Unsupervised Domain Adaptation with Debiased Contrastive Learning and Support-Set Guided Pseudo Labeling 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) DNN object detection performance. Unsupervised and Semi-supervised domain adaptation (DA) have been decent solutions to bridge the gap between two different distributions of datasets. The state-of-the-art pseudo-labeling 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 misleads the complete learning process. This paper proposes DCLDA (support-guided debiased contrastive learning for domain adaptation) to properly label the unlabeled target dataset and remove the bias toward target detection. We introduce (i) A support-set curated approach to generate high-quality pseudo-labels from the target dataset proposals, (ii) a reduced distribution gap across different datasets using domain alignment on local, global, and instance-aware features for remote sensing datasets, and (iii) 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 false negatives 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 mAP for DIOR, DOTA, Visdrone, and UAVDT datasets, respectively.**

***Index Terms*—Object Detection, Unsupervised Domain Adaptation, Debiased Contrastive Learning, UAV Images, Remote Sensing Analytics**

I. Introduction

Remote sensing images (RSI) 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.

The Department of Computer Science, Texas State University, San Marcos TX 78666 US; e-mail:debojyoti_biswas@txstate.edu, jtesic@txstate.edu.
This work is partially supported by the NAVAIR SBIR N68335-18-C-0199 and NVIDIA. 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 US Government paper references.

Fig. 1: Visual difference between consumer [5] and remote sensing images [9].

Examples of low variability frames in consumer data and high variability in overhead structures of similar pixel size are illustrated in Figure 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. Figure 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 vs. 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. Figure 2 (b) (left) contains well-defined objects, while Figure 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. Figure 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 Non-max 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 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 an-

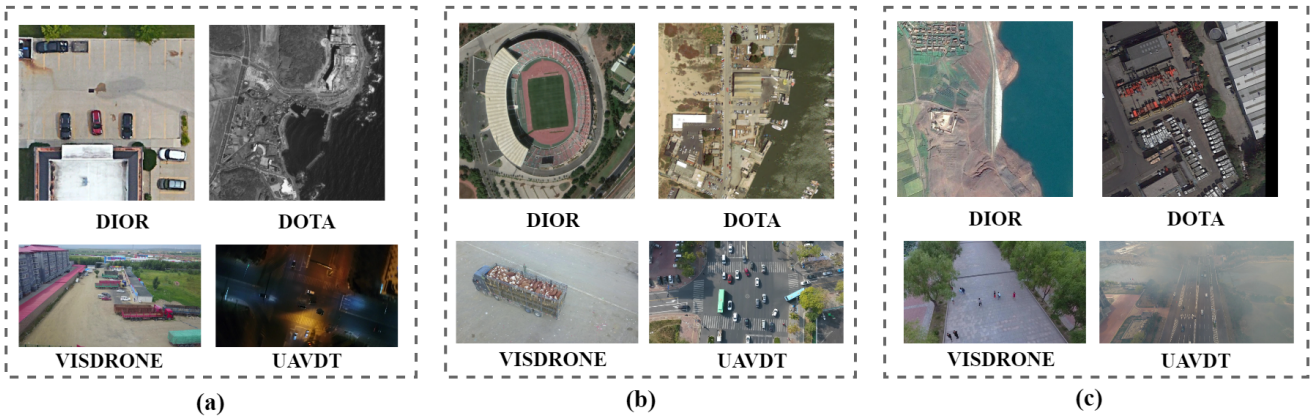


Fig. 2: High-variability remote sensing frames: (a) lighting conditions variations, (b) variations in object shape and scale, and (c) high variability due to geographical and weather changes.

other 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 unsupervised domain adaptation 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 (MMD) [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 task [12], [19], [31]. To our knowledge, few object detection benchmarks exist for remote-sensing images. The dataset's highly dense and variable nature hinders the progress of pseudo-labeling and optimal object detection performance of the RS Images. Xiong et al. 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 [22]. On the other hand, Yan et al. introduce a semantics-guided contrastive network to transfer semantic information for classes that have not been previously encountered [7]. Chen et al. presented a cross-domain adaptation object detection network that is rotation-invariant and relation-aware [10]. 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 this [10] work. Most unsupervised domain adaptation techniques require labeling the target datasets for instance-level domain adaptation and feature alignment. The existing pseudo-labeling 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 (HSIC) tasks. These works [8], [13] use Graph-based semi-supervised 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 hyperspectral images. 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 vs. all for the target dataset pseudo-labeling. 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. Secondly, the current contrastive learning approach follows the INFONCE [20] loss function with a single positive instance. Two problems are involved with this technique: (i) the INFONCE loss itself does not restrict the false negative image being selected as the negative case. For example, while performing local and global domain adaptation, the negative cases are selected randomly, and an image similar to the query image (See Figure 5) may be selected as a negative case. (ii) 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. So, 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 [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. Novel framework to address the high variability of remote sensing images for the object detection and labeling task in previously unseen datasets.

2. Efficient pseudo-labeling 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 improves the target domain detection performance.

3. Debiased contrastive learning for imbalanced remote sensing data, which is very important to produce domain-invariant, 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 *False Negative* examples that can disturb the learning process and result in poor performance.

4. Positive multi-sampling of *N-variants* positive samples in domain adaptation [17].

The rest of this article is organized as follows. Section II summarizes related work, and Section III introduces the proposed DCLDA method describing the debiased contrastive learning 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 summarizes the quantitative findings and outlines future works.

II. Related Work

Full potential use of Deep Neural Networks 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 [27], [28], [25] 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.

Object Detection in Remote Sensing Images: Shi et al. propose an anchor-free-based detector called Centerness-Aware Network (CANet), which captures the symmetrical shape of objects in remote sensing videos [29]. Biswas and Tešić suggest a strong custom backbone and an image difficulty scoring technique [30] to help detect small and complex objects. Xin et al. [32] use the local and global contrast information to effectively detect small bright and dark objects from Infrared images. Authors embed a small-sized U-Net into a larger U-Net backbone, which allows the multi-level and multi-scale representation learning of objects. Zhang et al. find that context-based feature extraction is more effective for detecting complex objects and scenes in the overhead imagery [33]. Global Context-Weaving Network incorporates a global context aggregation module and feature refinement module [34], and transformer-based CNN 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. modify darknet53 backbone with Cross Stage Partial DenseNet and add a transformer head in the detection layer, which gains state-of-the-art results of overhead drone images [25]. 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.

