{sim}(Q, K_i^-)/\tau)} \tag{1}$$

where τ is a hyperparameter known as temperature to put penalties on the calculated similarities [45], [47].

The similarity can be calculated using cosine, Euclidian, or Wasserstein distance functions. The cosine similarity score is used in the experiments and calculated as $\mathrm{sim}(x, y)$ for two features x and y and is $\mathrm{sim}(x, y) = x^T/(||x|| * ||y||)$. We calculate query similarity CL in (1) as a normalized sum of the similarity of query vector Q to N negative samples. In the baseline paper [43], the authors used (1) for the LDA and GDA, where only a single augmented image was used as the positive case. However, earlier research shows that the work in [17] including more than one positive case in contrastive learning can better generalize the feature representation. Based on this idea, we modify the loss

TABLE II
INSTANCE DISTRIBUTION STATISTICS (TEST SET) OF THE DIOR [9], DOTA2.0 [48], [49], VISDRONE [50], AND UAVDT [51] DATASETS OVER DIFFERENT CATEGORIES

Class Name	# of instances DIOR	# of instances DOTA	# of instances Visdrone	# of instances UAVDT
Bridge	176	1039	–	–
Vehicle	2079	85479	–	–
Harbor	254	5704	–	–
Storage.T	2623	5416	–	–
Baseball	250	516	–	–
Car	–	–	14064	222650
Track	138	417	–	–
Basketball	171	358	–	–
Tennis	580	1662	–	–
Truck	–	–	750	4979
Stadium	40	393	–	–
Bus	–	–	251	6553
Airport	25	153	–	–

function in (1) as follows:

$$\mathrm{CL} = -\log \frac{\sum_{i=1}^{M} \exp(\mathrm{sim}(Q, K_i^+)/\tau)}{M * \sum_{j=1}^{N} \exp(\mathrm{sim}(Q, K_j^-)/\tau)} \tag{2}$$

where M is the number of augmented positive samples for the query. We perform a cross product between the query and positive cases following this operation $Q(1, \text{size}) \times K^+(M, \text{size})' = \text{Sim}(1, M)$, which gives a column vector with a dimension equal to positive cases (M). Then, we average all the logits and compute a single scalar value as the final similarity score. It is shown in Section IV that adding more than one positive case significantly improved the performance across different datasets.

Another challenge for contrastive learning is imbalance classes. Table II shows that the real datasets are highly imbalanced. As samples for contrastive learning are selected randomly, we cannot control which class instances are picked in a mini-batch. This raises the chances of getting FN picked as the negative samples, as illustrated in Fig. 3. Earlier DA methods for consumer datasets do not deal with this problem because consumer datasets are usually nearly balanced. On the other hand, remote sensing datasets are often dominated by some major classes that require extra effort to gain optimal results. The number of FNs increases as we increase the number of negative samples in a mini-batch

$$\mathrm{DCL} = -\log \frac{\frac{1}{M}\sum_{i=1}^{M} \exp(\mathrm{sim}(Q, K_i^+)/\tau)}{\sum_{j=1}^{N} \exp(D\_K_j^-/\tau)}. \tag{3}$$

In this light, we propose to filter out negative samples with high similarity scores with the query sample. In Fig. 3, three out of four images have a similarity score below 0.2, and one image is highly similar to the query image. DCL in (3) summarizes the process. First, reject the FN case that 70% matches the query. Next, replace the value with the remaining average score in the mini-batch for better consistency and stable learning. Here, DK_j^- is calculated using the following formula:

$$DK_j^- = \begin{cases} \mathrm{sim}(Q, neg), & \text{if } \mathrm{sim}(Q, neg) \leq 0.7 * \mathrm{sim}(Q, pos) \\ \mathrm{Avg.}(sim(Q, negs)), & \text{otherwise.} \end{cases}$$

D. Debiased Local Contrastive Learning

Local adaptation is a class-agnostic adaptation because we extract features at the pixel level of the source and target domains. From the architecture of our proposed model in Fig. 4, we can see that the first step toward LDA is to generate synthesized images from both the source (S) and target (T) images in a mini-batch. For that, we use CycleGAN and pass both the source and target images to generate translated source (S') and translated target (T'), respectively. Then, pass S, T', T, and S' to the backbone for feature extraction. Local features are saved from the earlier layers of the backbone in the dimension of $256 \times 100 \times 100$. Next, pass parts into the bottleneck block, which reduces the feature dimension to $32 \times 100 \times 100$, where dimensions are C, W, and H, respectively. Finally, we feed the output of the bottleneck layer to the multilayer perceptron (MLP) block and transform the final feature vector with a length of 1024. The minimal size of each feature reduces the necessity of GPU memory.

Let us represent the local features from the S, T', T, and S' as α_i^S, $\alpha_i^{T'}$, α_i^T, and $\alpha_i^{S'}$, respectively, where i is the index of the mini-batch. As we are going to perform bidirectional adaptation, for the adaptation of S and T', we select a local feature $\alpha_i^S \in \alpha^S$ as a query and choose different augmentations of the corresponding feature from $\alpha_i^{T'} \in \alpha^{T'}$ as the positive cases. On the other hand, negative cases are all other local features $\alpha_j^{T'} \in \alpha^{T'}$ in the mini-batch, where $j \neq i$. The bidirectional local contrastive loss between (S and T') and (T and S') can be calculated from the following equations:

$$\begin{aligned} \mathrm{DCL}_{\mathrm{local}}^{S,T'} = & -\log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} \exp(\mathrm{sim}(\alpha_i^S, \alpha_m^{T'})/\tau)}{D(\sum_{j=1}^{\nu} \exp(\mathrm{sim}(\alpha_i^S, \alpha_j^{T'})/\tau))} \\ & -\log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} \exp(\mathrm{sim}(\alpha_i^{T'}, \alpha_m^S)/\tau)}{D(\sum_{j=1}^{\nu} \exp(\mathrm{sim}(\alpha_i^{T'}, \alpha_j^S)/\tau))}, \quad j \neq i \end{aligned} \tag{4}$$

$$\begin{aligned} \mathrm{DCL}_{\mathrm{local}}^{T,S'} = & -\log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} \exp(\mathrm{sim}(\alpha_i^T, \alpha_m^{S'})/\tau)}{D(\sum_{j=1}^{\nu} \exp(\mathrm{sim}(\alpha_i^T, \alpha_j^{S'})/\tau))} \\ & -\log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} \exp(\mathrm{sim}(\alpha_i^{S'}, \alpha_m^T)/\tau)}{D(\sum_{j=1}^{\nu} \exp(\mathrm{sim}(\alpha_i^{S'}, \alpha_j^T)/\tau))}, \quad j \neq i. \end{aligned} \tag{5}$$

