

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 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. However, the pseudo-labeling process involved in existing works is prone to background noise, hindering the optimal performance in target datasets. Also, 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 issue toward target detection. Firstly, a support-set curated approach is proposed to generate high-quality pseudo-labels from the target dataset proposals. Next, we attempt to reduce the distribution gap across different datasets by performing domain alignment on local, global, and instance-aware features for remote sensing datasets. Finally, we introduce the debiased contrastive loss function with multiple positive examples, learning to make the model more generalized of the variable appearance of a particular class over images and domains. Our proposed debiased contrastive learning pivots on class probabilities to address the challenge of false negatives in the unsupervised framework. We show the advantages of the proposed model on satellite (DIOR and DOTA2.0) and drone (Visdrone and UAVDT) image datasets.

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

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/or 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 paper references.



Fig. 1: Visual difference between consumer [5] and remote sensing images [6].

Examples of low variability frames in consumer data and high variability in overhead frames 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 area of capture or drone flight. 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 significantly.

2. **Variation in Object Size** is great 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 between two datasets.

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 [7]. Today, only a few aerial datasets exist that cover real scenario object class diversity and a sufficient number of training examples.

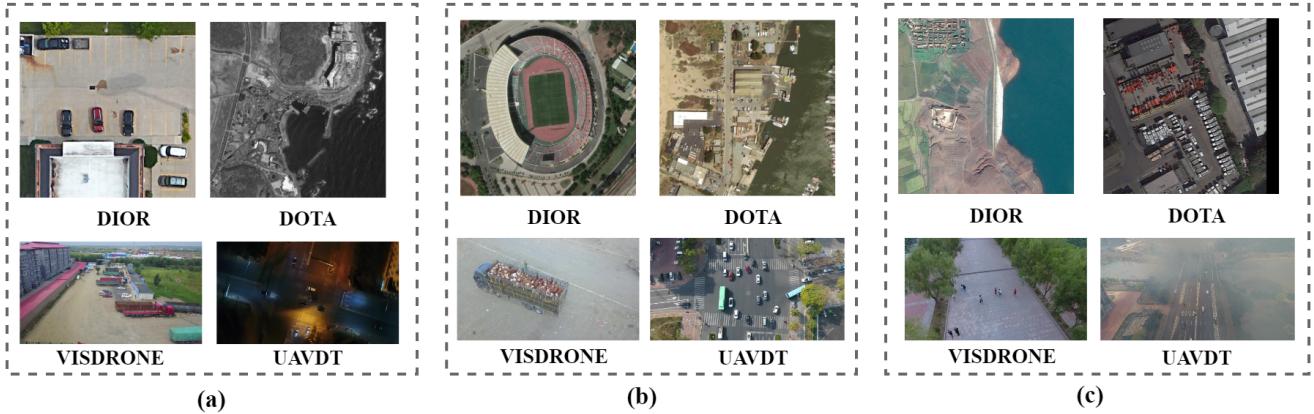


Fig. 2: High-variability remote sensing frames: (a) lighting conditions variations, (b) variations in object shape and scale, and (c) high variability due to geographical and weather changes.

A common technique to generalize a model is to train on one source dataset, and finetuning 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 unsupervised domain adaptation methods have been widely used in the classification and segmentation tasks of RSIs [8], [9]. These techniques mainly focus on mitigating the disparity by leveraging semantic feature alignment between the source and target domains. Later, Maximum mean discrepancy (MMD) [10] was utilized to preserve the main statistical properties across domains by minimizing the distance of the distribution between the source and target domains.

Although several domain adaption techniques were used over the years for the improvement of the classification and semantic segmentation task, to the best of our knowledge, there are very few object detection benchmarks for remote sensing images. The highly dense and variable nature of the dataset 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 [11]. On the other hand, Yan et al. introduce a semantics-guided contrastive network aimed at transferring semantic information for classes that have not been previously encountered [12]. Furthermore, Chen et al. [13] presented 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. Firstly, most UDA object detection techniques require labeling the target datasets. It is required for the instance-aware domain adaptation and feature alignment.

