Domain Adaptation with Contrastive Learning for Object Detection in Satellite Imagery

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

***Abstract*—State-of-the-art object detection methods applied to satellite and drone imagery largely fail to identify cross-domain small and dense objects. The high content variability in the overhead imagery is due to the different sensors, terrestrial regions, lighting conditions, and the image acquisition time of the day. Moreover, the number and size of objects in aerial imagery are very different than in the consumer data. We propose a small object detection pipeline that improves the feature extraction process by spatial pyramid pooling, cross-stage partial networks, and heatmap-based region proposal networks. Next, we propose the instance-aware image difficulty score that adapts the overall focal loss to improve object localization and identification. Finally, we add the two progressive domain adaptation blocks using contrastive learning in the pipeline. The blocks align the local and global features extracted from the customized CSP Darknet backbone, as the different levels of feature alignment alleviate the degradation of object identification in previously unseen datasets. We create a first-ever domain adaptation benchmark using contrastive learning for the object detection task in highly imbalanced satellite datasets with significant domain gaps and dominant small objects from existing satellite benchmarks—the proposed method results in up to a 7.4% and 4.6% increase in mAP over the best state-of-art method for the DOTA and NWPU-VHR10 datasets, respectively.**

***Index Terms*—Object Detection, Domain Adaptation, Satellite Imagery, Debiased Contrastive Learning, Remote Sensing Image**

I. Introduction

Deep learning has succeeded in diverse computer vision applications [1], [2], [3], [4]. Still, there is a growing need for automated object localization and identification systems for overhead imagery for traffic control, national parks, wilderness areas, natural disaster surveillance, agriculture, maritime piracy, etc. Research efforts are underway in precision agriculture [5], emergency rescue [6], terrestrial and naval traffic monitoring [7], and industrial surveillance [8] to integrate accurate automated object localization and identification in overhead systems. Oriented object detection has been used for the precise localization of target objects [9], [10]. The challenge lies in the fact that due to high ground sample distance (GSD), the aerial imagery content varies significantly within the same area of capture or drone flight. Several factors are responsible for this dramatic change, such as significant changes in light conditions (time of day, season, weather) and the type of terrestrial terrain captured in the imagery. The

Fig. 1: Visual difference between consumer images [11] and satellite images [12]

variation between datasets, including multiple dates, terrains, missions, object distribution and sizes, and lighting conditions, is even higher. Figure 3, (a) shows the variations due to image capture time and lighting conditions, Whereas Figure 3(b) illustrates that an object can be as small as 0.01% and as large as 70% of an image. Figure 3(b) also shows a loosely packed nature vs. the densely packed small object characteristics in satellite images. Further, we claim that geographical variance is a critical challenge for domain adaptation tasks on satellite images in Figure 3 (c); we also see that geographical variance can also exist in a singular experimental dataset due to image capture from regions of the world. Furthermore, in Figure 2, we provide empirical evidence by plotting the validation set image features fetched from the source (DIOR) and both target domain (DOTA2.0, NWPU-VHR) datasets. To plot the features in 2D feature space, first, we extracted the image features from the backbone of the baseline CenterNet2 network. Next, we used T-SNE to plot the 1024 dimension features in 2D space. It is clear from Figure 2(a) and 2(b) that the domains are pretty different due to the above-mentioned variations, and both target datasets are reasonable candidates for domain adaptation tasks.

The examples of consumer images vs. overhead images are illustrated in Figure 1. Advancements in deep neural networks and greater availability of computational resources have led to enhanced object detection techniques in consumer images [13], and the improvements include improving the Region Proposal Network, better backbone, and integration of the weight-based loss function for hard-example mining. State-of-the-art object detection and domain adaptation (DA) modeling approaches developed for consumer images do not translate to overhead imagery due to visual variation within the cross-domain image, the relative object size w.r.t the image, and the object density in an image [14]. Recent advances in model detection in overhead images do not address the domain shift and the challenges due to high GSD in an unsupervised setting with a single pipeline [15]. The object identification models for UAV/drone images tend to perform better in low-altitude datasets with relatively

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.

fewer small and dense objects compared to our experimental datasets [16].

Deep neural networks require large and diverse amounts of annotated training data to guarantee reliable object localization in unseen datasets. Collecting and annotating aerial datasets has proven difficult and complex due to many small and dense objects per image [17]. Only three datasets with rich class distribution are being used to benchmark the results to date: DOTA2.0 [18], and DIOR [12] are two satellite image collections, and VisDrone is the overhead drone collection [19]. Domain adaptation successfully shares acquired knowledge (regarding annotations and learned models) in the source domain with the target domain. Domain adaptation has risen as one of the approaches to speed up the pseudo-labeling of objects in the target domain using source labels. Domain adaptation for object recognition in consumer image datasets successfully addresses weather, lighting conditions, geological variance, variation in image quality, and cross-camera adaptation by aligning the global feature distribution of data from the origin and target domains [20]. Recent State-of-the-art (SOTA) work of unsupervised domain adaptation for aerial imagery uses the reconstructed feature alignment method instead of adversarial-based feature alignment to avoid background noise alignment [21]. Nevertheless, limited progress in the domain adaptation task has been focused on satellite imagery. Therefore, there is a pressing need to delve into new ideas and techniques to achieve more improved and favorable results.

In this work, we intend to explore contrastive learning on local and global image features to perform feature alignment on task-specific layers. Contrastive learning is a technique that evaluates pair-to-pair relationships by measuring the similarities between different pairs, such as query-positive or query-negative. It groups similar features closely and dissimilar features at a distance in the feature space. We use a random image feature as the query sample and its augmented version as the positive sample. Negative samples are the image features in a mini-batch that are not part of the query and positive samples. This paper introduces the first domain adaptation benchmark for large-scale satellite image datasets. To reduce the global gap between the source and target domains, we create two intermediate domains using the CycleGAN modeling [22]. Then, we extracted the local and global features extraction from feature pyramid network (FPN) layers using the adapted domain adaptation approach. Next, we introduce difficulty-weighted focal loss (DWFL), which uses the number of foreground proposals and amount of neuron activation and assigns a difficulty score for a particular image. Finally, we introduce the noise-contrastive estimation (InfoNCE) [23] loss to produce domain-invariant features. The methodology is outlined in Section IV, and the novelties of the proposed methods are:

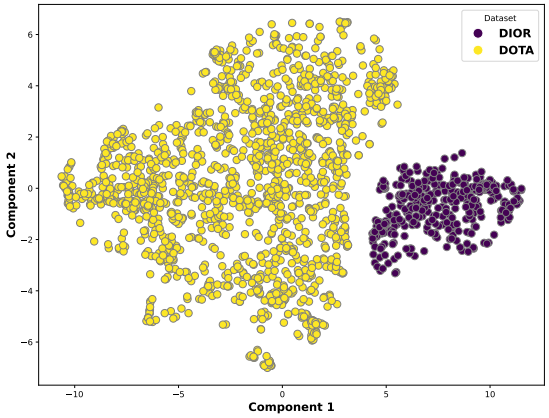
1) Using the local-global feature alignment from the source and target datasets using contrastive learning-based domain adaptation.
2) Integration of a novel difficulty estimation method and DWFL in the detection pipeline.
3) *Multiple* numbers of negative samples for debiased contrastive learning and object detection tasks.
4) Progressive domain adaptation by creating an intermediate domain and minimizing the domain gap between source and target datasets.
5) Finally, with the help of DWFL and local-global feature alignment, the proposed method results in up to a 7.4% and 4.6% increase in mAP over the best SOTA method for the DOTA and NWPU-VHR10 datasets, respectively.

The rest of this article is organized as follows. Section II summarizes related work, and Section III introduces the proposed difficulty-based small object detection pipeline. Next, Section IV describes the contrastive learning approach and the different feature alignment modules in the architecture. In Section V, we evaluate the proposed framework using the latest cross-domains detection benchmarks over three different high-altitude (DIOR, NWPU VHR-10, and DOTA2.0) remote sensing datasets and finally summarize the findings in Section VI.