In (4) and (5), D stands for *Debiased*, and m denotes the mth augmentation out of μ number of augmentations for a particular image. Finally, the number of negative examples drawn from a mini-batch is denoted with ν. The total bidirectional local DA loss can be formulated by accumulating the loss for all the query images in a mini-batch, as follows:

$$\begin{aligned} \mathrm{DCL}_{\mathrm{local}} = & W_1 * \mathrm{DInfoNCE}_{\mathrm{local}}^{S,T'} \\ & + W_1 * \mathrm{DInfoNCE}_{\mathrm{local}}^{T,S'}. \end{aligned} \tag{6}$$

E. Debiased Global Contrastive Learning

GDA focuses more on the abstract view of object features. Global image features are collected from the last layer of the

backbones; by this, we get features with very high details on lower spatial resolutions. Like the local adaptation, we also pass the images of dimension $256 \times 25 \times 25$ to the bottleneck layer and reduce the dimension to $3 \times 25 \times 25$. Next, features are fed to the MLP block, and a feature vector with 1024 dimensions is computed. Following the same notational format from previous Section III-D, we can define the global features from S, T', T, and S' as β_i^S, $\beta_i^{T'}$, β_i^T, and $\beta_i^{S'}$, respectively. Again, i is the index number in a mini-batch. Therefore, the bidirectional global contrastive loss between (S and T') and (T and S') can be presented as follows:

$$\mathrm{DCL}_{\mathrm{global}}^{S,T'} = -\log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} \exp(\mathrm{sim}(\beta_i^S, \beta_m^{T'})/\tau)}{D(\sum_{j=1}^{\nu} \exp(\mathrm{sim}(\beta_i^S, \beta_j^{T'})/\tau))} - \log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} \exp(\mathrm{sim}(\beta_i^{T'}, \beta_m^S)/\tau)}{D(\sum_{j=1}^{\nu} \exp(\mathrm{sim}(\beta_i^{T'}, \beta_j^S)/\tau))}, \quad j \neq i \tag{7}$$

$$\mathrm{DCL}_{\mathrm{global}}^{T,S'} = -\log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} \exp(\mathrm{sim}(\beta_i^T, \beta_m^{S'})/\tau)}{D(\sum_{j=1}^{\nu} \exp(\mathrm{sim}(\beta_i^T, \beta_j^{S'})/\tau))} - \log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} \exp(\mathrm{sim}(\beta_i^{S'}, \beta_m^T)/\tau)}{D(\sum_{j=1}^{\nu} \exp(\mathrm{sim}(\beta_i^{S'}, \beta_j^T)/\tau))}, \quad j \neq i. \tag{8}$$

The total bidirectional GDA loss can be formulated by accumulating the loss for all the query images in a mini-batch, as follows:

$$\mathrm{DCL}_{\mathrm{global}} = W_2 * \mathrm{DCL}_{\mathrm{global}}^{S,T'} + W_2 * \mathrm{DCL}_{\mathrm{global}}^{T,S'}. \tag{9}$$

F. Debiased Instance Contrastive Learning

LG contrastive learning helps to create domain-invariant features, as shown in Fig. 5; it is visible in the figure that *image-level* adaptation can remove the domain boundary and create a uniform domain feature space for source and target datasets. No class discrepancy is maintained at the image-level alignment, and there is an overlap between different class instances in the feature space. To solve this issue, we propose to perform debiased instance contrastive learning for the source and target datasets and achieve class discrepancy in features. The effect of this learning is illustrated in Fig. 5, where we can see a moderate separation line between the two classes.

Instance-level features are extracted from the RPN and fed into the instance domain adaptation (IDA) block. It is important to note that we do not perform strong feature alignment for samples near the decision boundaries. Instead, we perform weak feature alignment to maintain classwise discriminant in visual features. Instances near decision boundaries may look very similar but belong to different classes.

Notation for the source region proposals is Γ_i^S, and for the target region proposals is Γ_i^T. The corresponding class set for the source is C_i^S, and for the target is C_i^T; i is the proposal index among P proposals. For instance-level contrastive learning, the formula can be formulated from the following equations:

TABLE III
ABLATION STUDY FOR DIFFERENT MODULES OF OUR DCLDA METHOD

Method	CGAN	LDA	GDA	IDA	DOTA	UAVDT
Baseline					35.4	26.4
w/CGAN	✓				37.2	27.8
w/LDA	✓	✓			41.6	30.2
w/GDA	✓		✓		44.5	34.6
w/IDA	✓			✓	46.9	36.8
DCLDA W3 = 0.01	✓	✓	✓	✓	**50.6**	**41.5**
DCLDA W3 = 0.1	✓	✓	✓	✓	**48.2**	**37.9**
DCLDA W3 = 0.5	✓	✓	✓	✓	**42.5**	**30.3**

Here, CGAN = CycleGAN transfer learning, LDA = local domain adaptation, GDA = global domain adaptation, and IDA= instance-level domain adaptation.

$$\mathrm{DCL}_{\mathrm{Ins}}^{S} = -\log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} \exp(\mathrm{sim}(\Gamma_{(qc,i)}^S, \Gamma_{(pc,m)}^S)/\tau)}{D(\sum_{n=1}^{\nu} \exp(\mathrm{sim}(\Gamma_{(qc,i)}^S, \Gamma_{(nc,n)}^S)/\tau))} \quad i \neq m \text{ and } i \neq n \tag{10}$$

$$\mathrm{DCL}_{\mathrm{Ins}}^{T} = -\log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} \exp(\mathrm{sim}(\Gamma_{(qc,i)}^T, \Gamma_{(pc,m)}^T)/\tau)}{D(\sum_{n=1}^{\nu} \exp(\mathrm{sim}(\Gamma_{(qc,i)}^T, \Gamma_{(nc,n)}^T)/\tau))} \quad i \neq m \text{ and } i \neq n. \tag{11}$$

Equations (10) and (11) represent the source and target instance losses, respectively. Here, μ and ν stand for the number of positive and negative samples, respectively, and i stands for $i\text{th} \in the P$ proposal in the proposal set P. We define the class id of the query, positive, and negative samples using qc, pc, and nc, respectively. The total instance contrastive loss can be formulated by accumulating the loss for all the region proposals in a mini-batch, as follows:

$$\mathrm{DCL}_{\mathrm{Ins}} = W_3 * \mathrm{DCL}_{\mathrm{Ins}}^{S} + W_3 * \mathrm{DCL}_{\mathrm{Ins}}^{T}. \tag{12}$$