The existing pseudo-labeling techniques are mostly cluster-based, which does not address the possible background noise being considered as foreground objects. Secondly, the current contrastive learning approach follows the INFONCE [14] loss function with a single positive instance. There are two problems involved with this technique, (i) the INFONCE loss itself does not restrict the false negative image being selected as the negative cases. For example, while performing local and global domain adaptation, the negative cases are selected randomly, and it is highly possible an image similar to the query image (See Figure 5) is selected as a negative case. (ii) the default INFONCE loss works with only a single positive example, however, it is very important to consider positive samples with variable appearance for a particular class over images and domains. To relieve the limitations of the existing domain adaptation algorithms, we propose DCLDL with support-guided pseudo labeling that removes the noise attracted from the background object proposals. We use the N-shot approach to filter out unnecessary proposals from future representation learning. Next, to the best of our knowledge, we proposed debiased contrastive learning with more than one positive case for the imbalanced RSI datasets for the first time. The previous work on debiased contrastive learning [15] focuses only on balanced datasets. However, our experimental datasets are highly imbalanced; thus, this approach does not hold valid for our task. In a nutshell, we make the following contributions:

- 1) we propose a novel framework to address the high variability of remote sensing images for the object detection and labeling task in previously unseen datasets.
- 2) We propose an efficient pseudo-labeling process that depends on N-shot learning to remove the unwanted background noise from the target object proposals. Our experiments show that fine curation of target proposals significantly improves the target domain detection performance.
- 3) For the first time, we introduce debiased contrastive learning for imbalanced remote sensing datasets. We show that it 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) We employ the idea of N positive samples in domain adaptation as it proved very successful [15] in other representation tasks.

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 playing a crucial role in solving recent consumer applications [16], [17]. Recent advantages in the field show that the object detection task can be successfully solved for the Drone captured Visdrone dataset[18] and the COCO consumer image benchmark dataset [19]. 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 [20], [21], [22], [18] 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 [23]. Biswas and Tešić suggest a strong custom backbone and an image difficulty scoring technique [24] to help detect small and complex objects. Zhang et al. find that context-based feature extraction is more effective for detecting complex objects and scenes in the overhead imagery [25]. Global Context-Weaving Network incorporates a global context aggregation module and feature refinement module [26], and transformer-based CNN encoders are used for better feature extraction [27]. 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 [18]. 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 detectors to capture every small object from different levels of features.

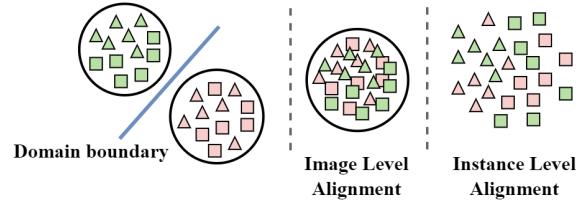


Fig. 3: Contrastive Learning alignments: different color represents different domains, and shapes represent different categories.

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. For a labeled source dataset and an unlabeled target dataset, unsupervised domain adaptation methods generalize the model by aligning source and target [28]. Cheng adjusts the decision boundary biased towards the target data source domain and adds adversarial training in conjunction with image-to-image translation techniques [29]. 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 [11]. Ma et al. [30] minimize domain discrepancy between source and target using a progressive domain mixup technique. On the other hand, Xu et al. have introduced a semantic-aware mixup (SAM) for domain generalization, where whether to perform a mixup depends on the semantic and domain information [31]. Mattolin et al. implement the confidence-based mixing 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 [32]. A novel SemantIc-complete Graph MAtching (SIGMA) [33] framework was proposed for the Domain Adaptation task, which completes mismatched semantics and reformulates the adaptation with graph matching. Mostly, 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. propose a probabilistic model to analyze the influence of the negative sampling ratio on training sample informativeness [34]. Yan et al. propose a semantics-guided contrastive network to transfer semantic information for classes not seen before [12]. 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 [35]. Li et al. use contrastive learning on overhead imagery for the semantic segmentation task [36]. Biswas et al. perform contrastive learning for object detection on the image-level feature alignment [37]. However,