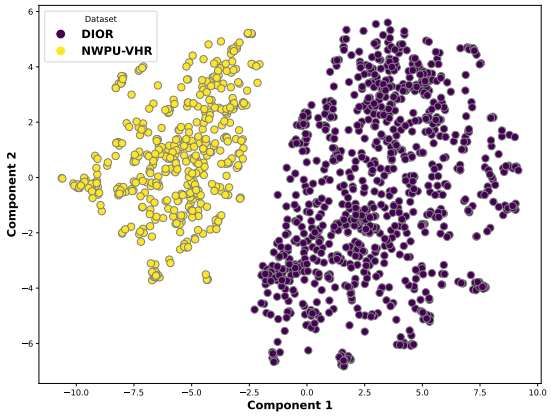
II. Related Work

Object Detection The latest object detection techniques are classified into single-stage or multistage detectors. As the name suggests, single-stage detectors aim to predict object bounding boxes and class labels directly from a single network pass [25], [15], [26]. Single-stage detectors do not have separate neural network modules to generate object proposals and rely on anchor boxes of varying scales and aspect ratios. Single-stage object detection architectures may struggle with accurately localizing small or densely packed objects due to the limited receptive field of the network [18], [27]. However, Cheng et al. [28] proposed an efficient approach to generate anchor-free region proposals for accurate localization using oriented object detection. Multistage detectors are often more accurate and computationally expensive than single-stage detectors due to their use of region proposal networks (RPN) and the non-maximum suppression technique (NMS) to refine the regions of interest in images [29], [30]. NMS filters out positive instances by rejecting overlapped proposal regions in the image with the help of IOU [31], [32] threshold. CenterNet2 [24] is a heatmap-based two-stage approach with balanced positive/negative samples per batch. It uses a Gaussian filter to create a heatmap peak at the object's center to define the proposal regions [24]. The anchor to the object is the region's center based on location. Thus, the one anchor per object eliminates the need for the non-maximum suppression filtering of the overlapping proposals without affecting the quality of the proposal. Due to its superior characteristics in finding densely packed and small objects, we chose CenterNet2 as our baseline architecture.

Small and dense Object Detection As the object's size decreases, the chances of losing local information in deep layers increase significantly. The outcome of the small object detection depends on how well the backbone network [1], [3], [33] captures the region features from the input image. Next, different scale features from various stages of the backbone have been successfully used for other scale predictions. The feature pyramid network (FPN) layer also helps to strengthen standard spatially rich features by combining semantically



(a) Yellow circles represent the target dataset (DOTA), and blue circles represent the source dataset (DIOR)



(b) Yellow circles represent the target dataset (NWPU-VHR), and blue circles represent the source dataset (DIOR)

Fig. 2: High domain gaps illustration using image features extracted from source and target domain validation set.

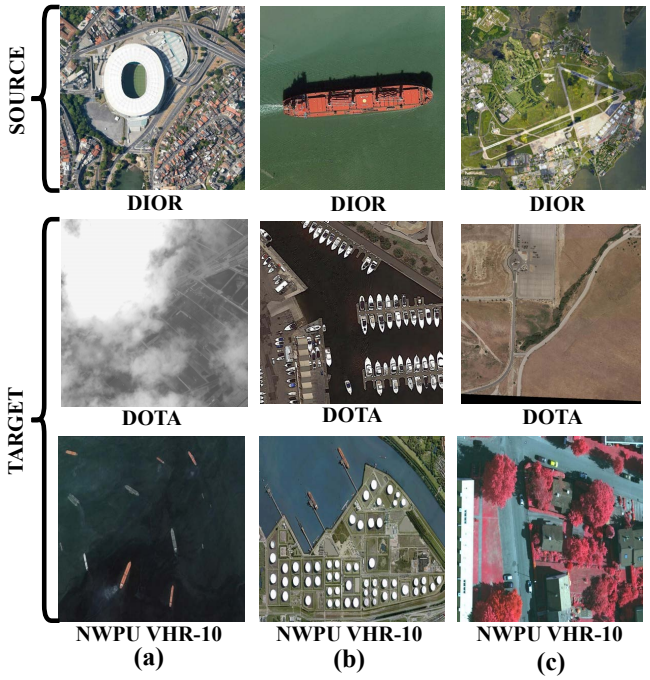


Fig. 3: Domain variations from different aspects in our experimental datasets: (a) lighting conditions, (b) object shape and scale, and (c) variation due to geographical change

rich features by combining low-level and high-level features with the fuse connection and up-sample method. An improved FPN module uses a similarity-based fusion method capable of extracting information for various sizes of instances [34]. Authors argue [35], pixel-level appearance features do not contain enough information to localize small objects in an image, the global context aggregation module and the feature refinement module to build Global Context-Weaving Network are required for optimal performance in small object detections [36]. Hence, context-based feature extraction is more robust for complex object and scene detection and performs better in benchmark datasets [37]. Yang et al. [38] propose *querydet*, which first predicts the coarse locations of small objects on low-resolution features and then computes the accurate detection results using high-resolution features sparsely guided by those crude positions. This work does not rely much on low-level feature queries because it is hard to distinguish small objects with very low resolution on feature maps. The above discussion verifies that small object detection requires global/high-level features to classify objects accurately, which motivates us to use features with high receptive field from later layers in conjunction with the Multi-layered perceptions (MLP) module to calculate the difficulty score for an image (see details in section III).

Domain Adaptation for Object Detection Domain Adaptation techniques are used to handle the problem of domain change between source and target data sets. In the last few years, the unsupervised Generative Adversarial Network (GAN) has been critical in solving the domain shift problem. The GAN-based approach expands object detection in consumer images to other domains. The GAN-based domain adaptation model uses a Gradient Reversal Layer (GRL) to learn domain invariant features. Hsu et al.[39] use adversarial learning to align the distribution of characteristics between domains and perform progressive domain adaptation to address the problem of significant domain gaps. Saito et al. show domain adaptation should not be rigorous and uniform at every point in feature space, and careful inspection is required to align features that are very close to the inter-class boundary to achieve optimal results [40].

On the other hand, mixing-based strategies have succeeded in consumer image domain adaptation tasks. Mattolin et al. [41] employed the confidence-based mixing (ConfMix) technique to integrate source and target domain images. This integration process involves the computation of instance proposal confidence, which is determined by considering both the objectness score and bounding box uncertainty score associated with each instance proposal within the image. Song et al. [42] propose a similar network to our work; they propose to use

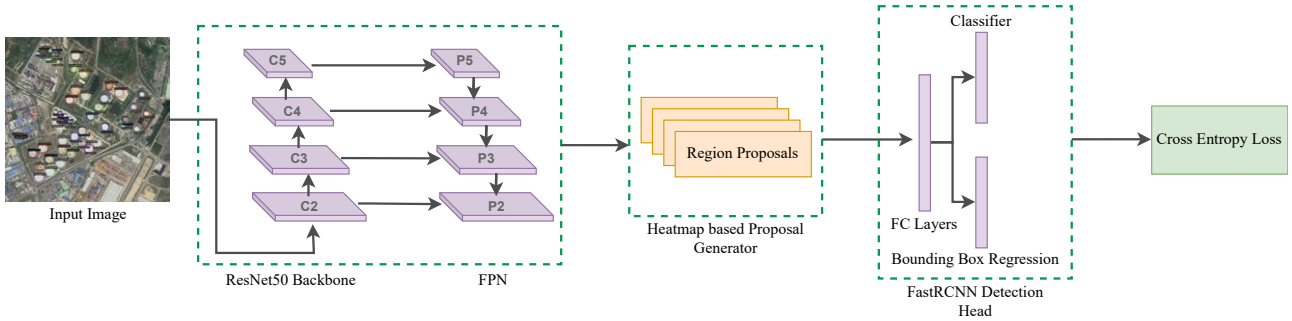


Fig. 4: *CenterNet2*: Heatmap based multistage small object detection model used as a baseline [24]

local object-aware feature alignment and cross-scale attention-aware feature alignment to highlight more foreground information and achieve better accuracy. The Semantic-complete Graph Matching (SIGMA) framework, introduced in the study by Li et al. [43], represents a novel approach for the Domain Adaptation task. This methodology addresses incongruities in semantics and redefines adaptation through graph matching. The major component within SIGMA, the Graph-embedded Semantic Completion module (GSC), specifically targets mismatched semantics by generating hallucination graph nodes corresponding to the absent categories.