Also, confidence tends to be less reliable at the early stage of the adaptation. The feature quality and objectness score from the RPN for the target dataset are generally less reliable due to the large domain gap. In this light, we use weights W_1, W_2, and W_3 in (6), (9), and (12), respectively, to perform progressive adaptation and give less weight during the early stage of transformation and progressively increase the focus with an increased object confidence score and quality features. Earlier works show that LDA and GDA work well with an initial weight of 0.1, so we keep $W1$ and $W2 = 0.1$. For the IDA, we tried different values of $W3$ as presented in Table III. However, the optimal result was achieved with an initial value of 0.01. The total loss for the detection and adaptation process can be calculated by summarizing all the loss components outlined as follows:

$$\mathrm{TotalLoss} = \mathrm{SWFL}(x, p_t, y) + \mathrm{DCL}_{\mathrm{local}} + \mathrm{DCL}_{\mathrm{global}} + \mathrm{DCL}_{\mathrm{Ins}}. \tag{13}$$

IV. Experiments

This section evaluates our proposed DCL model against current SOTA DA methods on four RSI datasets. The experimental setup is described in Section IV-A, the comparison findings are summarized in Section IV-D, and the extensive ablation studies over different factors and parameters are outlined in Section IV-F.

A. Implementation Details

We use the object classification pipeline similar to [43]: 1) Darknet53 as the backbone as it is shown to preserve semantic information from the small objects than the residual-based feature extractor networks [27], [54]; 2) RPN heatmap-based approach to identify dense small objects and remove NMS; and 3) the detection block is faster-region-based convolutional neural network (RCNN) [55]. We have used Python with PyTorch as the deep learning framework to implement the project. Our code implementation is heavily based on an open-source computer vision library *Detectron2* [56] and some part of *SOD* [30] implementations. With DCL, we implemented three new DA modules for LDA, GDA, and IDA. Also, we implemented a Cythonized K-means++ that is much faster than the Python implementation, and the clustering time is recorded in Table I.

B. Hyperparameter Settings

In CycleGAN network [57], load 800 and crop 640 were used for the data augmentation. To train our DCLDA model, we have resized all the images to 800×800 pixels and set eight as the mini-batch size in each epoch. Therefore, in total, we send $8 \times 4 = 32$ images in a mini-batch to train the DCLDA model. The PyTorch color-jitter augmentation technique was used to create multiple augmented copies of the synthesized images for image-level contrastive learning positive cases. During the support-guided pseudolabeling, we chose five samples (n) per class and created the five-shot support set. For the feature curation, we tried different values as the cosine similarity threshold and found that 70% cosine similarity threshold achieves optimal performance across most of the experiments. Other important hyperparameters were set: IOU $= 0.5$, NMS $= 0.6$, L.Rate $= 0.003$, POST_NMS_TOPK for IDA $= 64$, and POSITIVE_FRACTION $= 0.40$. We have used NVIDIA 2 x RTX 6000 GPU with 49 GB of memory, 11th-generation Intel CoreTM i9-11900K @ 3.50 GHz $\times$ 16 CPU, and 167 GB of system memory to carry out all experiments.

C. Datasets and Evaluation Metrics

1) Datasets: The *DIOR* dataset originally consisted of 24 500 Google Earth images from 80 countries. After selecting only common classes, the reduced dataset has 11 402 images. The images varied in quality and were captured in different seasons and weather conditions. The number of pictures in the training set is 10 888; in the testing set, we have 512 images. The *DOTA* dataset comprises 2430 overhead images with image sizes ranging from 800×800 to $29\,200 \times 27\,620$ pixels. The ground sample distance in the dataset ranges from 0.1 to 0.87 m, and each image contains an average of 220 objects. For experiments, we split high-resolution images into patches of size 1024×1024 pixels with an overlap of 200 pixels. Considering only the common ten classes, the DOTA2.0 training set has 11 551 images, and the testing set has 3488 images. *Visdrone* is a unmanned aerial vehicle (UAV) dataset containing over 10 000 image frames from more than 6 h of videos, making it one of the largest drone datasets available. The experimental dataset includes three common object categories, and the images have different resolutions ranging from 540p to 1080p. The training and testing sets contain 6883 and 546 images, respectively. The *UAVDT* dataset contains over 80 000 frames in 179 videos captured by UAVs, making it one of the largest datasets available for object detection. The experimental dataset contains 10 000 images with three object categories with different image resolutions ranging from 540p to 1080p. The dataset covers various weather conditions, including sunny, cloudy, and rainy. The $\rightarrow$ symbol is illustrating the direction of DA: *source* $\rightarrow$ *target*.

2) Evaluation Metrics: To assess the effectiveness of our proposed approach in the target domain, we measure its precision, recall, and average precision (AP) by considering both precision and recall for each object category. The mean average precision (mAP) is then calculated as the average AP across all the object categories. The mAP for all the experiments was calculated with an IOU of 0.5 at the NMS stage.

D. Method Performance Comparisons

We compare our DCLDA method with several current SOTA techniques for the adaptive object detection task on two high-variability video image datasets and two high-variability image datasets. Specifically, we have used the CenterNet2 [26] as the source-only baseline, which is trained only with labeled source data, serving as the performance lower bound for comparisons. On the other hand, the *oracle* method is trained with labeled target data, serving as the performance upper bound. We have used feature alignment DA methods, such as MGADA [52] and SAPNet [53], and a spatial-attention-based DA network for the performance measurements. A novel SIGMA method [39] is also introduced in the model comparison to have better diversity in the plans. Finally, we introduced ConfMix [38], a sample mixing-based paradigm of DA for SOTA comparisons. Fig. 7 presents the qualitative analysis and the detection performance of DCLDA trained on DIOR source data and tested on the DOTA target dataset. In detection figures, we illustrate the GT, foreground-focused saliency map, and object detection for different samples. It is evident from Fig. 7 that our DCLDA model well adopted the variation in lighting conditions, object sizes, and foggy-weather conditions between source and target datasets. Table IV presents the quantitative performance comparison for DIOR and DOTA satellite image datasets. This table shows classwise performance for the target dataset and overall performance for both the source and target datasets. We can see from Table IV that our baseline method achieves mAP of 66.6 and 35.4 in the source and target datasets, respectively. We improve the baseline model with image-level LDA and GDA and pseudolabeling-based instance adaptation, which helps us