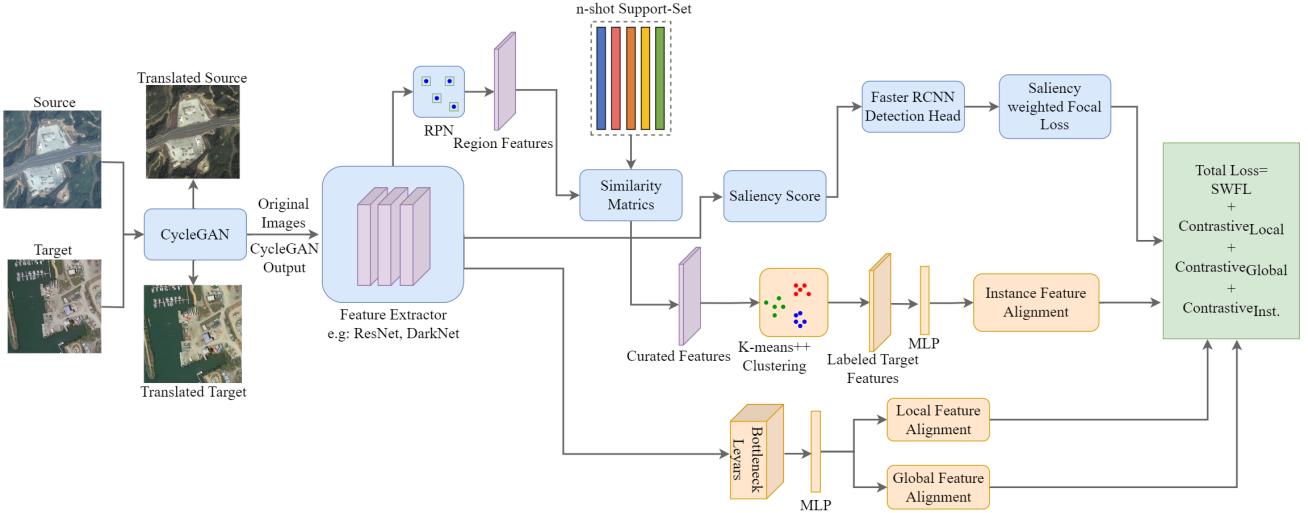


Fig. 4: Unsupervised Domain Adaptation Architecture with Debiased Contrastive Learning(DCLDA).

these works do not address the noise introduced in the pseudo-labeling process. Also, the mentioned contrastive learning approaches are not well suited for highly imbalanced datasets where debiasing is required to reduce false negative samples.

III. METHODOLOGY

The baseline detection architecture is built on [37], as illustrated in Figure 4. The baseline proved to play an efficient feature extraction strategy and the saliency-weighted custom focal loss function for remote-sensing images [37]. The saliency information from each image is used to calculate the difficulty score of each image as we track the number of neuron activations and the number of objects per image to derive the final saliency score. Based on this saliency/difficulty score, the Loss function assigns more penalties on difficult images and less on easy images [37].

Contrastive learning evaluates pair-to-pair relationships by measuring the similarities between different sample pairs, such as query-positive or query-negative. Here, Query is the subject feature, whereas positive samples are augmented features similar to the subject, and negative samples are randomly selected features dissimilar to the subject feature. Performing only image-level contrastive domain adaptation is a strong 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 to simultaneously perform the image and instance level alignment, as shown in Figure 3(right).

Previous works use traditional methods such as k-means or one vs. all for the target dataset pseudo-labeling. In this work, we use an advanced clustering technique *K-means++* [38] for generating target labels due to its proven performance [39] in high-dimensional data. Moreover, instead of using the single example as the positive sample, we propose to use N numbers of positive samples for contrastive learning. It was observed that using more than one positive case increases

the performance significantly. Finally, we perform progressive debiasing with our custom contrastive loss to remove the False Negatives (FNs) from the target negative samples.

A. Unsupervised Domain Adaptation

The idea of DA comes from the scarcity of available annotated datasets and the different factors that introduce domain gaps among the datasets. The Remote-sensing images captured from other parts of the world give different geographical variances and background difficulties as illustrated in Figure 2(c). Camera orientations, weather conditions, and changes in illumination challenges degrade the performance of an object detection model trained on a different dataset. Unsupervised domain adaptation minimizes the domain gap between two datasets, called source and target datasets. It is assumed that we have annotations for only the source dataset, and a large domain gap exists between them. The goal is to generate domain-invariant features at different levels of image features and perform better in unseen/target datasets.

In this paper, we perform unsupervised domain adaptation with a contrastive learning technique to align domains at local, global, and instance levels. 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 . We use the CycleGAN network to produce 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 represents a more abstract version of objects. The authors performed only local-global domain adaptation in the baseline paper [37]. However, we take it further to instance-level adaptation with pseudo-labeling in the target dataset.