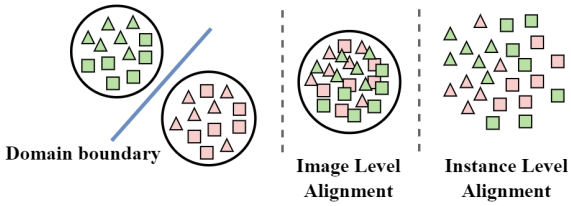


Fig. 3: Contrastive Learning alignments: different colors represent different domains, and shapes represent different categories.

Unsupervised Domain Adaptation: Training data for RS images can differ significantly from the source domain to the target domain regarding geographical, illumination, and visual characteristics. Besides RGB images, Hyperspectral remote sensing images 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, unsupervised domain adaptation methods generalize the model by aligning source and target [36]. Cheng adjusts the decision boundary biased towards the target data source domain and adds adversarial training in conjunction with image-to-image translation techniques [37]. Xiong et al. 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 [22].

On the other hand, Mattolin et al. implement the confidence-based mixing (ConfMix) of source and target domain images, where the confidence of an instance proposal is calculated based on the objectness score and the bounding box uncertainty score of each instance proposal from the image [38]. 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 (GSC) 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 remote sensing images.

Contrastive Learning for Domain Adaptation: It is hard to discriminate object classes in high-variable remote sensing images. 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.

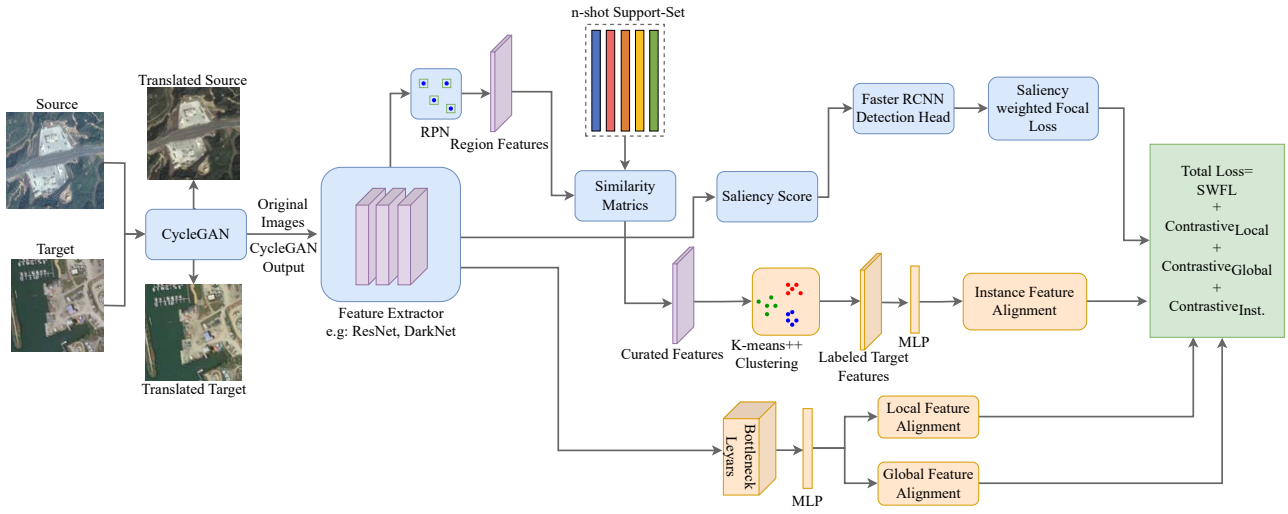


Fig. 4: Unsupervised Domain Adaptation Architecture with Debiased Contrastive Learning(DCLDA).

propose a probabilistic model to analyze the influence of the negative sampling ratio on training sample informativeness [40]. Yan et al. propose a semantics-guided contrastive network to transfer semantic information for classes not seen before [7]. Bai et al. propose a strategy called RefosNet to a representation focus shift network (RefosNet), which adds the rotation transformations to CL methods to improve the robustness of representation [41]. Li et al. use contrastive learning on overhead imagery for the semantic segmentation task [42]. Biswas et al. perform contrastive learning for object detection on the image-level feature alignment [43]. However, these works do not address the noise introduced in the pseudo-labeling process. Also, the mentioned contrastive learning approaches are unsuited for highly imbalanced datasets where debiasing is required to reduce false negative samples.

III. Methodology

The baseline detection architecture is built on [43], as illustrated in Figure 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.

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 domain adaptation 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 the instance level discriminability, as illustrated in Figure 3(middle). Hence, our goal is simultaneously aligning the image and instance level, as shown in Figure 3(right).

A. Unsupervised Domain Adaptation

In this paper, we perform unsupervised domain adaptation 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 Figure 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 domain adaptation with contrastive learning is performed bi-directional 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, 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 domain adaptation in the baseline paper [43]. However, we take it further to instance-level transformation with pseudo-labeling in the target dataset.

B. Support-Set Guided Pseudo Labeling

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 pseudo labeling, the

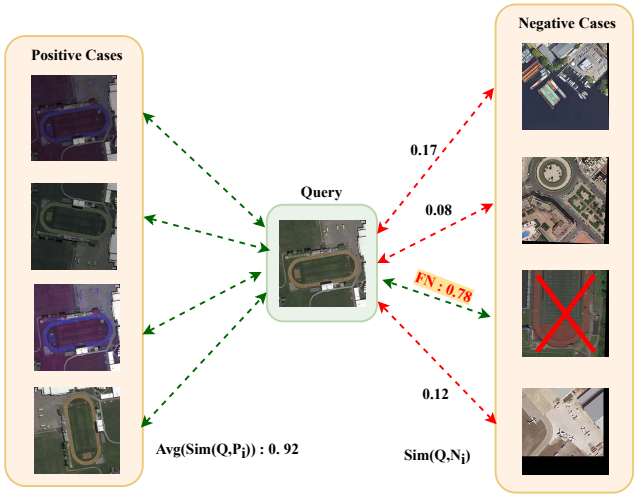


Fig. 5: Contrastive learning with multiple positive cases and false negative filtering. Here, green connections denote higher similarity, and red connections denote lower similarity with the query case.

target domain instance feature vectors in a mini-batch are collected from the RPN module (see Figure 4).

Early-stage target feature vectors are prone to background noise and mistake many background scenes as foreground objects. So, 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 a *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 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 pseudo-labeling task. The curated features are then used for target pseudo-labeling 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 [44]. Thus, we use K-means++ to generate pseudo labels through clustering from deep features. The clustering performance of the K-means++, as shown in Figure 6 and the value of K for clustering, is selected empirically. The selection process of K is described later in Subsection IV-F and Table IX.