The maximum mean and central moment discrepancy approaches successfully produced domain-invariant features through the alignment of feature space. Long et al. proposed DAN [44], which matches the mean embedding of different domain distributions from different task-specific layers in CNN, and Zeillinger et al. [45] proposed to use means of order-wise moment differences to match the higher-order central moments of probability distributions. State-of-the-art guides knowledge transfer between domains while maintaining consistency of the relevant semantics before and after adaptation [46]. Class-level distribution alignment across the source and target domains was achieved using the Easy-to-Hard Transfer Strategy and a Prototype Feature Alignment Network [47]. However, some adversarial adaptation methods aim to mitigate the need for uniform alignment among all samples; their effectiveness remains constrained by the challenge of achieving accurate unsupervised adaptation.

Contrastive Learning for Domain Adaptation The idea of representation learning in contrastive fashion was first introduced by Hadsell et al. [48] for the dimensionality reduction task. Later, contrastive learning became more popular with the InfoNCE [23] loss for clustering similar features together and dissimilar features at a distance in feature space. Contrastive learning has gained much attention for feature representation learning for tasks such as domain adaptation [49], video anomaly detection[50], and video-text alignments [51] due to its straightforward work process to produce similar features across domains. However, it maintains discriminating characteristics in task-specific features to represent different classes in a singular domain. It uses a similarity function such as cosine similarity or Euclidean distance to measure the similarity between two vectors. As we also work with feature vectors at both image and instance-level for satellite images, we decided to follow feature learning with contrastive loss for our unsupervised domain adaptation task.

Several versions of InfoNCE loss were derived based on different applications. Some focus on the number of negative examples we need to design optimal contrastive learning. Wu et al. [52] argued that real-world datasets often introduce noise, and incorporating too many noisy negative examples yields sub-optimal results. Chuang et al. [53] propose reducing bias in contrastive learning by carefully selecting negative examples. The selection of *False Negative* examples can disturb the learning process and harm the overall performance. Contrastive learning is successful not only in single-source domain adaptation but also in multi-source domain adaptation [54]. Kang et al. [54] consider class information and label the target dataset using the K-means clustering method. Kalantidis et al. [55] use *hard negative mixing* strategy to amplify the effect of negative samples with very minimum overhead computation. Motivated by these works, we propose a new version of InfoNCE loss for the domain adaptation task. We present contrastive learning that is less susceptible to false negatives from highly imbalanced datasets by carefully selecting negative examples.

III. Small Object Detection

The satellite image has a maximum of 400 million pixels and objects are frequently smaller than 100 pixels. A typical patch of the image is 1024×1024, which equals 1.05 million pixels. If an object is 10×10 or 100 pixels, the object's size is < 0.0001 of the area of the image. The success of object detection is contingent on the reliability of the pixel- and object-feature extraction, as well as the RPN-based proposal network within the DNN architecture. An increase in the number of small, densely packed objects raises the possibility of losing pixel-level information during feature extraction. The RPN-based proposal network can miss small objects early in the processing, leading to difficulty in detection at later stages [56]. Furthermore, dense object arrangements result in extra noise from surrounding information during input and more post-processing operations.

A. Baseline Model for SOD

Figure 4 shows the *Base* model architecture, a pipeline adopted from the CenterNet2 model [24] with 3 components:

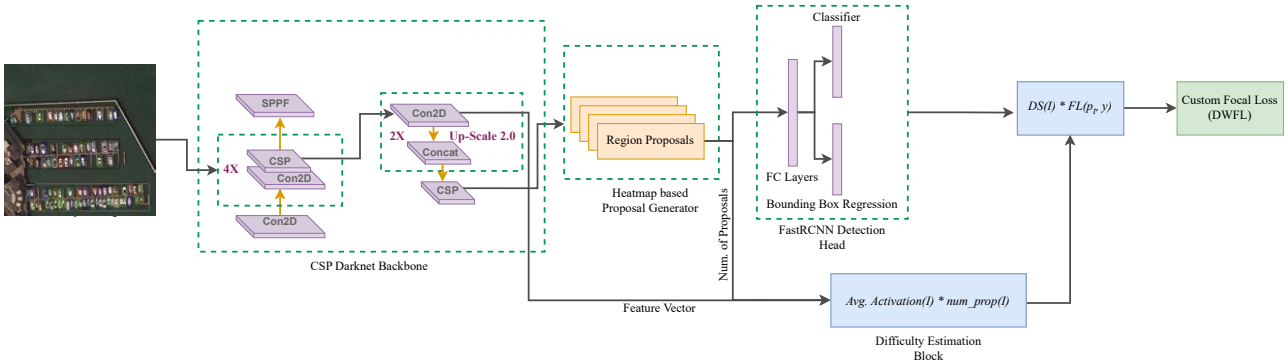


Fig. 5: *SOD* model: Small Object Detection model with updated backbone, new loss function, and the difficulty estimation block.

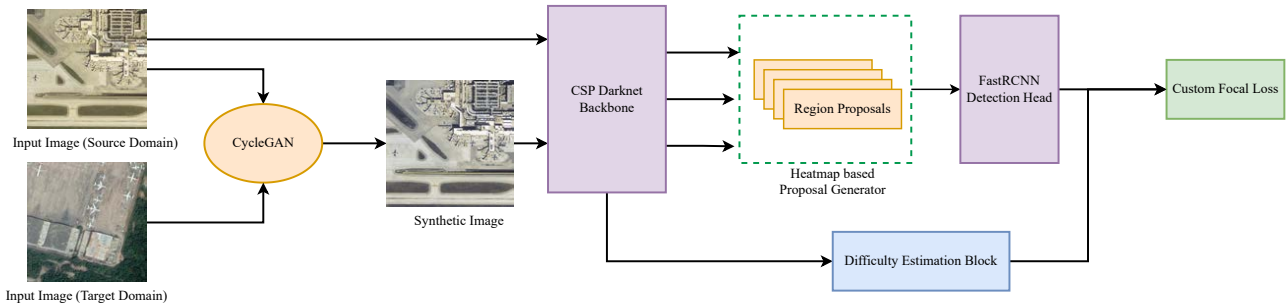


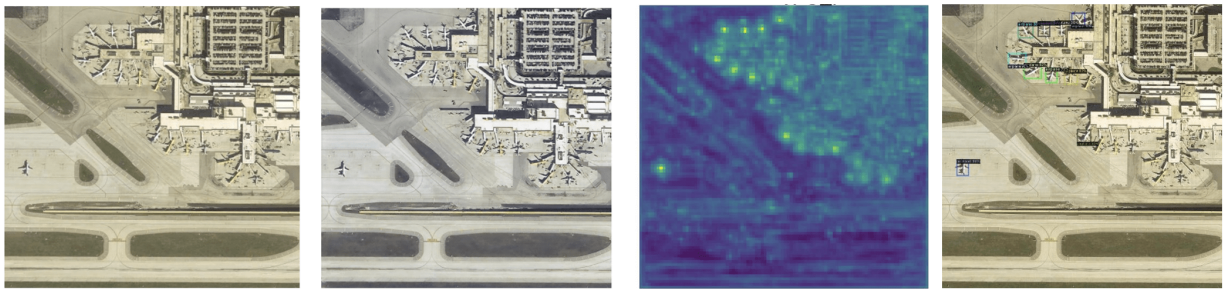
Fig. 6: *HeatDA*: Heat Domain Adaptation model with transfer learning and CycleGAN translated image domain adaptation.