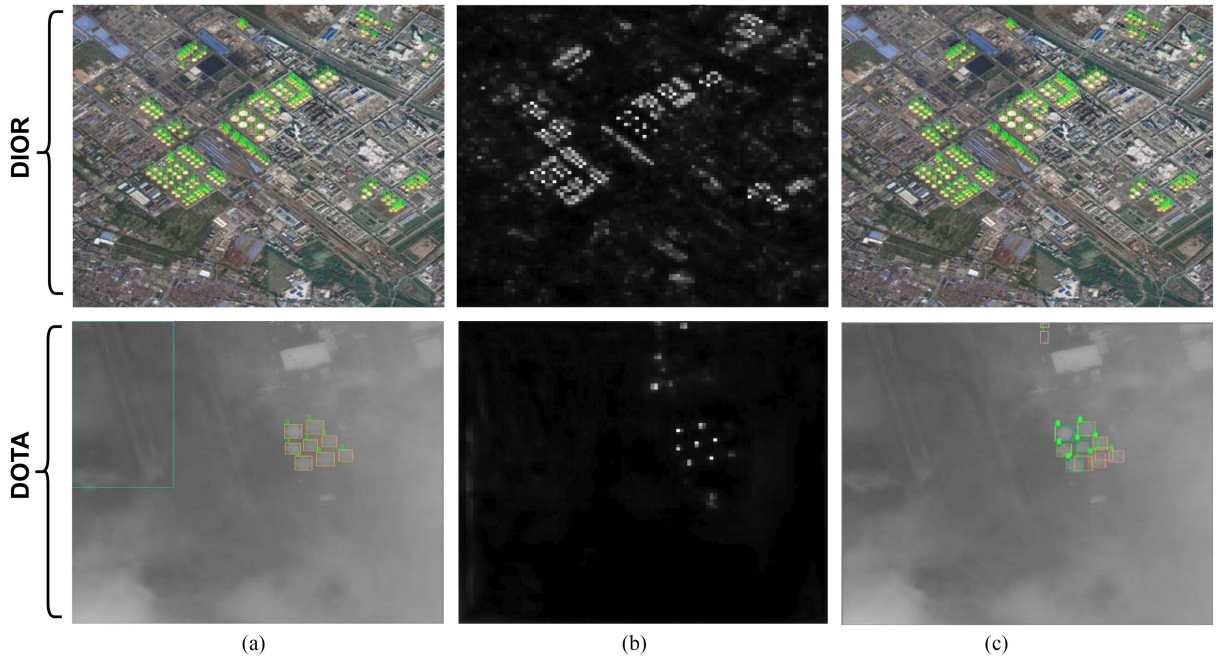

Fig. 7. Detection results from DIOR (source) and DOTA (target) dataset using our DCLDA method. (a) GT. (b) Saliency map. (c) Prediction.

TABLE IV
CLASSWISE PERFORMANCE COMPARISONS (MAP) FOR DIOR → DOTA BENCHMARK(IOU = 0.5), AS MEASURED BOTH ON THE DIOR (SOURCE) AND ON THE DOTA (TARGET) DATASETS

Method	Detector+ Backbone	Bridge	Vehicle	Harbor	Storage Tank	Baseball Field	Track Field	B.Ball Court	Tennis Court	Stadium	Airport	DIOR → mAP	DOTA mAP
Baseline [26]	CenNet2 ResNet-50	10.1	9.7	46.7	42.9	50.1	34.9	49.3	77.6	0.0	33.0	66.6	35.4
MGADA [52]	FCOS VGG-16	13.3	11.3	46.8	47.2	47.4	38.4	50.0	85.8	0.0	37.0	68.2	37.7
SAPNET [53]	FCOS ResNet	7.8	9.2	18.1	20.2	35.5	24.7	29.2	74.7	0.0	19.3	55.1	26.5
MGADA [52]	Faster-RCNN ResNet-101	15.9	12.0	50.7	46.5	47.6	39.3	52.3	89.6	0.0	37.9	73.1	39.2
SIGMA [39]	FCOS ResNet50	27.0	32.6	64.5	65.0	55.4	56.6	62.3	91.9	1.3	34.7	77.2	47.1
ConfMix [38]	YOLOv5 CSP Darknet53	27.2	32.0	65.9	65.1	56.3	56.3	61.5	93.5	1.0	34.9	78.8	47.4
DCLDA*	CenNet2 ResNet-50	27.0	28.7	68.1	66.6	52.4	51.1	63.0	90.2	5.6	36.0	81.4	49.1
DCLDA	CenNet2 CSP Darknet53	30.1	28.8	70.0	65.8	55.4	52.5	62.2	93.2	7.9	37.3	**82.7**	**50.6**
Oracle	Baseline	46.4	40.1	83.1	65.8	64.4	60.0	77.7	94.9	27.2	54.3	62.7	**62.8**

to outperform other SOTA models by a minimum margin of 3.2% on the target dataset. Moreover, the gap between the DCLDA and Oracle results is now narrowed to 12.2% from 27.4%. From the classwise performance, we notice that while other methods ultimately failed to affect the stadium class, our DCLDA method showed a significant gain of 7.9% mAP of this particular class. It is also visible that CSP-Darnet53 can perform better than the ResNet50 model with +1.5% of target mAP improvement. Finally, the precision, recall, and F1 scores are presented in Table V.

The Visdrone and UAVDT video datasets are two high-variability videos captured from the UAV in Table VI. The qualitative analysis of Visdrone and UAVDT datasets is presented in Fig. 8. Visdrone and UAVDT pose critical domain gaps due to illumination, low light, and foggy conditions. Fig. 8 shows samples with shadows due to high buildings and sunlight angles. In addition, we see some samples where the objects are overexposed with traffic lights, and some are underexposed due to low illuminations. It can be seen from Fig. 8 that our proposed DCLDA can tackle all these critical scenarios and detect objects successfully. Next, we evaluate the target dataset performance over three different categories. We have not only shown excellent performance on the target dataset but have also achieved a 59.2% mAP (see Table VI) on the source dataset, which is noteworthy.