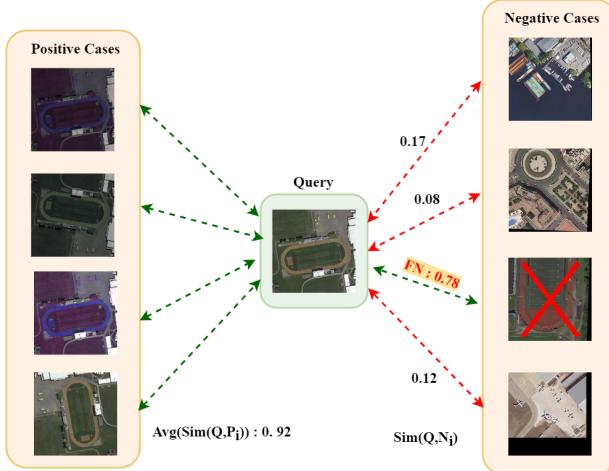


Fig. 5: Contrastive learning with multiple positive cases and false negative filtering. Here, green connections denote higher similarity, and red connections denote lower similarity with the query case.

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. As we do not have any GT for the target dataset, we must generate labels for the target proposals to guide contrastive learning. To perform pseudo labeling, the 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 [40]. 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-C and Table VI.

C. Debiased Contrastive Learning

Contrastive learning is a process of matching different distributions based on Query (Q) and Key (K) embeddings [41], [42]. 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 [43], [44].

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

The similarity can be calculated using cosine, Euclidian, or Wasserstein distance functions. The cosine similarity score is used in the experiments and calculated as $\text{sim}(x, y)$ for two features x , and y is $\text{sim}(x, y) = x^T / (\|x\| * \|y\|)$. 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 [37], 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 [15] 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:

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

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 $Q(1, \text{size}) \times K^+(M, \text{size})' = \text{Sim}(1, M)$, 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 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 dataset 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.

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

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. Debiased Contrastive Learning (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, D_K^- is calculated using below formula,

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

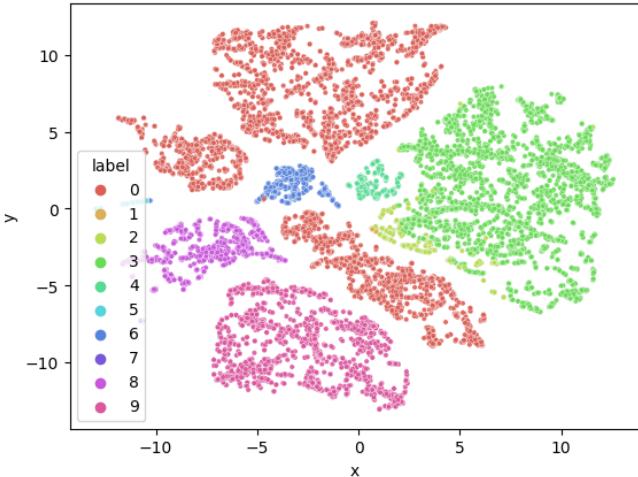


Fig. 6: Clustering visualization for pseudo labeling in 12,000 features over 10 classes of DOTA dataset.

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 features 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

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 [6], DOTA2.0 [45], [46], Visdrone [47], and UAVDT [48] datasets over different categories.

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 $\alpha_i^S, \alpha_i^{T'}, \alpha_i^T$, 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.

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

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

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 $\beta_i^S, \beta_i^{T'}, \beta_i^T$, 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.

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

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

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

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

F. Debiased Instance Contrastive Learning

Local-Glocal (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. However, it is also evident that 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.

Toward instance adaptation, 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 class-wise discriminability in features. Instances near decision boundaries may look very similar but belong to different classes.

Let's denote the source and target region proposals as Γ_i^S and Γ_i^T , respectively, and corresponding classes as C_i^S and C_i^T , where i is the proposal index among P proposals. For instance-level contrastive learning, the formula can be formulated from Equation 10 and 11.