Backbone, RPN, and Detection Head. To enhance performance on overhead datasets, the image size, output channels, and IOU were optimized in the FastRCNN Detection Head. The **Backbone** employs ResNet50 as a feature extractor and FPN for multi-scale predictions, combining features from prior layers via residual connections to prevent vanishing gradients problems. The Base model, with ResNet50, achieved leading results on COCO [11] and LVIS [30] datasets. The FPN layer extracts features at different scales from different Backbone layers, shown in the FPN block in Figure 4. The Resnet(C3), Resnet(C4), and Resnet(C5) blocks represent strides of 8, 16, and 32 in the network. P3, P4, and P5 show three scale prediction FPN outputs, which are fed into the RPN block described in the following paragraph for scale predictions.

Region Proposal Network (RPN) in the *Base* model [24] creates region suggestions through heatmaps by applying Gaussian kernels on features from the FPN at different scales. The heatmaps are produced by comparing the max-pool input and the Gaussian kernel output element-wise. Max-pooling highlights each pixel in the feature except for local maxima, which have a value of 1. Each peak in the heatmap represents the center of an object, as shown in Figure 10 (a) and (c). The features at each key point are used to determine the size of objects, leading to accurate bounding boxes even when objects are close or overlap. To improve performance on overhead imagery, the baseline model requires advanced *image augmentation* techniques, enhanced *feature extraction* methods, and an increased *proposals per image* and *detection per image*. **Detection Head** is based on the Faster-RCNN detector [31]. It takes filtered region proposals from the RPN as input. It processes them as follows: 1) Convert proposals into 7×7 grids with the same number of channels via region-of-interest (ROI) pooling, 2) Flatten the pooling output and feed it into FCN for final detection output, (N, C) for class predictors for C classes, and $(N, 4)$ for N region proposals bounding boxes.

B. Small Object Detection pipeline

An extended version of our *SOD* model for small object detection [57] improves upon the Base model by optimizing the pipeline for detecting small objects. We replaced the backbone with CSP Darknet [26] and added the Difficulty Estimation block. We also switched to a modified focal loss instead of cross-entropy loss (Figure 5). The effectiveness of the RPN module depends on the backbone's performance. If the backbone fails to extract meaningful features for the small object in the image, the RPN module will likely fail to include the small object in the region proposals. Our findings found that 75% of RPN proposals in the Base model were trivial and repetitive (backgrounds, partial objects) as explained in [57]. To enhance the model, we used the CSP Darknet backbone for preserving better semantic information in deeper CNN layers [26], [58]. The research introduces a partial cross stage for low/high-res aggregation and replaces max-pooling with spatial pyramid-pooling for finer feature extraction.



(a) Source Domain (b) Target Domain(CycleGAN) (c) Heatmap HeatDA (d) HeatDA Prediction

Fig. 7: Domain translation from Source(S) to Source as Target(SaT) using CycleGAN [22] and Object Detection from HeatDA model.

The proposed small object detection model is illustrated in Figure 5. The pipeline concatenates several low-level features with high-level features channel-wise, thus propagating the semantic information from the lower to higher levels. The region proposal network considers the multi-resolution features from different CSP-Bottleneck layers for proposal generation, as illustrated in Figure 5. **Difficulty Estimator (DE)** module numerically captures the complexity of an image feature based on the number of foreground object proposals and the amount of neuron activation information in the network for an image. It calculates the overall difficulty of each image by taking feature input from different stages of the feature network. The difficulty score (DS) for an FPN feature level with a resolution of $C \times W \times H$ for the image I is calculated in Equation 1.

$$DS(I) = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{C*W*H}\sum_{c=1}^{C}\sum_{w=1}^{W}\sum_{h=1}^{H} f_{c,w,h}(I)\right) * \ num\_prop(I) \quad (1)$$

The feature output channels, width, and height at FPN level (l) are represented by *C, W, H* respectively. In contrast, L represents the number of FPN levels used to calculate difficulty. The difficulty score (DS) at the FPN level is calculated by dividing the sigmoid linear unit (SiLU) activation values $f_{c,w,h}(I)$ at all pixels in an image (I) by the total dimension of C, W, H. Using this block we calculate the number of total neurons fired for a single image in the forward pass. The difficulty score DS is derived from three FPN levels, averaged, and multiplied by the number of proposals (See Equation 1). An image can contain a single big object can have a high amount of neuron activation and a high-intensity value. The high amount of region proposals is more likely to correspond to a higher number of dense small objects. The intensity values are not always reliable for the overhead imagery. Thus, we compute the difficulty score DS as a multiplication of the intensity value with the neuron activation. The DS after multiplication can be very high; for smooth learning, the final DS value for an image (I) is obtained by normalizing the DS value between $[0.7, 1.4]$. The ultimate goal of the DS is to perform as the weight parameters in the focal loss function (See Equation 2). So, we do not penalize the loss very much to avoid longer convergence time as well as vanishing gradient decent issues due to very small weight updates. We found that an easy example with minimum DS=0.7 and a difficult example with highest DS=1.4 works well for experimental datasets. The complexity increase for the DS block is minimal with constant time PyTorch operation, and its computational time, expressed in Big *(O)* notation, is *O(r)*, where r is the number of iterations during training.

$$FL(p_t, y) = \alpha_t * (1-p_t)^{\gamma} * CE(p, y), \quad DWFL(x, p_t, y) = DS(I) * FL(p_t, y), \quad (2)$$

Difficulty weighted Focal Loss is calculated from the difficulty scores for each image, and we propose replacing the loss of cross-entropy with the loss of custom focus, as illustrated in Figure 5. The difficulty scores are calculated using Equation 1 for each image by a difficulty estimator block as a weight factor to focus more on complex images with many small objects and a high variation in pixel-level features. The basic form of the focal loss function is outlined in Equation 2.

$$\forall c \in C, \alpha'_c = -1 * log\left(\frac{|C_c|}{|C_1 \cup C_2 \cup ...|}\right) \Rightarrow \alpha_c = \beta * \frac{\alpha'_c - min(\alpha_c)}{max(\alpha_c) - min(\alpha_c)} \quad (3)$$

The p_t is the probability distribution of the target *t*, and y is the ground truth of the object being a specific class, γ is the modulating factor, α_t is used as a weighting factor, and CE represents the cross-entropy function. We propose a new measure, the Difficulty Weighted Focal Loss (DWFL) in Equation 2 as a product of difficulty score, $DS(I)$ in Equation 1, and focal loss for the image, $FL(p_t, y)$ in Equation 2. The value α is used in the $FL(p_t, y)$ calculation to control the class imbalance problem in our source and target data sets. The α_c is calculated as in Equation 3 for each class, where the modulating factor α'_c depends on the frequency $|C_c|$ of a particular class in the data set and $|C_1 \cup C_2 \cup C_3...|$ is the total number of all instances of all classes in the data set. The normalized α_c values from Equation 3 are used across different classes c, $c \in C$ to mitigate the imbalance of object class labels. In the experiment section V we confirm that the proposed normalization of α_c in Equation 3 is more effective and gives a stable loss calculation for a highly unbalanced class count in the data set.

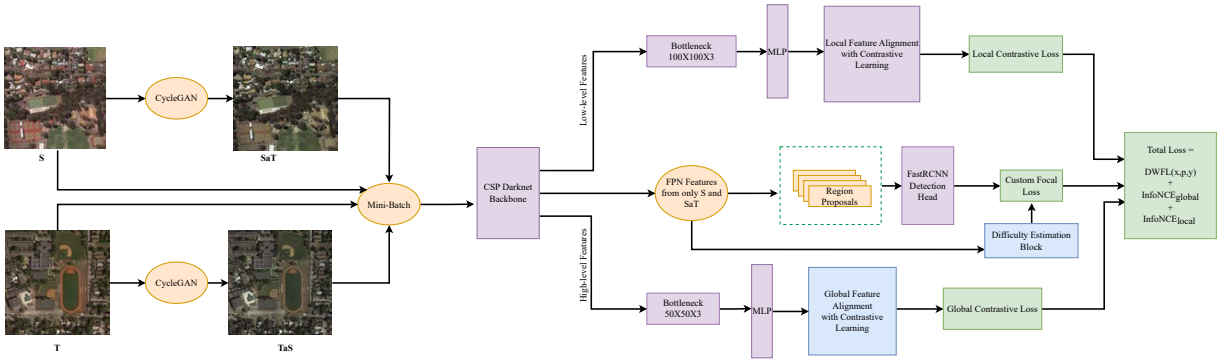


Fig. 8: Local and Global Domain Adaptation(LGDA) model with contrastive learning.