TABLE V
COMPARISON OF PRECISION, RECALL, AND F1 SCORE BETWEEN THE CLOSEST SOTA COMPETITOR AND OUR PROPOSED MODEL FOR THE EXPERIMENTAL DATASETS

	Precision				Recall				F1			
Dataset	DIOR	DOTA	Visdrone	UAVDT	DIOR	DOTA	Visdrone	UAVDT	DIOR	DOTA	Visdrone	UAVDT
ConfMix	80.1	62.7	68.4	44.8	75.5	48.8	50.2	46.3	77.7	54.4	57.9	45.7
DCLDA	**85.9**	65.0	72.3	47.4	**78.5**	48.5	53.4	50.6	**82.0**	55.5	61.4	48.9

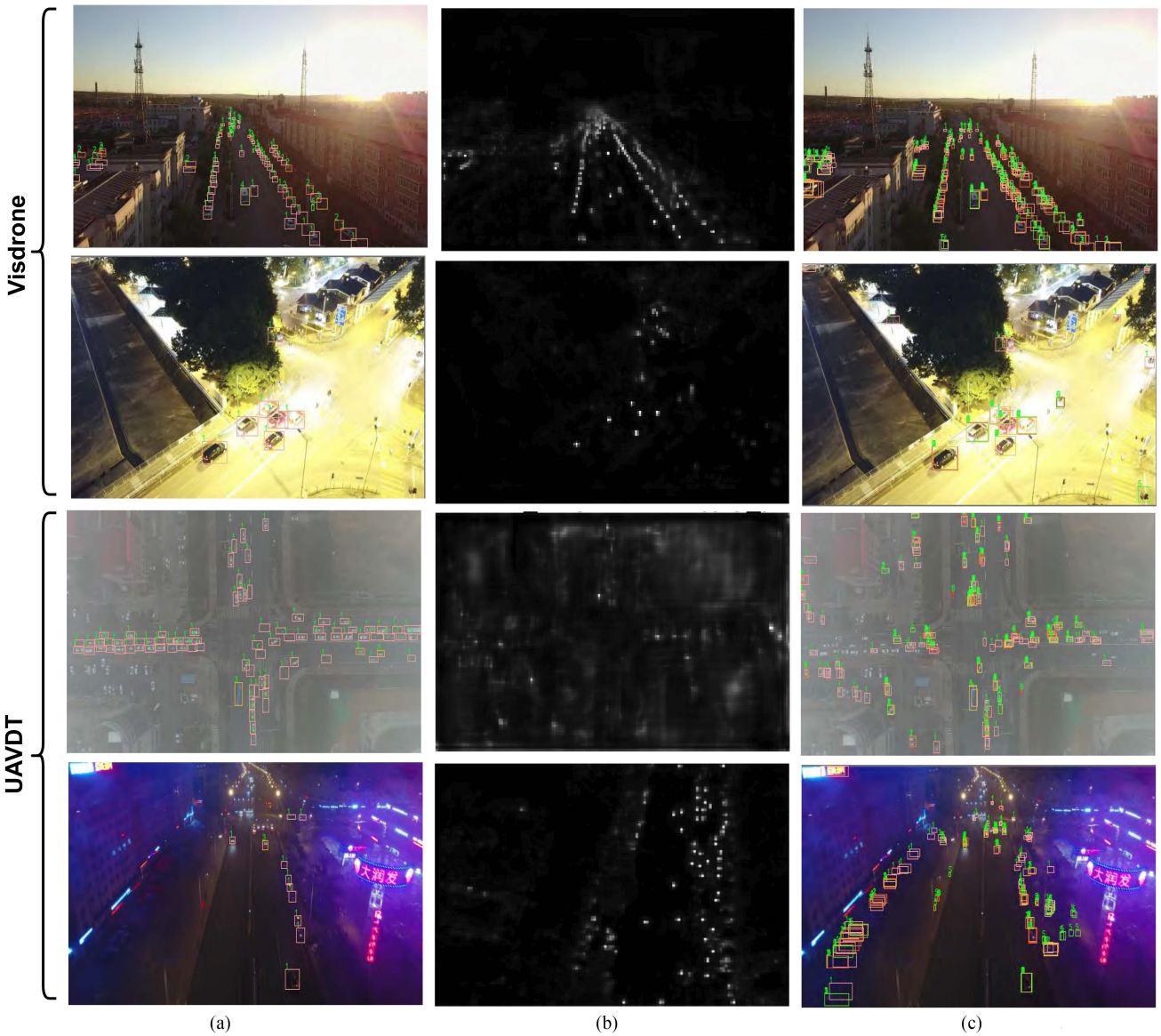

Fig. 8. Detection results from Visdrone (source) and UAVDT (target) datasets using our DCLDA method. (a) GT. (b) Saliency map. (c) Prediction.

Our baseline method trained on only source data gives 26.4% of mAP, whereas our DCLDA method achieves 41.5% of mAP using DCL and pseudolabeling. Also, we have a +2.1% gain margin compared to the best SOTA *ConfMix* method. Moreover, using DCL, we could shrink the performance gap between the oracle and our model from 30.5% to 15.4% compared to the baseline model. Table VI also demonstrates that a well-designed backbone can enhance performance by around +2.7% on the video target domain with dominated dense objects.

In Fig. 9, we present a qualitative analysis of DCLDA with other competitive SOTA methods. The green boxes denote correct foreground object detection, and the yellow boxes refer to missed object detection. From Fig. 9, we can see that our DCLDA performs significantly better in detecting challenging small and dark objects. However, we found some missing detection from DCLDA when the object has a uniform color distribution (e.g., green field or tennis court). On the other hand, the SIGMA, MGADA, and ConfMix methods can do well on