C. Debiased Contrastive Learning

Contrastive learning is a process of matching different distributions based on query (Q) and key (K) embeddings [45], [47]. 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 Figure 3. The most used formula for contrastive learning is outlined in Equation 1, where τ is a hyper-parameter known as temperature to put penalties on the calculated similarities [46], [48].

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

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

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

In Equation 2, 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 $Q(1, size) \times K^+(M, size)' = Sim(1, M)$, 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 I 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 False Negative (FN) picked as the negative samples, as illustrated in Figure 5. Earlier domain adaptation methods for consumer datasets does not deal with this problem because consumer datasets are usually nearly balanced. On the other hand, RS datasets are often dominated by some major classes that require extra effort to gain optimal results. The number of false negatives (FNs) increases as we increase the number of negative samples in a mini-batch.

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

In this light, we propose to filter out negative samples with high similarity scores with the query sample. In Figure 5, three out of four images have a similarity score below 0.2 and one image is highly similar to the query image. **D**ebiased **C**ontrastive **L**earning (DCL) in Equation 3 summarizes the

process. First, reject the false negative 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 below formula,

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

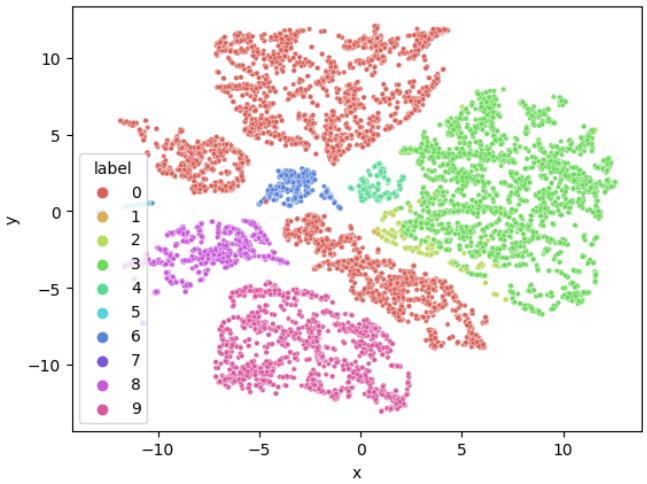


Fig. 6: Clustering visualization for pseudo labeling in 12,000 features over 10 classes of DOTA dataset.

Class Name	# of Ins. DIOR	# of Ins. DOTA	# of Ins. Visdrone	# of Ins. 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	-	-

TABLE I: Instance distribution statistics (Test Set) of the DIOR [9], DOTA2.0 [49], [50], Visdrone [51], and UAVDT [52] datasets over different categories.

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 domain. From the architecture of our proposed model in Figure 4, we can see that the first step toward local domain adaptation is to generate synthesized images from both source (S) and target (T) images in a mini-batch. For that, we use CycleGAN and pass both source and target image to generate translated source (S') and translated target (T'), respectively. Then, pass S, T', T, 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 Multi-Layers-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.

Lets 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 bi-directional adaptation, for the adaptation of the 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 bi-directional local contrastive loss between (S and T') and (T and S') can be calculated from the Equation 4 and 5.

$$DCL_{local}^{S,T'} = -log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} exp(sim(\alpha_i^S, \alpha_m^{T'})/\tau)}{D(\sum_{j=1}^{\nu} exp(sim(\alpha_i^S, \alpha_j^{T'})/\tau))} - log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} exp(sim(\alpha_i^{T'}, \alpha_m^S)/\tau)}{D(\sum_{j=1}^{\nu} exp(sim(\alpha_i^{T'}, \alpha_j^S)/\tau))}, j \neq i \quad (4)$$

$$DCL_{local}^{T,S'} = -log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} exp(sim(\alpha_i^T, \alpha_m^{S'})/\tau)}{D(\sum_{j=1}^{\nu} exp(sim(\alpha_i^T, \alpha_j^{S'})/\tau))} - log \frac{\frac{1}{\mu}\sum_{m=1}^{\mu} exp(sim(\alpha_i^{S'}, \alpha_m^T)/\tau)}{D(\sum_{j=1}^{\nu} exp(sim(\alpha_i^{S'}, \alpha_j^T)/\tau))}, j \neq i \quad (5)$$

In the above Equation 4 and 5, D stands for *Debiased*, m denotes the m^{th} 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 domain adaptation loss can be formulated by accumulating the loss for all query images in a mini-batch, as follows:

$$DCL_{local} = W_1 * DInfoNCE_{local}^{S,T'} + W_1 * DInfoNCE_{local}^{T,S'} \quad (6)$$

E. Debiased Global Contrastive Learning

Global domain adaptation 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 these $256 \times 25 \times 25$ 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 the 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. So, the bi-directional global contrastive loss between (S and T') and (T and S') can be presented as in Equation 7 and 8.

$$DCL^{S,T'}_{global} = -log\frac{\frac{1}{\mu}\sum_{m=1}^{\mu} exp(sim(\beta^S_i, \beta^{T'}_m)/\tau)}{D(\sum_{j=1}^{\nu} exp(sim(\beta^S_i, \beta^{T'}_j)/\tau))} -log\frac{\frac{1}{\mu}\sum_{m=1}^{\mu} exp(sim(\beta^{T'}_i, \beta^S_m)/\tau)}{D(\sum_{j=1}^{\nu} exp(sim(\beta^{T'}_i, \beta^S_j)/\tau))}, j \neq i \quad (7)$$

$$DCL^{T,S'}_{global} = -log\frac{\frac{1}{\mu}\sum_{m=1}^{\mu} exp(sim(\beta^T_i, \beta^{S'}_m)/\tau)}{D(\sum_{j=1}^{\nu} exp(sim(\beta^T_i, \beta^{S'}_j)/\tau))} -log\frac{\frac{1}{\mu}\sum_{m=1}^{\mu} exp(sim(\beta^{S'}_i, \beta^T_m)/\tau)}{D(\sum_{j=1}^{\nu} exp(sim(\beta^{S'}_i, \beta^T_j)/\tau))}, j \neq i \quad (8)$$

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

$$DCL_{global} = W_2 * DCL^{S,T'}_{global} + W_2 * DCL^{T,S'}_{global} \quad (9)$$

F. Debiased Instance Contrastive Learning

Local-Global (LG) contrastive learning helps to create domain invariant features as shown in Figure 3; 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 the different class instances in the feature space. To solve this issue, we propose to perform debiased instance contrastive learning for the source and target dataset and achieve class discrepancy in features. The effect of this learning is illustrated in Figure 3, 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 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^S_i and for the target region proposals is Γ^T_i. The corresponding class set for the source is C^S_i, and for the target is C^T_i; i is the proposal index among P proposals. For instance-level contrastive learning, the formula can be formulated from Equation 10 and 11.