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

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

The above 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 $i^{th} \in P$ 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 region proposals in a mini-batch, as follows:

$$DCL_{Ins} = W_3 * DCL_{Ins}^S + W_3 * DCL_{Ins}^T \quad (12)$$

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 adaptation, 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 W_1 and $W_2 = 0.1$. For the instance domain adaptation, we tried different values of W_3 as presented in Table ??, but 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

In this section, we evaluate 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-B, and the extensive ablation studies over different factors and parameters are outlined in Section IV-C.

Method	Detector+ Backbone	Bridge	Vehicle	Harbor	Storage	Baseball Track	B.Ball	Tennis	Stadium	Airport	DIOR → mAP	DOTA mAP
		Tank	Field	Field	Court	Court						
Baseline [19]	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
MGADA [49]	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
SAPNET [50]	FCOS ResNet	7.8	9.2	18.1	20.2	35.5	24.7	29.2	74.7	0.0	19.3	55.1
MGADA [49]	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
SIGMA [33]	FCOS ResNet50	27.0	32.6	64.5	65.0	55.4	56.6	62.3	91.9	1.3	34.7	77.2
ConfMix [32]	YOLOv5 Darknet53	27.2	32.0	65.9	65.1	56.3	56.3	61.5	93.5	1.0	34.9	78.8
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
DCLDA	CenNet2 Darknet53	30.1	28.8	70.0	65.8	55.4	52.5	62.2	93.2	7.9	37.3	82.7
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.

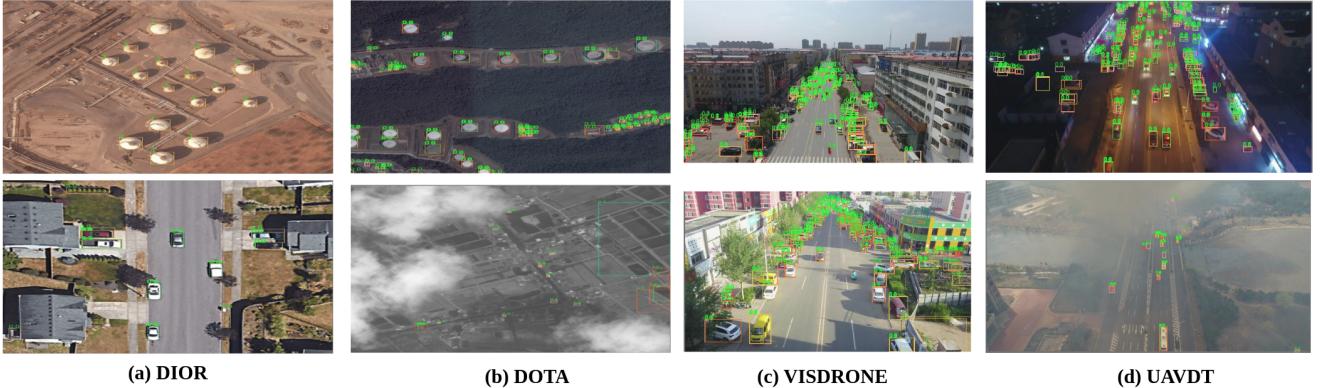


Fig. 7: Source and Target domain detection results using our DCLDA method.

A. Implementation Details and Setup

Implementation We use the object classification pipeline similar to [37]: Darknet53 as the backbone as it is shown to preserve semantic information from the small objects than the residual-based feature extractor networks [21], [51]; RPN heatmap-based approach to identify dense small objects and remove NMS; and the detection block is Faster-RCNN [52]. 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** [53] and some part of **SOD** [24] 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 VI.

In CycleGAN network [54], 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. We have used NVIDIA 2 x RTX 6000 GPU with 49GB of memory, 11th generation Intel® CoreTM i9-11900K @ 3.50GHz × 16 CPU, and 167GB of system memory to carry out all experiments.

Datasets **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 $29,200 \times 27,620$

Method	Car	Truck	Bus	VISDRONE → UAVDT
Baseline [19]	34.2	7.6	29.3	48.1
MGADA [49]	42.0	15.6	36.4	51.9
SAPNET [50]	31.5	7.9	22.7	26.3
MGADA [49]	39.2	12.6	35.8	54.6
SIGMA [33]	50.1	20.9	45.5	46.3
ConfMix [32]	51.3	20.4	46.0	46.5
DCLDA*	52.5	23.0	41.1	58.4
DCLDA	54.2	24.4	46.3	59.2
Oracle	72.4	37.8	60.5	45.6
				56.9

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

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 Average Precision (AP) by considering both precision and recall for each object category. The mean AP (mAP) is then calculated as the average AP across all 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).