IV. Domain Adaptation Methods

Different overhead image datasets are usually taken at different geographical locations, and different types of satellites were used to capture images with different orientations under various weather and lighting conditions. There are a handful of annotated overhead image collections [59], and they all have different object class annotations, both in frequency and assigned object labels. This contributes to a large domain gap between our source and target data sets. Object detection performance on target degrades drastically when the domain gap is very w.r.t source dataset. Domain adaptation (DA) methods are key to solving this problem. Using domain adaptation methods, we can perform better in unseen datasets not introduced during the training phase. The self-supervised or unsupervised domain adaptation aims to produce invariant features for a particular class across domains. In the experiment section V, we confirm that the domain adaptation of the source in the training process improves the object detection performance.

A. HeatDA *model: Heat Domain Adaptation Model*

Here, we propose a pre-processing step for the pipeline outlined in Figure 5 and map the source domain into the target domain first, as illustrated in Figure 6, as we have found that closing the source and target gap using progressive domain adaptation leads to better object detection in the previously unseen overhead imagery. Using source and target image examples, we train the CycleGAN domain discriminator [22], [60]. The resulting domain discriminator model translates the source image to the target domain, as illustrated in Figure 7. This additional CycleGAN domain discriminator model illustrated in Figure 6 allows us to align pixel-level features between two domains and use the source image and the translated source image to train the SOD network, as illustrated in Figure 6. The conversion of the source to the target domain allows us to incorporate target-like domain characteristics without relying on the object-level annotations that might or might not be present. Training the *HeatDA* using target-like images helps to align pixel-level semantic information for the source and target domain, thus improving the detection performance (see Figure 7) in the target dataset.

B. LGDA *Model: Local Global Domain Adaptation Model*

Contrastive learning [48], [61] is a simple process of measuring pair-to-pair relationships based on the similarities between different pairs, such as query-positive or query-negative. Figure 9 illustrates the functional strategy of contrastive loss. Feature representation of the source and target objects differs in the feature space, and there is a huge gap due to lightning, geographic, weather, and acquisition differences, and the difference is illustrated by a green line in Figure 9. Contrastive learning brings similar points to close together and pushes dissimilar points separate from each other by calculating similarities between pairs [54], [20]. A pair of feature vectors with high similarity are placed close together, and vector pairs with low similarity are placed distantly in feature space. In the ideal case, the contrastive domain adaptation maps the feature space of the source dataset to the target dataset so that the features representing objects in the same class in the source and the target domain dataset are closer together. In this light, we propose to enforce contrastive learning on local features to minimize the domain gap w.r.t the local characteristics in the image i.e. color and texture captured by deep features of the pixel and its nearest surroundings. Here, we produce domain-invariant object features for source and target domains by increasing similarities between *Query-Positive* pair and decreasing similarities between *Query-Negative* pairs. We introduce the Informative Noise Contrastive Estimation (InfoNCE) loss measure of finding similarities and dissimilarities between features in Equation 4. Similarity of two features u and v is captured by a cosine score $sim(u,v) = u^T/(||u|| * ||v||)$. The *Query* is from the source, the *Positive* example is from the translated source domain denoted as SaT in Figure 8, and negative examples are with different classes than the *Query* example and from the target domain. Contrastive learning increases the similarity between *Query* and the *Positive* sample and increases the dissimilarity between *Query* and N negative samples $Negative_n, n \in [1, N]$ s outlined in Equation 4.

$$InfoNCE = -log \frac{exp(sim(Query, Pos)/\tau)}{\sum_{n=1}^{N} exp(sim(Query, Neg_n)/\tau)} \quad (4)$$

The Informative Noise Contrastive Estimation (InfoNCE) loss measure is low when the similarity between the *Query* and the *Positive* example is high and when then the similarity

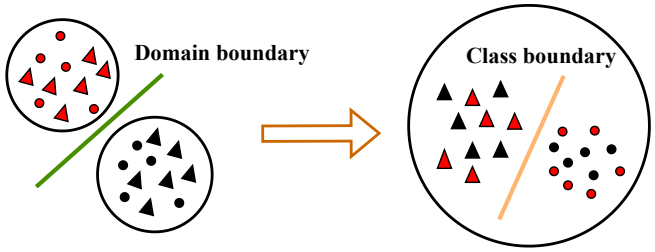


Fig. 9: Domain Adaptation with contrastive learning. Here, different colors indicate different domain distributions, and the different shapes represent different classes in a domain. The Green and Orange lines represent the domain and class boundaries, respectively.

between Query to all *Negative* examples is low. Using this loss, we are learning domain invariant features. N is the mini-batch size during the training phase, and τ is the temperature that controls the strength of penalties in hard negative samples. Our implementation ensures that we find a similar example as *Positive* case and dissimilar examples as *Negative* cases. Figure 8 illustrates the proposed Local Global Domain Adaptation *LGDA* model and incorporates the proposed approach in the small object detection framework. This architecture focuses on performing domain adaptation on a highly class-label imbalance dataset where labeled objects are small compared to the image size. We added two modules to two new modules: Local Feature Alignment and Global Feature Alignment for contrastive learning. Our proposed model takes input from four different distributions; among them, two are the source and target domains, and the other two are new intermediate domains, Source as Target (SaT) and Target as Source(TaS), from the source and target datasets, respectively, generated from the CycleGAN [22] network to reduce the gap between the source (S) and target (T) domains.

As shown in Figure 8, mini-batch inputs are passed into *CSP DarkNet* backbone. Next, local and global features extracted from the backbone are fed into Local Feature Alignment and Global Feature Alignment modules for calculating *Local Contrastive Loss* and *Global Contrastive Loss*, respectively. We use the generalized Equation 5 and 6 for calculating the losses, where F represents the specific type of feature used for contrastive learning. Finally, RPN produces region proposals from only Source (S) and Source-as-Target (SaT) domain features as we have ground truth for these two domains. The later part of the architecture is a traditional RCNN-style object detector with classification and regression modules.

$$InfoNCE^{S,SaT}_{Features} = -log\frac{exp(sim(F^S_i, F^{SaT}_i)/\tau)}{\sum_{j=1}^{N} exp(sim(F^S_i, F^{SaT}_j)/\tau)} - log\frac{exp(sim(F^{SaT}_i, F^S_i)/\tau)}{\sum_{j=1}^{N} exp(sim(F^{SaT}_i, F^S_j)/\tau)}, j \neq i \quad (5)$$

$$InfoNCE^{T,TaS}_{Features} = -log\frac{exp(sim(F^T_i, F^{TaS}_i)/\tau)}{\sum_{j=1}^{N} exp(sim(F^T_i, F^{TaS}_j)/\tau)} - log\frac{exp(sim(F^{TaS}_i, F^T_i)/\tau)}{\sum_{j=1}^{N} exp(sim(F^{TaS}_i, F^T_j)/\tau)}, j \neq i \quad (6)$$

Local Domain Adaptation focuses on the local features in an image and assumes there is no ground truth object labeling for the target dataset, only for the source dataset. Local features capture low-level descriptions of a pixel and its neighbors in an image. The images from a mini-batch pass through the backbone, and local features are saved from the earlier layers of the backbone. The saved local features are $256 \times 100 \times 100$, where dimensions are C, W, and H, respectively. To reduce the necessity of computational power and GPU memory and improve similarity computation performance, we pass the features into the bottleneck module as shown in Figure 8 and downgrade the shape to $3\times100\times100$. In Figure 8, S and T are images drawn from the source and target datasets, respectively, where SaT and TaS are the corresponding images of S and T in translated form produced by the CycleGAN network. Let us denote the local feature vectors from *S, SaT, T and TaS* as L^S_k, L^{SaT}_k, L^T_k, and L^{TaS}_k, respectively. Where k is the index of the mini-batch. For the adaptation of the S and SaT domain, we select a local feature $L^S_i \in L^S$ as a query and the corresponding feature from $L^{SaT}_i \in L^{SaT}$ as the positive case. Negative cases are all other local characteristics $L^{SaT}_j \in L^{SaT}$, where $j \neq i$. The local domain loss between *(S, SaT)* and *(T, TaS)* is calculated using the previously defined Equation 5 and 6. We replace the F in Equation 5 and 6 by local feature *L*. The local domain loss between *(S, SaT)* and *(T, TaS)* can be denoted by $InfoNCE^{S,SaT}_{local}$ and $InfoNCE^{T,TaS}_{local}$, respectively.