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

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

Equation 10 and 11 represent the source and target instance loss, respectively. Here, μ and ν stand for the number of positive and negative samples, respectively, and i stands for $i^{th} \in theP$ 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 region proposals in a mini-batch, as follows:

$$DCL_{Ins} = W_3 * DCL^S_{Ins} + W_3 * DCL^T_{Ins} \quad (12)$$

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 is generally less reliable due to the large domain gap. In this light, we use weights W_1, W_2, and W_3 in Equation 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 local and global domain adaptation works well with an initial weight of 0.1, so we keep $W1$ and $W2 = 0.1$. For the instance domain adaptation, we tried different values of $W3$ as presented in Table VII. 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 loss components outlined in Equation 13.

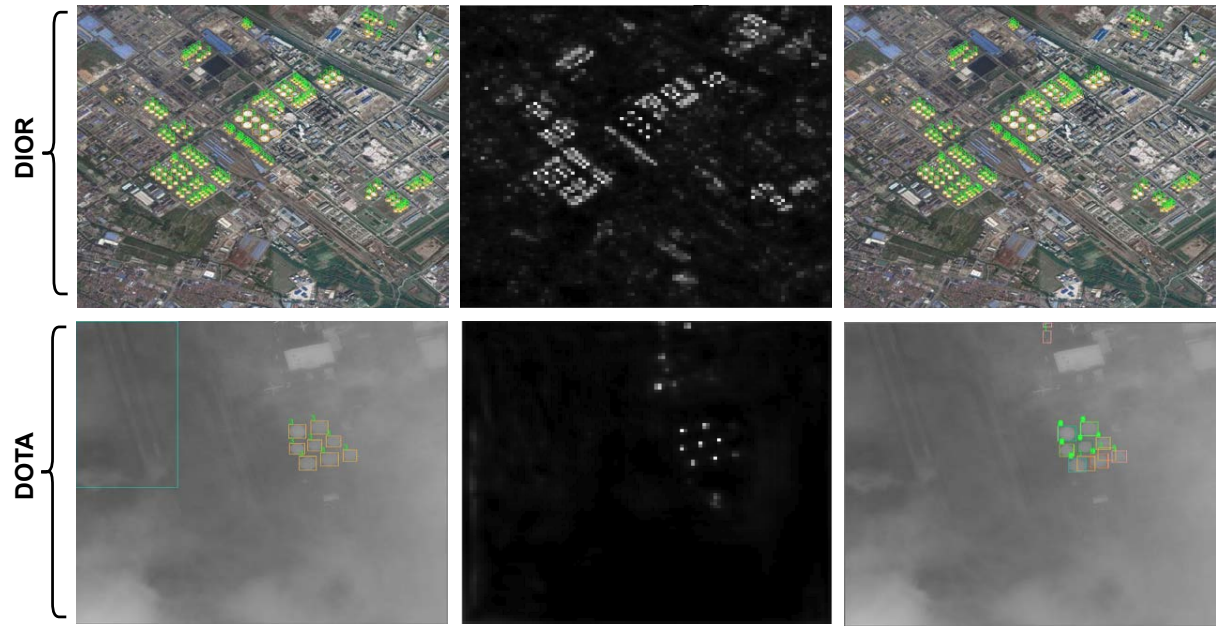
$$TotalLoss = SWFL(x, p_t, y) + DCL_{local} + DCL_{global} + DCL_{Ins} \quad (13)$$

IV. EXPERIMENTS

This section evaluates our proposed debiased contrastive learning model against current state-of-the-art domain adaptation methods on four remote-sensing image 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

Implementation We use the object classification pipeline similar to [43]: Darknet53 as the backbone as it is shown to preserve semantic information from the small objects than the residual-based feature extractor networks [27], [55]; RPN heatmap-based approach to identify dense small objects and remove NMS; and the detection block is Faster-RCNN [56]. 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** [57] and some part of **SOD** [30] implementations. With debiased contrastive learning, we implemented three new DA modules for local, global, and instance domain adaptation. Also, we implemented a Cythonized K-means++ that is much faster than the Python implementation, and the clustering time is recorded in Table IX.



(a) Ground Truth **(b) Saliency Map** **(c) Prediction**

Fig. 7: Detection results from DIOR (Source) and DOTA (Target) dataset using our DCLDA method.

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 [53]	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 [54]	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 [53]	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**

TABLE II: Classwise performance comparisons (mAP) for DIOR → DOTA benchmark(IOU=0.5), as measured both on the DIOR (source) and on the DOTA (target) datasets.

B. Hyper-parameter Settings

In CycleGAN network [58], load 800 and crop 640 were used for the data augmentation. To train our DCLDA model, we have resized all images to 800×800 pixels and set eight as the mini-batch size in each epoch. So, in total, we send $8 \times 4 = 32$ images in a mini-batch to train the DCLDA model. 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 pseudo labeling, we chose five samples (n) per class and created the 5-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 hyper-parameters 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, 11^{th} generation Intel® CoreTM i9-11900K @ 3.50GHz × 16 CPU, and 167GB of system memory to carry out all experiments.