# 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

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

B. Method 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 CenterNet2 [19] 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 [49] and SAPNet [50], a spatial attention-based domain adaptation network for the performance measurements. A novel Semantic-complete Graph Matching (SIGMA) method [33] is also introduced in the model comparison to have better diversity in the methods. Finally, we introduced ConfMix [32], a sample mixing-based paradigm of DA for state-of-the-art comparisons. Table II presents the 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 class-wise performance, we notice that while other methods ultimately failed to give any result on the stadium class, our DCLDA method showed a significant gain of 7.9% mAP of this particular class. Also, it is visible that CSP-Darnet53 can perform better than the ResNet50 model with +1.5% of target mAP improvement. Figure 7(a) and (b) shows the detection performance of DCLDA trained on DIOR source data and tested on the DOTA target dataset.

The Visdrone and UAVDT video datasets are two high-variability videos captured from UAV in Table III. Here we evaluate 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 the debiased contrastive learning and pseudo labeling. Also, we have a +2.1% gain margin compared

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 VI: 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.

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 and Figure 7(c) and (d) demonstrate the effectiveness of our method in detecting objects from challenging and less frequent categories, including trucks and buses. Table III also demonstrates that a well-designed backbone can enhance the performance by around +2.7% on the video target domain with dominated dense objects.

C. 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. However, when we used the support set to cure the noisy features and guide the IDA process, we not only gained higher mAP in the target dataset but also we were able to make a stable recovery 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 VI 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. However, we achieved our best performance gain by integrating the *IDA* module. We achieved a +11.5% and 10.4% increase in the mAP for DOTA and UAVDT datasets, respectively. Finally, we combined all proposed modules in our *DCLDA* architecture and recorded the optimal performance on both target datasets under careful hyperparameters selection.

Method	Backbone	CGAN	LDA	GDA	IDA	DOTA	UAVDT
Baseline						35.4	26.4
w/CGAN		✓				37.2	27.8
w/LDA	DarkNet53	✓	✓			41.6	30.2
w/GDA		✓		✓		44.5	34.6
w/IDA		✓			✓	46.9	36.8
DCLDA		✓	✓	✓	✓	50.6	41.5

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.

V. CONCLUSION

This paper proposes debiased 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 class-wise 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.