After accumulating loss for all query images in a minibatch, the total bidirectional local domain adaptation loss can be formulated as below in Equation 7.

$$InfoNCE_{local} = InfoNCE^{S,SaT}_{local} + InfoNCE^{T,TaS}_{local} \quad (7)$$

Global Domain Adaptation relies on the global alignment of features between the source and the target dataset. The global features represent a more abstract formation of objects in the image and are saved from the last layer of the backbone. The shape of the global features is $256 \times 25 \times 25$, where the dimensions are C, W, and H, respectively. Again, the same as for local features, we use a bottleneck module to reduce the size of global features to $3\times25\times25$. Global features are high-level features in the DNN pipeline. Global domain adaptation and feature alignment are also performed at the mini-batch level to restrict computational and GPU memory expense.

The global feature vectors of training mini-batch input: *S, SaT, T, and TaS* are indexed as G^S_k, G^{SaT}_k, G^T_k and G^{TaS}_k, where k is the index of the mini-batch. The global contrastive loss for S and SaT is calculated by selecting a query sample $G^S_i \in G^S$ and a positive case from the corresponding image feature $L^{SaT}_i \in L^{SaT}$ and vice versa. We take the negative cases as all the other global features $G^{SaT}_j \in G^{SaT}$, where $j \neq i$. The adaptation formula for *(S, SaT)* and *(T, TaS)* domains

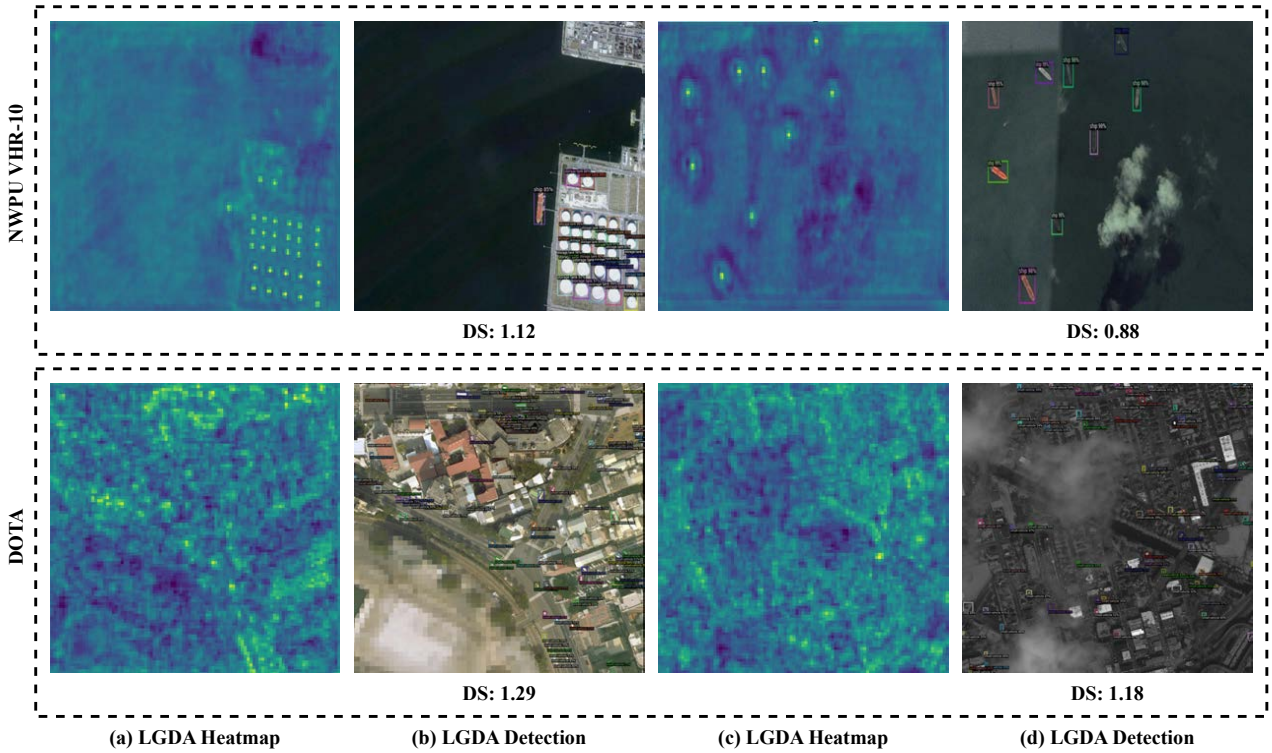


Fig. 10: Proposals heatmaps and detection results from the *LGDA* model for DOTA2.0 and NWPU VHR-10 target datasets. Here DS = Difficulty score for a particular image.

in global feature space can be derived from Equation 5 and 6, respectively, by replacing the F term with global features *G*.

The accumulated bi-directional global domain adaptation loss in a mini-batch is now calculated in Equation 8.

$$InfoNCE_{global} = InfoNCE_{global}^{S,SaT} + InfoNCE_{global}^{T,TaS} \quad (8)$$

Finally, the *LGDA* model combines the local and global contrastive loss with the detection loss and the final loss function is now calculated as in Equation 9:

$$TotalLoss = W_1 * InfoNCE_{global} + W_2 * InfoNCE_{local} + DWFL(x, p, y) \quad (9)$$

In the above Equation 9, the W_1 and W_2 denotes the weight we put on the two different modules.

V. Experiments

We evaluated the proposed approach in the three largest annotated satellite image collections, DIOR [12], DOTA2.0[18], and NWPU VHR-10 [62] with four state-of-the-art domain adaptation models. DIOR is the data set for the source domain, and we used the ground-truth annotation for DIOR in our detection module and model evaluation. DOTA2.0 and NWPU VHR-10 are the target data set that adapts to the local and global domains, as presented in Table I. We assume that the DOTA2.0 and NWPU VHR-10 annotations are unavailable at the object detection training time, and we use ground truth to evaluate the system's performance only at the testing time. In our experiments, we kept only common classes available in the DIOR and DOTA2.0 datasets. We assigned the same class label to each corresponding class in each dataset, as demonstrated for the reduced DIOR, DOTA2.0, and NWPU VHR-10 dataset in Table I. We describe the experimental dataset in Subsection V-A, the setup and implementation details in Section V-B, the performance comparison in Section V-C, and lastly, the ablation study in Section V-D.

Class Name	# of Ins. DIOR	# of Ins. DOTA	# of Ins. NWPU VHR-10
Bridge	176	1039	124
Vehicle	2079	85479	598
Harbor	254	5704	224
Storage.T	2623	5416	655
Stadium	40	393	-
Baseball	250	516	390
Track	138	417	163
Basketball	171	358	159
Tennis	580	1662	524
Airport	25	153	-

TABLE I: Test set instance distribution statistics of the DIOR, DOTA2.0, and NWPU VHR-10 datasets across common categories.