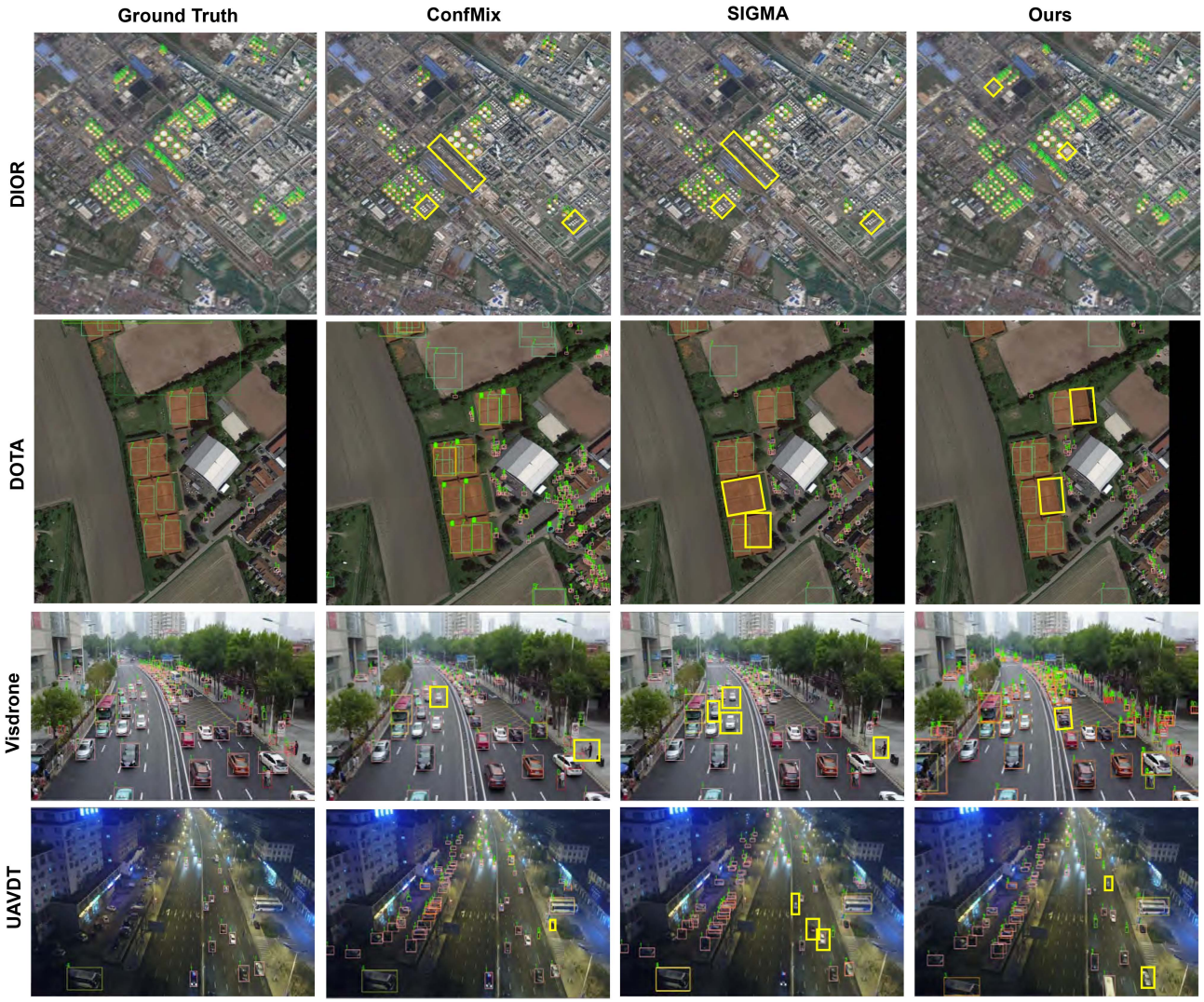

Fig. 9. Comparison of SOTA methods with our DCLDA method. In this figure, we present the comparison across different methods and datasets to illustrate the effectiveness of our model. The green boxes denote true predictions, and the yellow boxes denote missed detections.

TABLE VI
CLASSWISE PERFORMANCE COMPARISONS (MAP) FOR VISDRONE → UAVDT BENCHMARK (IOU = 0.5)

Method	Car	Truck	Bus	VISDRONE → UAVDT	
Baseline [26]	34.2	7.6	29.3	48.1	26.4
MGADA [52]	42.0	15.6	36.4	51.9	31.3
SAPNET [53]	31.5	7.9	22.7	26.3	20.8
MGADA [52]	39.2	12.6	35.8	54.6	29.2
SIGMA [39]	50.1	20.9	45.5	46.3	38.9
ConfMix [38]	51.3	20.4	46.0	46.5	39.4
DCLDA*	52.5	23.0	41.1	58.4	38.8
DCLDA	54.2	24.4	46.3	**59.2**	**41.5**
Oracle	72.4	37.8	60.5	45.6	**56.9**

regular-sized objects, but we can find that there are still several false alarms in the detection results, as they fail to align the source and target domains properly.

Finally, each dataset's precision, recall, and F1 scores are presented in Table V. We compared the performance of our DCLDA with the best competitor, ConfMix. We achieved better precision and recall for most datasets, except for DOTA, where ConfMix slightly outperforms DCLDA. We can also verify the recall performance from Figs. 7 and 8, which shows the foreground detection results from experimental datasets.

E. Computational Cost Comparisons

The DA methods are well known for their high computational cost (see Table VII). However, carefully designing the gradient computation tree helps our DCLDA method maintain reasonable computational stability with optimal detection performance. Table VII presents a computational comparison between the

TABLE VII
MULTIFACTOR COMPUTATIONAL COST COMPARISON BETWEEN OUR PROPOSED DCLDA AND RECENT SOTA METHODS

Method	# of parameters	GFLOPS	# of Layers
SIGMA	45354466	33.5	321
MGADA step 1	66974671	26.9	119
MGADA step 2	66974671	26.9	119
ConfMix step 1	7057387	16.1	270
ConfMix step 2	7047883	15.9	213
DCLDA	53592866	34.8	196

TABLE VIII
SOURCE AND TARGET DETECTION PERFORMANCE (MAP) WITH (W/) AND WITHOUT (W/O) IDA

Method	Curation	DIOR → DOTA		VISDRONE → UAVDT	
		mAP	mAP	mAP	mAP
w/o IDA	NA	84.2	42.7	58.4	36.1
w/ IDA	–	82.7	47.8	56.0	40.1
w/ IDA	✓	83.4	50.6	58.2	41.5

proposed and some closely competitive SOTA models. Among the models, our DCLDA and SIGMA are end-to-end trainable models. On the other hand, ConfMix and MGADA are two-step trainable methods. The MGADA is the most computationally expensive model, with 53.8 GFLOPS, whereas the ConfMix is the most computationally efficient, with 32 GFLOPS. Although our DCLDA requires 34.8 GFLOPS, it outperforms the ConfMix method in object detection tasks by 3.2% and 2.1% for DOTA and UAVDT target datasets, respectively. To reduce the learnable parameters and GFLOPS, we turn ON gradient updates only for the query vectors and no gradient updates for positive and negative vectors during contrastive learning. Also, we subsample the positive and negative keys throughout all contrastive learning to reduce training time and computation cost further. The training time for ConfMix and DCLDA is 12.3 and 13.4 h, respectively, for 50 epochs. Therefore, we can conclude that exploring contrastive learning for DA tasks is computationally convincing with the careful design of the gradient computation graph.