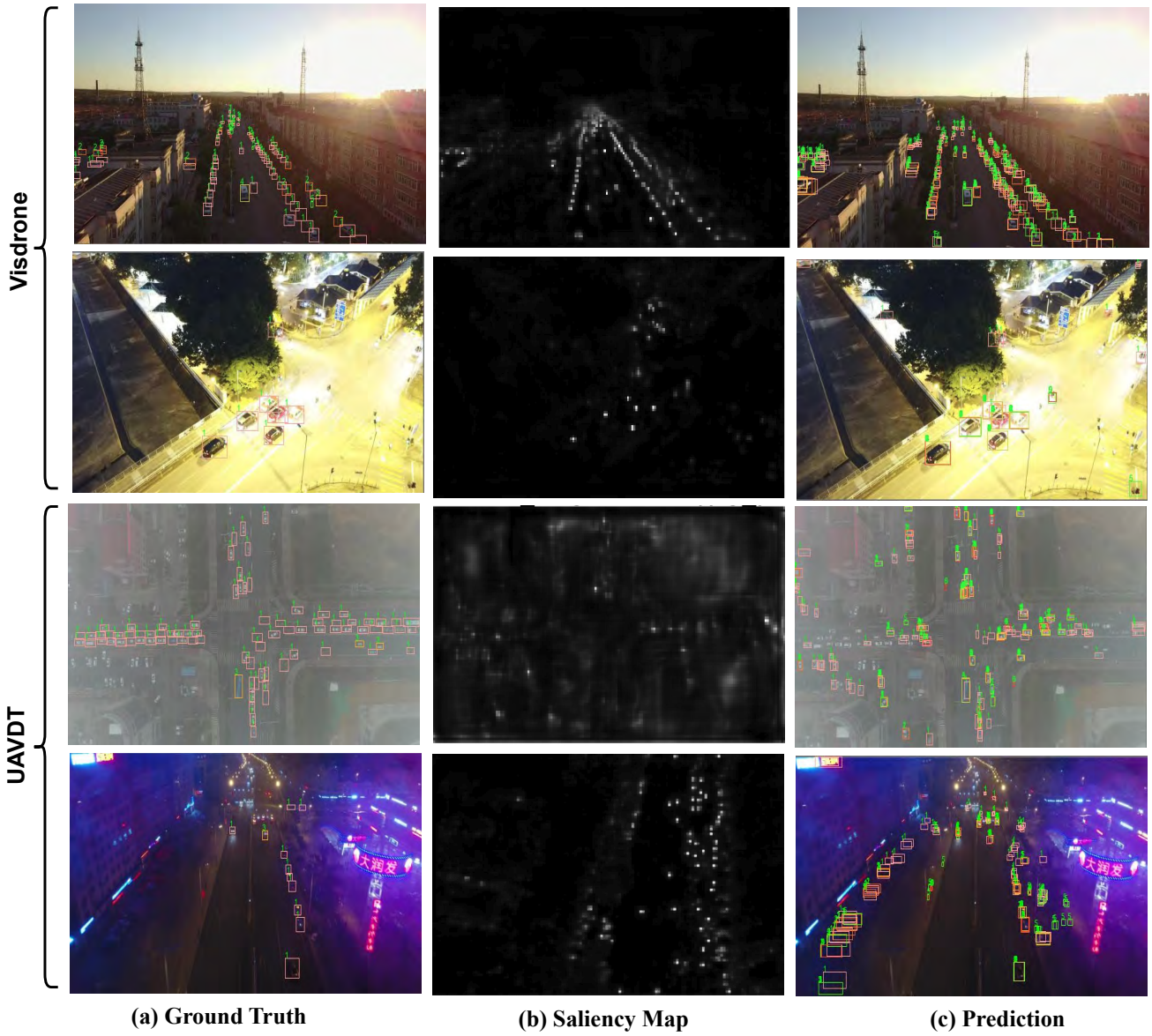


Fig. 8: Detection results from Visdrone (Source) and UAVDT (Target) dataset using our DCLDA method.

C. Datasets and Evaluation Metrics

DIOR data set 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. **DOTA** dataset comprises 2,430 overhead images with image sizes ranging from 800×800 to $29,200 \times 27,620$ pixels. The ground sample distance (GSD) in the data set 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 3,488 images. **Visdrone** is a UAV dataset containing over 10,000 image frames from more than 6 hours 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 set contains 6883 and 546 images, respectively. **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 domain adaptation: **source** $\rightarrow$ **target**.