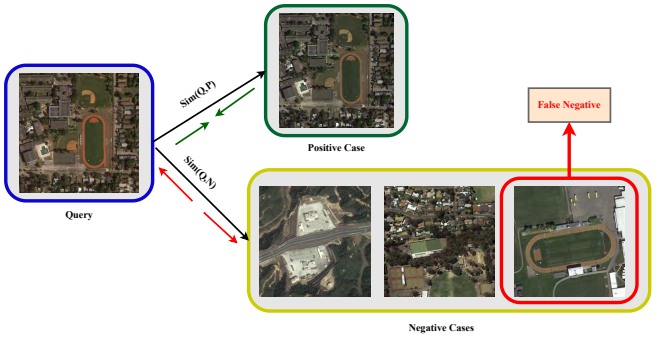


Fig. 11: Example of False Negative occurrence in a highly imbalanced dataset.

A. Datasets

The **DIOR** data set originally consisted of 24,500 Google Earth images from 80 countries. However, 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 dataset boasts a wide range of spatial resolutions, object sizes, object orientations, and a diverse class distribution, as shown in Fig. I. The spatial resolution of the images ranges from $[0.5m, 30m]$, with each image measuring 800×800 pixels. The data set has 97,450 annotated objects, classified into ten classes [12]. Out of the 11,402 images, 10,888 are in the training set, and the remaining 512 images are in the testing set.

The **DOTA2.0** dataset comprises 2,430 overhead images sourced from Google Earth, Gaofen-2 (GF-2), and Jilin-1 (JL-1) satellites [18]. The image sizes range from 800×800 to $29,200 \times 27,620$ pixels. The GSD in the data set ranges from 0.1 to 0.87 m, and each image contains an average of 220 objects. In the experiment, large images were split into overlapping tiles of 1024×1024 pixels with a 200-pixel overlap, resulting in 23,300 images from the original DOTA2.0 dataset. The reduced DOTA2.0 dataset has 11,551 images in the training set and 3,488 in the validation set, classified into ten classes (See Table I). In this paper, we interchangeably use DOTA and DOTA2.0 for referring to this dataset.

The **NWPU VHR-10** has in total 800 High-resolution images in the dataset, of which 715 images were collected from Google Earth, and the remaining are very-high-spatial-resolution pan-sharpened color infrared (CIR) images collected from the Vaihingen dataset [63]. The GSD in the dataset ranges from 0.5 to 2m, and the image size range from 800 to 1267. Among the 800 High-resolution images, 650 are positive images with the available target in the annotation. In contrast, the rest of the images do not contain the target object considered a negative image. We only use positive images from the dataset for our Domain Adaptation task experiments. The reduced NWPU VHR-10 dataset has 450 images in the training and 200 images in the test set, distributed among eight classes (See Table I). In this paper, we interchangeably use NWPU VHR-10 and NWPU for referring to this dataset.

B. Experimental Setup and Implementation

Source as Target (SaT) and Target as Source (TaS) synthesized data are created using the PyTorch implementation of the CycleGAN [22], [60] network with the setup: the learning rate was 0.001; the number of training epochs was 2; load size was 800; and the crop size was 640.

The proposed model for object detection architecture *LGDA* is illustrated in Figure 8 and is described in Section IV-B is an extension of the CenterNet2 [24] and SOD [57] model. The work in [57] uses Darknet53 as the backbone as it is shown to preserve better semantic information from the small objects with the help of Cross-Stage-Partial(CSP) network than the residual-based feature extractor networks [26], [58]; RPN heatmap-based approach to identify dense small objects and remove NMS; and the detection block is Faster-RCNN [31].

We developed our code implementation by leveraging an open-source computer vision library **Detectron2** [67] and some part of **CenterNet2** [24]. We implemented two new DA modules for local and global domain adaptation using the contrastive learning technique. To train our **LGDA** model, we have resized all images to 800×800 pixels, and the mini-batch size in each epoch is set to 8. The loss of InfoNCE in Equation 4 requires the same image from different domains as in the query and the positive case. To achieve this goal, we created a custom data loader in PyTorch that fetches the exact image of the different domains in a mini-batch. This custom data loader ensures that the Query and Positive examples are in the mini-batch sample during the training. In our experiments, eight is the size of the mini-batch, and the $8 \times 4 = 32$ images were fed into the *LGDA* model in each mini-batch. The number of negative cases and the temperature were set to 7 and 0.12, respectively, for contrastive learning. Our research found that passing the S and SaT for the object detection task during the full training time makes the model over-fitted to the source domain. To eliminate this problem, we implemented random sample selection, randomly selecting 8 of 16 images from S and SaT and passing them on to the object detection module. Li et al. [68] found that Global Average Pooling (GAP) loses important context information when working with satellite images, so we replaced GAP with the bottleneck and MLP module for channel reduction. 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.

C. Performance Measures and Comparisons

We use Average Precision (AP), and Average Precision (mAP) to evaluate and compare the models. The precision (P), recall (R), F1 score ($F1$), and Mean Average Precision (mAP) are computed using Equation 10 and 11. True positives (TP) are instances correctly predicted by the model, false positives (FP) are instances missed by the model, and true negatives (TN) are instances incorrectly predicted by the model. We use an IOU threshold of 0.5 for all experiments, and the number of proposals per image is 256. Precision

Method	Detector+Backbone	Bridge	Vehicle	Harbor	Storage Tank	Baseball Field	Track Field	B.Ball Court	Tennis Court	Stadium	Airport	DIOR → mAP	DOTA mAP
Baseline [24]	CenNet2 ResNet50	10.5	8.9	42.1	40.6	46.5	31.4	46.7	74.2	0.0	28.3	64.8	32.1
QueryDET [38]	RetinaNet ResNet50	14.1	14.5	38.2	50.8	43.0	33.4	46.6	77.5	5.3	35.1	69.7	35.8
EPM [64]	FCOS ResNet101	10.1	10.8	40.6	47.7	46.2	34.8	48.7	81.9	1.2	35.5	65.5	35.7
MGADA [65]	FCOS VGG16	13.1	10.8	45.9	48.5	46.0	37.7	50.1	84.3	0.0	37.2	66.9	37.3
SAPNET [66]	FCOS ResNet50	10.9	11.0	23.5	24.4	35.3	27.8	32.2	74.1	0.0	22.7	54.7	26.1
MGADA [65]	F-RCNN ResNet101	15.9	14.0	48.1	46.5	47.6	39.3	52.6	87.2	1.8	37.9	72.6	39.4
SOD	CenNet2 Darknet53	13.9	15.8	36.7	48.1	46.3	31.7	45.3	78.6	4.1	32.9	70.1	35.2
HeatDA	CenNet2 Darknet53	14.4	17.4	39.1	50.9	46.1	35.5	48.0	80.1	4.8	37.6	70.1	37.4
LGDA*	CenNet2 ResNet50	22.0	26.8	48.6	59.7	56.9	44.6	56.7	85.4	5.5	38.2	**74.6**	**44.3**
LGDA	CenNet2 Darknet53	24.5	27.9	51.3	62.0	59.2	47.9	58.6	87.8	6.1	38.7	**76.9**	**46.7**
Oracle	Baseline	46.2	40.4	82.6	65.8	64.4	60.0	77.2	93.5	27.2	54.3	**59.7**	**61.2**

TABLE II: Quantitative performance comparisons (mAP) across classes for DIOR → DOTA benchmark(IOU=0.5), where DIOR is considered as the source and DOTA as the target dataset. Class-wise performance is presented only for target dataset.

(P) represents the fraction of relevant instances recovered by the model, while recall (R) measures the fraction of relevant instances correctly identified by the model among all relevant instances.

$$P = \frac{TP}{TP+FP},\ R = \frac{TP}{TP+FN},\ F1 = \frac{2PR}{P+R} \quad (10)$$

$$AP = \sum_{k=0}^{k=n-1} [R(k) - R(k+1)] * P(k),$$
$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP(k) \quad (11)$$

The $F1$ score provides a single measure of the model's performance when given a class imbalance dataset. The mAP is calculated as shown in Equation 11, where n is the number of classes in the test set, and $AP(k)$ is the Average Precision (AP) of class k in the test set. Here, AP is the weighted sum of precision at each threshold (n is the number of thresholds), and the weight is the increase in recall (Equation 11).