F. Ablation Study

In this section, we answer several questions. The first one is: *Does the instance-level adaptation help on target data?* Table VIII shows that the IDA improves mAP 7.9% and 5.4% recorded for the DOTA and UAVDT target datasets, respectively. The performance on the source dataset dropped slightly by 1.5% for the DIOR dataset after IDA (w/o curation) due to the increased number of loss functions and noise from target instance labels. When we used the support set to cure the noisy features and guide the IDA process, we gained higher mAP in the target dataset. We could recover from the source dataset performance drop (see Table VIII). The second question we want to answer is: *how much we benefit from using multiple positive cases?* We claim that the single sample of positive cases for contrastive learning does not work for high-variability overhead videos and imagery. Table IX illustrates the performance gain, and even for two positive samples, improves the overall performance by roughly 2.0% for both the target datasets.

TABLE IX
QUANTITATIVE PERFORMANCE COMPARISONS (MAP) FROM *DCLDA* MODEL FOR VARIOUS NEGATIVE AND POSITIVE CASE VALUES

# neg	# pos	DIOR → DOTA		Visdrone → UAVDT	
		mAP	mAP	mAP	mAP
4	1	77.3	46.5	53.7	38.0
4	2	78.2	48.2	53.1	39.9
15	8	80.5	47.4	55.6	38.3
7	4	**82.7**	**50.6**	**59.2**	**41.5**

More positive and negative examples can introduce more noise and ultimately hamper the results, as illustrated in Table IX for 15 negative and eight positive cases. The study found that using seven negative and four positive points gives the optimal results for each dataset. The third question is: *how many clusters do we set for pseudolabeling?* and Table I shows that pseudolabeling with five clusters for DOTA and two for UAVDT can achieve up to 7.5% and 5.8% increase, respectively. Table II shows that five significant classes dominate the DOTA dataset labels. For UAVDT, a single class with two minor classes separates the dataset into two clusters for target labeling.

Finally, we answer the efficacy of different modules of the proposed DCLDA model. Table III shows that each integrated module has some performance gain in our target dataset. We recorded the mAP performance against the experimental dataset. We first integrated CycleGAN-based synthetic image for transfer learning, and we can see that it gains +1.8% and +1.4% mAP on DOTA and UAVDT datasets, respectively. Next, we integrated three contrastive learning modules (e.g., *LDA*, *GDA*, and *IDA*) incrementally, and the performance is presented in Table III. Integrating the *IDA* module obtains the best performance gain. The proposed model gains +11.5% and 10.4% increase in the mAP on the DOTA and UAVDT datasets, respectively. Finally, we combined all the proposed modules in our *DCLDA* architecture and ran experiments with different hyperparameter values (W3). DCLDA is very sensitive to W3, and we noticed a significant performance drop when weighing the IDA close to 50%. The optimal performance on both the target datasets was recorded by carefully selecting all the hyperparameters and setting W3 equal to 0.01 or 1%.

V. CONCLUSION

This article proposes specialized contrastive learning with support-set guided pseudolabeling for the UDA task. We show that remote sensing video frames and images have significant domain shifts due to lighting conditions, weather changes, and geographical variance. A careful design of the detection pipeline and the instance-aware DA method is required for optimal performance. Our proposed contrastive learning method consists of two significant improvements. The first is DCL to remove FN samples using the classwise probability logits. The second

introduces multiple augmented positive cases for more stability from object size and scale variation over images and datasets. Next, we show that a faster and support-guided pseudolabeling technique can improve the target instance learning performance by eliminating noisy object features with little training time overhead. Specifically, our method takes only a second to label 4000 target features in a mini-batch. Finally, we validate our approach in four challenging high-variability datasets that showed significant performance gain over available SOTA methods. For the UAVDT and DOTA target datasets, we outperformed the latest SOTA *ConfMix* method by +2.1% and +3.2% mAP, respectively. We hope our work can inspire future exploration of DA tasks in remote sensing imagery using DCL. In the future, we plan to make the model more computationally efficient and further pursue the category imbalance problem in RSIs for improved detection performance. Besides, we plan to introduce the first-ever multimodal image-text-based DA pipeline for RSI imagery.

ACKNOWLEDGMENT

The NVIDIA RTX 6000 GPU used for this research was donated by NVIDIA Corporation. This article's views, opinions, and findings are those of the authors. They should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. government.

REFERENCES

[1] M. Kumar, P. Singh, and P. Singh, "Machine learning and GIS-RS-based algorithms for mapping the groundwater potentiality in the Bundelkhand region, India," *Ecol. Informat.*, vol. 74, 2023, Art. no. 101980.

[2] J. Valente, B. Sari, L. Kooistra, H. Kramer, and S. Mücher, "Automated crop plant counting from very high-resolution aerial imagery," *Precis. Agriculture*, vol. 21, pp. 1366–1384, 2020.

[3] S. Workman and N. Jacobs, "Dynamic traffic modeling from overhead imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12315–12324.

[4] D. Biswas, M. Rahman, Z. Zong, and J. Tešić, "Improving the energy efficiency of real-time DNN object detection via compression, transfer learning, and scale prediction," in *Proc. IEEE Int. Conf. Netw., Archit. Storage*, 2022 pp. 1–8.

[5] BookingHunterTV, *New York City Walking Tour Part 1—Midtown Manhattan*, Dec. 2019. [Online]. Available: https://youtu.be/-IpXdtWfneI

[6] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2007, pp. 1027–1035.

[7] C. Yan, X. Chang, M. Luo, H. Liu, X. Zhang, and Q. Zheng, "Semantics-guided contrastive network for zero-shot object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, doi: 10.1109/TPAMI.2021.3140070.

[8] Y. Ding et al., "Self-supervised locality preserving low-pass graph convolutional embedding for large-scale hyperspectral image clustering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536016.

[9] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159 pp. 296–307, 2020.

[10] Y. Chen, Q. Liu, T. Wang, B. Wang, and X. Meng, "Rotation-invariant and relation-aware cross-domain adaptation object detection network for optical remote sensing images," *Remote Sens.*, vol. 13, 2021, Art. no. 4386.

[11] D. Lam et al., "xView: Objects in context in overhead imagery," 2018, *arXiv:1802.07856*.

[12] J. Chen, J. Zhu, Y. Guo, G. Sun, Y. Zhang, and M. Deng, "Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5616915.

[13] I. Georgoulas, E. Protopapadakis, K. Makantasis, D. Seychell, A. Doulamis, and N. Doulamis, "Graph-based semi-supervised learning with tensor embeddings for hyperspectral data classification," *IEEE Access*, vol. 11, pp. 124819–124832, 2023.