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

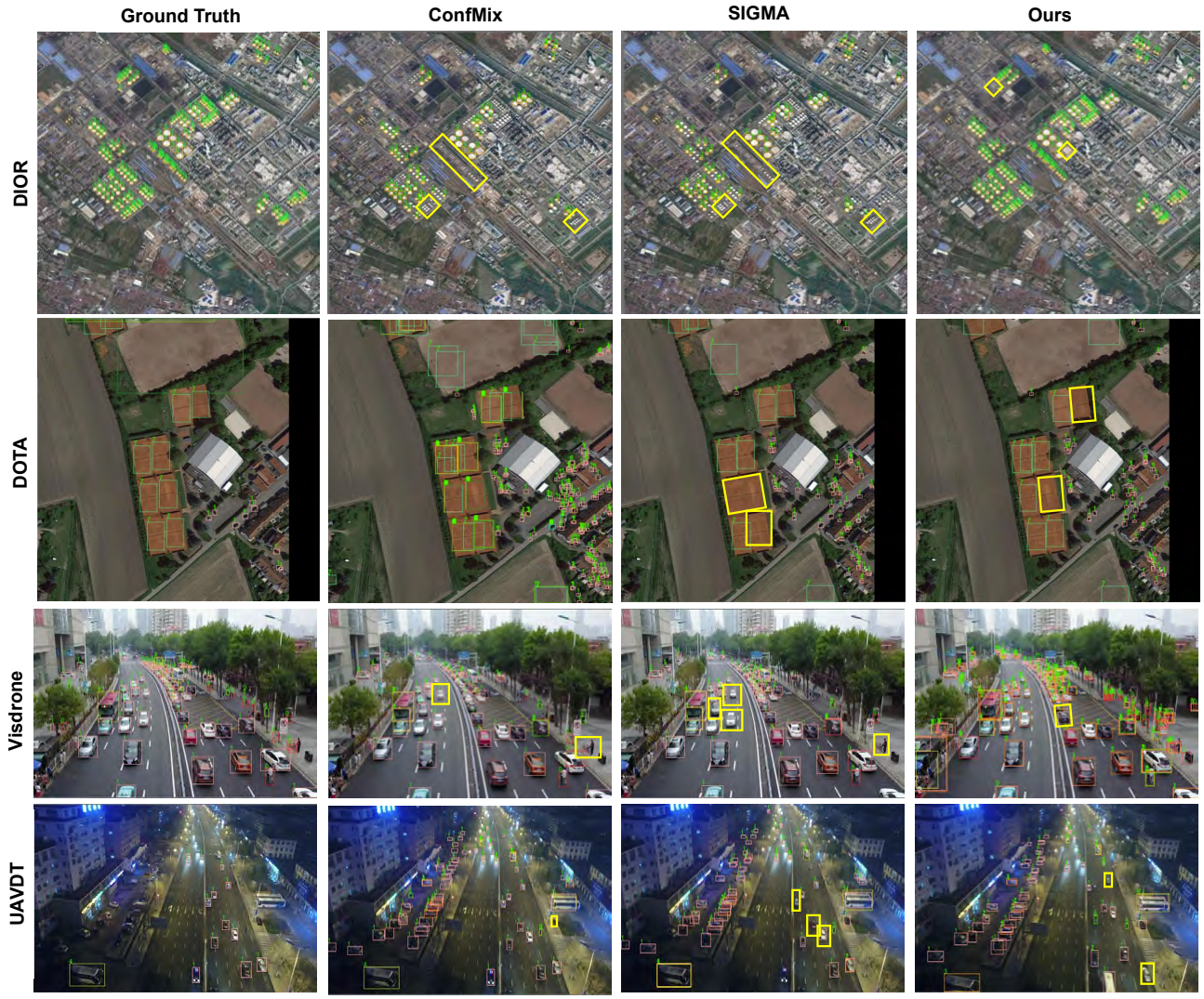


Fig. 9: Comparison of SOTA methods with our DCLDA method. In this figure, we present the comparison across different methods and datasets to illustrate the effectiveness of our model. The green boxes denote true predictions, and the yellow boxes denote missed detections.

Method	Car	Truck	Bus	VISDRONE → UAVDT
Baseline [26]	34.2	7.6	29.3	48.1	26.4
MGADA [53]	42.0	15.6	36.4	51.9	31.3
SAPNET [54]	31.5	7.9	22.7	26.3	20.8
MGADA [53]	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**

TABLE III: Classwise performance comparisons (mAP) for VISDRONE → UAVDT benchmark(IOU=0.5).

object categories. The mAP for all experiments was calculated with an IOU of 0.5 at the Non-Maximal suppression stage.

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

TABLE IV: Source and target detection performance (mAP) with(w/) and without(w/o) Instance Domain Adaptation (IDA).

D. Method Performance Comparisons

We compare our DCLDA method with several current state-of-the-art 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

		DIOR → DOTA		Visdrone → UAVDT	
# neg	# pos	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**

TABLE V: Quantitative performance comparisons (mAP) from **DCLDA** model for various negative and positive case values.

Method	# of Params.	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 VI: Multi-factor computational cost comparison between our proposed DCLDA and recent SOTA methods.

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 [53] and SAPNet [54], a spatial attention-based domain adaptation network for the performance measurements. A novel SemantIc-complete Graph MAtching (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 state-of-the-art comparisons. Figure 7 present 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 ground truth, foreground-focused saliency map, and object detection for different samples. It is evident from Figure 7 that our DCLDA model well-adopted the variation in lighting conditions, object sizes, and foggy-weather conditions between source and target datasets.Table II presents the quantitative performance comparison for DIOR and DOTA satellite images datasets. This table shows classwise performance for the target dataset and overall performance for both source and target datasets. We can see from Table II 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 local and global domain adaptation and pseudo-labeling-based instance adaptation, which helps us to outperform other state-of-the-art 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 VIII.

The Visdrone and UAVDT video datasets are two high-variability videos captured from UAV in Table III. The qualitative analysis of Visdrone and UAVDT datasets are presented in Figure 8, respectively. Visdrone and UAVDT pose critical domain gaps due to illumination, low light, and foggy conditions. Figure 8 shows samples with shadows due to high buildings and sunlight angles. Additionally, 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 Figure 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% mean average precision (mAP) (see Table III) on the source dataset, which is noteworthy. Our baseline method trained on only source data gives 26.4% of mAP, whereas our DCLDA method achieves 41.5% of mAP using debiased contrastive learning and pseudo labeling. Also, we have a +2.1% gain margin compared to the best state-of-the-art *ConfMix* method. Moreover, using debiased contrastive learning, we could shrink the performance gap between the oracle and our model from 30.5% to 15.4% compared to the baseline model. Table III. Table III also demonstrates that a well-designed backbone can enhance performance by around +2.7% on the video target domain with dominated dense objects.

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**

TABLE VII: Ablation study for different modules of our DCLDA method. Here, CGAN= CycleGAN Transfer Learning, LDA= Local domain adaptation, GDA= Global domain adaptation, and IDA= Instance-level domain adaptation.

In Figure 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 Figure 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 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 domain properly.

Finally, each dataset's precision, recall, and F1 scores are presented in Table VIII. 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 Figure 7 - 8, which shows the foreground detection results from experimental datasets.

E. Computational Cost Comparisons

The domain adaptation methods are well-known for their high computational cost (see Table VI). However, carefully designing the gradient computation tree helps our DCLDA method maintain reasonable computational stability with optimal detection performance. Table VI presents a computational comparison between the 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 2-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 sub-sample 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 hours and 13.4 hours, respectively, for 50 epochs. So, 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 IV shows that the instance domain adaptation 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 IV). 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 V illustrates the performance gain, and even for two positive samples, improves the overall performance by roughly 2.0% for both target datasets.

More positive and negative examples can introduce more noise and ultimately hamper the results, as illustrated in Table V for 15 negative and eight positive cases. The study found 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 pseudo labeling?* and Table IX shows that pseudo labeling with five clusters for DOTA and two for UAVDT can achieve up to 7.5% and 5.8% increase, respectively. Table I 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 VII 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 VII. 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 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 target datasets was recorded by carefully selecting all hyperparameters and setting W3 equal to 0.01 or 1%.

V. Conclusion

This paper proposes specialized contrastive learning with Support-Set guided pseudo-labeling for the unsupervised domain adaptation task. We show that remote-sensing video frames and images have significant domain shifts due to lighting conditions, weather changes, and geographical variance. Careful design of the detection pipeline and instance-aware domain adaptation method is required for optimal performance. Our proposed contrastive learning method consists of two significant improvements. The first is the debiased contrastive learning to remove false negative 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 pseudo-labeling 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 state-of-the-art methods. For the UAVDT and DOTA target dataset, we outperformed the latest state-of-the-art *ConfMix* method by +2.1% and +3.2% mAP, respectively. We hope our work can inspire future exploration of domain adaptation tasks in RS imagery using

	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

TABLE VIII: Comparison of Precision, Recall, and F1 score between the closest SOTA competitor and our proposed model for the experimental datasets.

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

TABLE IX: Target detection performance(mAP) with/without Aggregated Pseudo Labeling, clustering time, and the number of clusters. The clustering time is given for a mini-batch of 4000 features from both target datasets.

debiased contrastive learning. 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 domain adaptation pipeline for RSI imagery.