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] M. Kumar, P. Singh, and P. Singh, "Machine learning and gis-rs-based algorithms for mapping the groundwater potentiality in the bundelkhand region, india," *Ecological Informatics*, p. 101980, 2023.
- [2] J. Valente, B. Sari, L. Kooistra, H. Kramer, and S. Mücher, "Automated crop plant counting from very high-resolution aerial imagery," *Precision Agriculture*, vol. 21, no. 6, pp. 1366–1384, 2020.
- [3] S. Workman and N. Jacobs, "Dynamic traffic modeling from overhead imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12315–12324.
- [4] D. Biswas, M. M. 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 *2022 IEEE International Conference on Networking, Architecture and Storage (NAS)*, 2022, pp. 1–8.
- [5] BookingHunterTV, "New york city walking tour part 1 - midtown manhattan," December 2019. [Online]. Available: <https://youtu.be/-IpXdtWfneI>
- [6] 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 Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.
- [7] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord, "xview: Objects in context in overhead imagery," *arXiv preprint arXiv:1802.07856*, 2018.
- [8] 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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
- [9] J. Zheng, Y. Zhao, W. Wu, M. Chen, W. Li, and H. Fu, "Partial domain adaptation for scene classification from remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2022.
- [10] I. O. Tolstikhin, B. K. Sriperumbudur, and B. Schölkopf, "Minimax estimation of maximum mean discrepancy with radial kernels," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [11] L. Xiong, M. Ye, D. Zhang, Y. Gan, and Y. Liu, "Source data-free domain adaptation for a faster r-cnn," *Pattern Recognition*, vol. 124, p. 108436, 2022.
- [12] C. Yan, X. Chang, M. Luo, H. Liu, X. Zhang, and Q. Zheng, "Semantics-guided contrastive network for zero-shot object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [13] 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 Sensing*, vol. 13, no. 21, p. 4386, 2021.
- [14] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
- [15] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 8765–8775, 2020.
- [16] T. D. Jui, G. M. Bejarano, and P. Rivas, "A machine learning-based segmentation approach for measuring similarity between sign languages," in *sign-lang@ LREC 2022*. European Language Resources Association (ELRA), 2022, pp. 94–101.
- [17] M. R. Babavalian and K. Kiani, "Learning distribution of video captions using conditional gan," *Multimedia Tools and Applications*, pp. 1–23, 2023.
- [18] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2778–2788.
- [19] X. Zhou, V. Koltun, and P. Krähenbühl, "Probabilistic two-stage detection," *arXiv preprint arXiv:2103.07461*, 2021.
- [20] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
- [23] L. Shi, L. Kuang, X. Xu, B. Pan, and Z. Shi, "Canet: Centerness-aware network for object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.
- [24] D. Biswas and J. Tešić, "Small object difficulty (sod) modeling for objects detection in satellite images," in *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 2022, pp. 125–130.
- [25] J. Zhang, Y. Shi, Q. Zhang, L. Cui, Y. Chen, and Y. Yi, "Attention guided contextual feature fusion network for salient object detection," *Image and Vision Computing*, vol. 117, p. 104337, 2022.
- [26] 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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.
- [27] Q. Li, Y. Chen, and Y. Zeng, "Transformer with transfer cnn for remote-sensing-image object detection," *Remote Sensing*, vol. 14, no. 4, p. 984, 2022.
- [28] Z. Deng, Q. Kong, N. Akira, and T. Yoshinaga, "Hierarchical contrastive adaptation for cross-domain object detection," *Machine Vision and Applications*, vol. 33, no. 4, pp. 1–13, 2022.
- [29] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8869–8878.
- [30] N. Ma, H. Wang, Z. Zhang, S. Zhou, H. Chen, and J. Bu, "Source-free semi-supervised domain adaptation via progressive mixup," *Knowledge-Based Systems*, vol. 262, p. 110208, 2023.
- [31] C. Xu and X. Tian, "Semantic-aware mixup for domain generalization," *arXiv preprint arXiv:2304.05675*, 2023.
- [32] G. Mattolin, L. Zanella, E. Ricci, and Y. Wang, "Confmix: Unsupervised domain adaptation for object detection via confidence-based mixing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 423–433.
- [33] W. Li, X. Liu, and Y. Yuan, "Sigma: Semantic-complete graph matching for domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5291–5300.
- [34] C. Wu, F. Wu, and Y. Huang, "Rethinking infonce: How many negative samples do you need?" *arXiv preprint arXiv:2105.13003*, 2021.
- [35] G. Bai, W. Xi, X. Hong, X. Liu, Y. Yue, and S. Zhao, "Robust and rotation-equivariant contrastive learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [36] H. Li, Y. Li, G. Zhang, R. Liu, H. Huang, Q. Zhu, and C. Tao, "Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
- [37] D. Biswas and J. Tešić, "Progressive domain adaptation with contrastive learning for object detection in the satellite imagery," 2023.
- [38] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Stanford, Tech. Rep., 2006.
- [39] M. M. Mahabubur Rahman and J. Tešić, "Hybrid approximate nearest neighbor indexing and search (hannis) for large descriptor databases," in *2022 IEEE International Conference on Big Data (Big Data)*, 2022, pp. 3895–3902.
- [40] D. Arthur and S. Vassilvitskii, "K-means++ the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.
- [41] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.
- [42] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [43] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
- [44] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

- [45] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.
- [46] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [47] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang *et al.*, "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.
- [48] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 370–386.
- [49] W. Zhou, D. Du, L. Zhang, T. Luo, and Y. Wu, "Multi-granularity alignment domain adaptation for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9581–9590.
- [50] C. Li, D. Du, L. Zhang, L. Wen, T. Luo, Y. Wu, and P. Zhu, "Spatial attention pyramid network for unsupervised domain adaptation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 481–497.
- [51] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CspNet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390–391.
- [52] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.
- [53] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," <https://github.com/facebookresearch/detectron2>, 2019.
- [54] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.



Debojoyti Biswas is a Computer Science Ph.D. student. He received his B.Sc degree from the Noakhali Science and Technology University, Bangladesh 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. Prior, 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.