[14] J. Zheng, Y. Zhao, W. Wu, M. Chen, W. Li, and H. Fu, "Partial domain adaptation for scene classification from remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601317.

[15] M. M. M. Rahman and J. Tešić, "Hybrid approximate nearest neighbor indexing and search (HANNIS) for large descriptor databases," in *Proc. IEEE Int. Conf. Big Data*, 2022, pp. 3895–3902.

[16] I. O. Tolstikhin, B. K. Sriperumbudur, and B. Schölkopf, "Minimax estimation of maximum mean discrepancy with radial kernels," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1938–1946.

[17] C. Chuang, J. Robinson, Y. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 8765–8775.

[18] Y. Ding et al., "Unsupervised self-correlated learning smoothy enhanced locality preserving graph convolution embedding clustering for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536716.

[19] M. Luo and S. Ji, "Cross-spatiotemporal land-cover classification from VHR remote sensing images with deep learning based domain adaptation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 191, pp. 105–128, 2022.

[20] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[21] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 1923–1938, Apr. 2019.

[22] L. Xiong, M. Ye, D. Zhang, Y. Gan, and Y. Liu, "Source data-free domain adaptation for a faster R-CNN," *Pattern Recognit.*, vol. 124, 2022, Art. no. 108436.

[23] T. Jui, G. Bejarano, and P. Rivas, "A machine learning-based segmentation approach for measuring similarity between sign languages," in *Proc. 10th Workshop Represent. Process. Sign Lang.: Multilingual Sign Lang. Resources*, 2022, pp. 94–101.

[24] M. Babavalian and K. Kiani, "Learning distribution of video captions using conditional GAN," *Multimedia Tools Appl.*, vol. 83, pp. 1–23, 2023.

[25] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2778–2788.

[26] X. Zhou, V. Koltun, and P. Krähenbühl, "Probabilistic two-stage detection," 2021, *arXiv:2103.07461*.

[27] A. Bochkovskiy, C. Wang, and H. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[28] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[29] L. Shi, L. Kuang, X. Xu, B. Pan, and Z. Shi, "CANet: Centerness-aware network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603613.

[30] D. Biswas and J. Tešić, "Small object difficulty (SOD) modeling for objects detection in satellite images," in *Proc. 14th Int. Conf. Comput. Intell. Commun. Netw.*, 2022, pp. 125–130.

[31] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[32] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32 pp. 364–376, 2022.

[33] J. Zhang, Y. Shi, Q. Zhang, L. Cui, Y. Chen, and Y. Yi, "Attention guided contextual feature fusion network for salient object detection," *Image Vis. Comput.*, vol. 117, 2022, Art. no. 104337.

[34] Y. Wu, K. Zhang, J. Wang, Y. Wang, Q. Wang, and X. Li, "GCWNet: A global context-weaving network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619912.

[35] Q. Li, Y. Chen, and Y. Zeng, "Transformer with transfer CNN for remote-sensing-image object detection," *Remote Sens.*, vol. 14, pp. 984, 2022.

[36] Z. Deng, Q. Kong, N. Akira, and T. Yoshinaga, "Hierarchical contrastive adaptation for cross-domain object detection," *Mach. Vis. Appl.*, vol. 33, pp. 1–13, 2022.

[37] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8869–8878.

[38] G. Mattolin, L. Zanella, E. Ricci, and Y. Wang, "ConfMix: Unsupervised domain adaptation for object detection via confidence-based mixing," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 423–433.
[39] W. Li, X. Liu, and Y. Yuan, "SIGMA: Semantic-complete graph matching for domain adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5291–5300.
[40] C. Wu, F. Wu, and Y. Huang, "Rethinking InfoNCE: How many negative samples do you need?" 2021, *arXiv:2105.13003*.
[41] G. Bai, W. Xi, X. Hong, X. Liu, Y. Yue, and S. Zhao, "Robust and rotation-equivariant contrastive learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, doi: 10.1109/TNNLS.2023.3243258.
[42] H. Li et al., "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618014.
[43] D. Biswas and J. Tešić, "Progressive domain adaptation with contrastive learning for object detection in the satellite imagery," 2022, *arXiv:2209.02564*.
[44] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1735–1742.
[45] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
[46] T. Chen, S. Kornblith, M. Norouzi, and G. A. Hinton, "simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[47] J. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.
[48] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022.
[49] G. Xia et al., "DOTA: A. large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
[50] D. Du et al., "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 213–226.
[51] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 370–386.
[52] W. Zhou, D. Du, L. Zhang, T. Luo, and Y. Wu, "Multi-granularity alignment domain adaptation for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9581–9590.
[53] C. Li et al., "Spatial attention pyramid network for unsupervised domain adaptation," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 481–497.
[54] C. Wang, H. Liao, Y. Wu, P. Chen, J. Hsieh, and I. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 390–391.
[55] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[56] Y. Wu, A. Kirillov, F. Massa, W. Lo, and R. Girshick, "Detectron2," 2019. [Online]. Available: https://github.com/facebookresearch/detectron2
[57] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

Debojyoti Biswas (Student Member, IEEE) received the B.Sc. degree in informaiton and communication engineering from the Noakhali Science and Technology University, Noakhali, Bangladesh, in 2018. He is currently working toward the Ph.D. degree in computer science with the Department of Computer Science, Texas State University, San Marcos, TX, USA.

He was a Lecturer with the Department of Computer Science, Leading University, Sylhet, Bangladesh, from 2019 to 2021. His research interests include computer vision, image processing, deep learning, and remote sensing image object detection.

Jelena Tešić (Member, IEEE) received the Dipl.Ing. degree in electrical engineering from the University of Belgrade, Belgrade, Serbia, in 1998, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara, CA, USA, in 1999 and 2004, respectively.

She is currently an Assistant Professor with Texas State University, San Marcos, TX, USA. Before that, she was a Research Scientist with Mayachitra Inc., Goleta, CA, USA, and IBM Watson Research Center, Yorktown Heights, NY, USA. She has authored more than 40 peer-reviewed scientific papers and holds six U.S. patents. Her research interests include advancing the analytic application of earth observation remote sensing, namely, object localization and identification at scale.

Dr. Tešić was an Area Chair for *ACM Multimedia*, IEEE International Conference on Image Processing, and IEEE International Conference on Multimedia and Expo. She was a Guest Editor for *IEEE Multimedia Magazine* for the September 2008 issue and a Reviewer for numerous IEEE and ACM journals.