ACKNOWLEDGMENTS

The NVIDIA RTX 6000 GPU used for this research was donated by NVIDIA Corporation. NAVAIR SBIR N68335-18-C-0199 partially supports this work. 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 US government.

REFERENCES

[1] Kumar, M., Singh, P. & Singh, P. Machine learning and GIS-RS-based algorithms for mapping the groundwater potentiality in the Bundelkhand region, India. *Ecological Informatics*. pp. 101980 (2023)

[2] Valente, J., Sari, B., Kooistra, L., Kramer, H. & Mücher, S. Automated crop plant counting from very high-resolution aerial imagery. *Precision Agriculture*. **21**, 1366-1384 (2020)

[3] Workman, S. & Jacobs, N. Dynamic traffic modeling from overhead imagery. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 12315-12324 (2020)

[4] Biswas, D., Rahman, M., Zong, Z. & Tešić, J. Improving the Energy Efficiency of Real-time DNN Object Detection via Compression, Transfer Learning, and Scale Prediction. *2022 IEEE International Conference On Networking, Architecture And Storage (NAS)*. pp. 1-8 (2022)

[5] BookingHunterTV New York City Walking Tour Part 1 - Midtown Manhattan. (Video file,2019,12), https://youtu.be/-IpXdtWfneI

[6] Arthur, D. & Vassilvitskii, S. k-means++: The advantages of careful seeding. (Stanford,2006)

[7] Yan, C., Chang, X., Luo, M., Liu, H., Zhang, X. & Zheng, Q. Semantics-guided contrastive network for zero-shot object detection. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. (2022)

[8] Ding, Y., Zhang, Z., Zhao, X., Cai, Y., Li, S., Deng, B. & Cai, W. Self-supervised locality preserving low-pass graph convolutional embedding for large-scale hyperspectral image clustering. *IEEE Transactions On Geoscience And Remote Sensing*. **60** pp. 1-16 (2022)

[9] Li, K., Wan, G., Cheng, G., Meng, L. & Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **159** pp. 296-307 (2020)

[10] Chen, Y., Liu, Q., Wang, T., Wang, B. & Meng, X. Rotation-Invariant and Relation-Aware Cross-Domain Adaptation Object Detection Network for Optical Remote Sensing Images. *Remote Sensing*. **13**, 4386 (2021)

[11] Lam, D., Kuzma, R., McGee, K., Dooley, S., Laielli, M., Klaric, M., Bulatov, Y. & McCord, B. xview: Objects in context in overhead imagery. *ArXiv Preprint ArXiv:1802.07856*. (2018)

[12] Chen, J., Zhu, J., Guo, Y., Sun, G., Zhang, Y. & Deng, M. Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention. *IEEE Transactions On Geoscience And Remote Sensing*. **60** pp. 1-15 (2022)

[13] Georgoulas, I., Protopapadakis, E., Makantasis, K., Seychell, D., Doulamis, A. & Doulamis, N. Graph-based Semi-supervised Learning with Tensor Embeddings for Hyperspectral Data Classification. *IEEE Access*. (2023)

[14] Zheng, J., Zhao, Y., Wu, W., Chen, M., Li, W. & Fu, H. Partial domain adaptation for scene classification from remote sensing imagery. *IEEE Transactions On Geoscience And Remote Sensing*. **61** pp. 1-17 (2022)

[15] Mahabubur Rahman, M. & Tešić, J. Hybrid Approximate Nearest Neighbor Indexing and Search (HANNIS) for Large Descriptor Databases. *2022 IEEE International Conference On Big Data (Big Data)*. pp. 3895-3902 (2022)

[16] Tolstikhin, I., Sriperumbudur, B. & Schölkopf, B. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances In Neural Information Processing Systems*. **29** (2016)

[17] Chuang, C., Robinson, J., Lin, Y., Torralba, A. & Jegelka, S. Debiased contrastive learning. *Advances In Neural Information Processing Systems*. **33** pp. 8765-8775 (2020)

[18] Ding, Y., Zhang, Z., Zhao, X., Cai, W., Yang, N., Hu, H., Huang, X., Cao, Y. & Cai, W. Unsupervised self-correlated learning smoothy enhanced locality preserving graph convolution embedding clustering for hyperspectral images. *IEEE Transactions On Geoscience And Remote Sensing*. **60** pp. 1-16 (2022)

[19] Luo, M. & Ji, S. Cross-spatiotemporal land-cover classification from VHR remote sensing images with deep learning based domain adaptation. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **191** pp. 105-128 (2022)

[20] Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *ArXiv Preprint ArXiv:1807.03748*. (2018)

[21] Hong, D., Yokoya, N., Chanussot, J. & Zhu, X. An augmented linear mixing model to address spectral variability for hyperspectral unmixing. *IEEE Transactions On Image Processing*. **28**, 1923-1938 (2018)

[22] Xiong, L., Ye, M., Zhang, D., Gan, Y. & Liu, Y. Source data-free domain adaptation for a faster R-CNN. *Pattern Recognition*. **124** pp. 108436 (2022)

[23] Jui, T., Bejarano, G. & Rivas, P. A machine learning-based segmentation approach for measuring similarity between sign languages. *Sign-lang@ LREC 2022*. pp. 94-101 (2022)

[24] Babavalian, M. & Kiani, K. Learning distribution of video captions using conditional GAN. *Multimedia Tools And Applications*. pp. 1-23 (2023)

[25] Zhu, X., Lyu, S., Wang, X. & Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. *Proceedings Of The IEEE/CVF International Conference On Computer Vision*. pp. 2778-2788 (2021)

[26] Zhou, X., Koltun, V. & Krähenbühl, P. Probabilistic two-stage detection. *ArXiv Preprint ArXiv:2103.07461*. (2021)

[27] Bochkovskiy, A., Wang, C. & Liao, H. Yolov4: Optimal speed and accuracy of object detection. *ArXiv Preprint ArXiv:2004.10934*. (2020)

[28] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. & Berg, A. Ssd: Single shot multibox detector. *European Conference On*

Computer Vision. pp. 21-37 (2016)

[29] Shi, L., Kuang, L., Xu, X., Pan, B. & Shi, Z. CANet: Centerness-aware network for object detection in remote sensing images. *IEEE Transactions On Geoscience And Remote Sensing*. **60** pp. 1-13 (2021)

[30] Biswas, D. & Tešić, J. Small Object Difficulty (SOD) Modeling for Objects Detection in Satellite Images. *2022 14th International Conference On Computational Intelligence And Communication Networks (CICN)*. pp. 125-130 (2022)

[31] Hong, D., Zhang, B., Li, H., Li, Y., Yao, J., Li, C., Werner, M., Chanussot, J., Zipf, A. & Zhu, X. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sensing Of Environment*. **299** pp. 113856 (2023)

[32] Wu, X., Hong, D. & Chanussot, J. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions On Image Processing*. **32** pp. 364-376 (2022)

[33] Zhang, J., Shi, Y., Zhang, Q., Cui, L., Chen, Y. & Yi, Y. Attention guided contextual feature fusion network for salient object detection. *Image And Vision Computing*. **117** pp. 104337 (2022)

[34] Wu, Y., Zhang, K., Wang, J., Wang, Y., Wang, Q. & Li, X. GCWNet: A Global Context-Weaving Network for Object Detection in Remote Sensing Images. *IEEE Transactions On Geoscience And Remote Sensing*. **60** pp. 1-12 (2022)

[35] Li, Q., Chen, Y. & Zeng, Y. Transformer with transfer CNN for remote-sensing-image object detection. *Remote Sensing*. **14**, 984 (2022)

[36] Deng, Z., Kong, Q., Akira, N. & Yoshinaga, T. Hierarchical contrastive adaptation for cross-domain object detection. *Machine Vision And Applications*. **33**, 1-13 (2022)

[37] Chen, C., Zheng, Z., Ding, X., Huang, Y. & Dou, Q. Harmonizing transferability and discriminability for adapting object detectors. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 8869-8878 (2020)

[38] Mattolin, G., Zanella, L., Ricci, E. & Wang, Y. ConfMix: Unsupervised Domain Adaptation for Object Detection via Confidence-based Mixing. *Proceedings Of The IEEE/CVF Winter Conference On Applications Of Computer Vision*. pp. 423-433 (2023)

[39] Li, W., Liu, X. & Yuan, Y. Sigma: Semantic-complete graph matching for domain adaptive object detection. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 5291-5300 (2022)

[40] Wu, C., Wu, F. & Huang, Y. Rethinking InfoNCE: How Many Negative Samples Do You Need?. *ArXiv Preprint ArXiv:2105.13003*. (2021)

[41] Bai, G., Xi, W., Hong, X., Liu, X., Yue, Y. & Zhao, S. Robust and Rotation-Equivariant Contrastive Learning. *IEEE Transactions On Neural Networks And Learning Systems*. (2023)

[42] Li, H., Li, Y., Zhang, G., Liu, R., Huang, H., Zhu, Q. & Tao, C. Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images. *IEEE Transactions On Geoscience And Remote Sensing*. **60** pp. 1-14 (2022)

[43] Biswas, D. & Tešić, J. Progressive Domain Adaptation with Contrastive Learning for Object Detection in the Satellite Imagery. (2023)

[44] Arthur, D. & Vassilvitskii, S. K-means++ the advantages of careful seeding. *Proceedings Of The Eighteenth Annual ACM-SIAM Symposium On Discrete Algorithms*. pp. 1027-1035 (2007)

[45] Hadsell, R., Chopra, S. & LeCun, Y. Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference On Computer Vision And Pattern Recognition (CVPR'06)*. **2** pp. 1735-1742 (2006)

[46] He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 9729-9738 (2020)

[47] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. *International Conference On Machine Learning*. pp. 1597-1607 (2020)

[48] Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M. & Others Bootstrap your own latent-a new approach to self-supervised learning. *Advances In Neural Information Processing Systems*. **33** pp. 21271-21284 (2020)

[49] Ding, J., Xue, N., Xia, G., Bai, X., Yang, W., Yang, M., Belongie, S., Luo, J., Datcu, M., Pelillo, M. & Zhang, L. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. pp. 1-1 (2021)

[50] Xia, G., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M. & Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. *The IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*. (2018,6)

[51] Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., Hu, Q., Peng, T., Zheng, J., Wang, X., Zhang, Y. & Others VisDrone-DET2019: The vision meets drone object detection in image challenge results. *Proceedings Of The IEEE/CVF International Conference On Computer Vision Workshops*. pp. 0-0 (2019)

[52] Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q. & Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. *Proceedings Of The European Conference On Computer Vision (ECCV)*. pp. 370-386 (2018)

[53] Zhou, W., Du, D., Zhang, L., Luo, T. & Wu, Y. Multi-granularity alignment domain adaptation for object detection. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 9581-9590 (2022)

[54] Li, C., Du, D., Zhang, L., Wen, L., Luo, T., Wu, Y. & Zhu, P. Spatial attention pyramid network for unsupervised domain adaptation. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. pp. 481-497 (2020)

[55] Wang, C., Liao, H., Wu, Y., Chen, P., Hsieh, J. & Yeh, I. CSPNet: A new backbone that can enhance learning capability of CNN. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition Workshops*. pp. 390-391 (2020)

[56] Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances In Neural Information Processing Systems*. **28** (2015)

[57] Wu, Y., Kirillov, A., Massa, F., Lo, W. & Girshick, R. Detectron2. (https://github.com/facebookresearch/detectron2,2019)

[58] Zhu, J., Park, T., Isola, P. & Efros, A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Computer Vision (ICCV), 2017 IEEE International Conference On*. (2017)

Debojyoti Biswas is a Computer Science Ph.D. student. He received his B.S.c degree from the Noakhali Science and Technology University, Bangladesh, in 2018. He worked as a Lecturer in the Computer Science department at Leading University, Bangladesh, from 2019 to 2021 before his Ph.D. studies. His research interests include computer vision, image processing, deep learning, and remote sensing image object detection.

Jelena Tešić, Ph.D. is an Assistant Professor at Texas State University. Before that, she was a research scientist at Mayachitra (CA) and IBM Watson Research Center (NY). She received her Ph.D. (2004) and M.Sc. (1999) in Electrical and Computer Engineering from the University of California Santa Barbara, CA, USA, and Dipl. Ing. (1998) in Electrical Engineering from the University of Belgrade, Serbia. Dr. Tešić served as Area Chair for ACM Multimedia 2019-present and IEEE ICIP and ICME conferences; she has served as Guest Editor for IEEE Multimedia Magazine for the September 2008 issue and as a reviewer for numerous IEEE and ACM Journals. She has authored over 40 peer-reviewed scientific papers and holds six US patents. Her research focuses on advancing the analytic application of EO remote sensing, namely object localization and identification at scale.