Object Detection Comparisons: To compare our proposed model with recent state-of-the-art (SOTA) models, we set a lower-bound and an upper-bound on the performance for each dataset. We use *CenterNet2* as the baseline/ lower-bound for comparisons, where we use annotations from only the *source* datasets during the training phase to evaluate the target dataset. On the other hand, *oracle* is the upper bound, which uses *CenterNet* as the detection model and uses annotation from *Target* dataset while training to evaluate the target dataset. We choose CenterNet2 as our baseline model because of its ability to work better with small and dense objects leveraging the power of heatmap-based RPN. We use our extended version of the SOD pipeline for small object detection architecture. We have compared our *SOD* model performance with a recent small object detection SOTA model *QueryDet* [38] in Table II and III. We can see that our model performs very close to the *QueryDet* method for the DOTA dataset, and outperforms the *QueryDet* by 2.1% of mAP on the NWPU VHR-10 dataset. Moreover, using 12GB GPU memory *QueryDet* trains 2 images per batch whereas using *SOD* we can set batch size up to 8 images, so we decided to continue further experiments using SOD as the small object pipeline. As the domain adaptation SOTA models, we used feature alignment DA methods such as MGADA [65] and a spatial attention-based domain adaptation network SAPNet [66] for the performance measurements. Also, we introduced EPM [64], a domain adaptation framework that accounts for each pixel via predicting pixel-wise centerness and objectness for state-of-the-art comparisons. Later we keep track of the performance improvement of our new proposed model *LGDA* and try to minimize the gap between *LGDA* and *oracle* for domain adaptation task. Except for the *oracle*, all other comparing models use annotation from only the *source* dataset and images from the *source* and *target* datasets during the training phase. Next, we evaluate the performance of the models based on the test set of both *source* and *target* datasets. By this, we show that our proposed UDA models perform satisfactorily on the target domain, and there is no performance degradation in the source domain due to induced noise from the DA operation.

We start our DA evaluation in Table II, using DIOR as the source and DOTA as the target. Table II shows that the CenterNet2-based baseline model gives an mAP of 64.8% on the source and 32.1% mAP on the target dataset. We improve the baseline model with the integration of *Custom Focal Loss*, *Difficulty Estimation Block* and *Strong Backbone* as illustrated in Figure 5 and propose a new model *SOD* verified to work better on satellite imagery [57]. Table II shows significant improvements on both datasets for small objects such as harbors, vehicles, and storage tanks. The range of α_c' values in the DIOR dataset is 0.2 to 0.79, and the range of α_c' values in the DOTA2.0 dataset is 0.15 to 0.96, which represents a very tight scaling factor for *FL* in both data sets. The SOD model also shows reasonable promise in performance improvement with a gain of 6.0% and 3.0% of mAP for the DIOR and DOTA datasets, respectively, by dealing with small objects and challenging images. The first step towards DA operation was introducing transfer learning and using CycleGAN-generated composite target images for the DA training. We name the model as *HeatDA*, which

Method	Detector+Backbone	Bridge	Vehicle	Harbor	Storage Tank	Baseball Field	Track Field	B.Ball Court	Tennis Court	DIOR → mAP	NWPU mAP
Baseline [24]	CenNet2 ResNet50	32.5	27.1	76.5	48.4	30.3	66.1	63.6	67.0	66.8	52.0
QueryDet [38]	RetinaNet ResNet50	43.2	33.9	80.4	54.7	37.1	69.5	70.1	72.6	72.4	57.6
EPM [64]	FCOS ResNet101	49.6	39.4	86.7	61.6	43.9	76.1	75.7	77.0	68.9	64.0
MGADA [65]	FCOS VGG16	48.8	37.1	85.5	59.6	42.3	76.8	72.6	77.2	66.5	62.6
SAPNET [66]	FCOS ResNet50	35.8	22.6	70.1	41.4	25.5	60.1	56.5	61.4	60.8	46.7
MGADA [65]	F-RCNN ResNet101	54.2	43.9	88.5	65.0	49.9	77.2	76.6	77.0	69.2	66.8
SOD	CenNet2 Darknet53	47.5	36.4	83.9	54.5	38.5	69.0	72.6	72.8	71.5	59.7
HeatDA	CenNet2 Darknet53	48.8	38.2	85.5	56.9	49.5	70.6	74.5	73.0	71.9	62.1
LGDA*	CenNet2 ResNet50	54.0	43.9	90.5	62.0	53.9	76.2	78.6	81.0	**73.8**	**67.9**
LGDA	CenNet2 Darknet53	57.6	51.5	90.4	65.7	60.1	79.8	81.3	84.7	**74.6**	**71.4**
Oracle	Baseline	69.5	60.9	96.5	79.4	64.3	96.1	93.6	97.8	**60.6**	**84.4**

TABLE III: Quantitative performance comparisons (mAP) across classes for DIOR → NWPU VHR-10 benchmark(IOU=0.5), where DIOR is considered as the source and NWPU as the target dataset. Class-wise performance is presented only for the target dataset.

uses the target domain pixel-level context and gives 5.3% mAP improvement on the target dataset. Our final proposed model *LGDA* is an extension of *HeatDA* model by adding local and global domain adaptation modules. Compared to the baseline model, our *LGDA* model gains 12.2% and 14.6% of mAP on the target dataset using ReseNet50 and CSP-Darknet53 backbone, respectively. The detection results from *LGDA* model are illustrated in Figure 10, where the proposals generated from heatmaps are shown in Figure 10(a) and (c), and the object detection performance is shown in Figure 10(b) and (d). Figure 10 also shows us the value of image difficulty for a particular image. It is evident that an image with a higher number of objects and pixel diversity score more difficulty value than an image with fewer objects and less pixel variations. Among the other SOTA models, MGADA performs best with an mAP of 39.4%, and SAPNet gives the lowest mAP of 26.1%. However, Our proposed novel contrastive learning method with a small object-focused pipeline helps us to outperform other SOTA models by a minimum margin of 7.3 % on the target dataset.

Method	Backbone	TL	LDA	GDA	DWFL	DOTA	NWPU
Baseline	DarkNet53					32.1	52.0
w/DWFL					✓	35.2	59.7
w/TL		✓			✓	37.4	61.8
w/LDA			✓		✓	40.6	64.2
w/GDA				✓	✓	41.5	66.6
LGDA		✓	✓	✓	✓	46.7	71.4

TABLE IV: Ablation study for our proposed LGDA method. Here, TL= Transfer Learning, DWFL= Difficulty Weighted Focal Loss, LDA= pixel-level local domain adaptation, and GDA= abstract-level global domain adaptation.

Next, we use Table III to demonstrate the DA performance and SOTA comparison for DIOR and NWPU VHR-10 datasets. The DIOR and NWPU VHR-10 are two high-variability image datasets, as shown in Table I captured from satellites. Here we evaluate target dataset performance over eight different categories. It is observed during experiments that we have not only demonstrated outstanding performance on the target dataset but also we have attained a notable 74.6% mAP(refer to Table III) on the source dataset. Our baseline CenterNet2 method trained on only source dataset archives 52.0% of mAP, whereas our LGDA method achieves 71.4% of mAP using contrastive learning with local and global domain adaptation. Also, we have a +4.6% gain margin compared to the best state-of-the-art *MGADA* method. Moreover, compared to the baseline model, we were able to shrink the performance gap between the oracle and our model from 32.4% to 13.0% using the Local-Golbal DA. Table III and Figure 10(b) and (d) demonstrate the effectiveness of our method in detecting objects from challenging and less frequent categories, including track, bridge, and basketball fields. Table III further illustrates that a meticulously designed backbone can augment the performance by approximately +3.5% on the target domain, mainly when dealing with densely populated objects.

Finally, we perform an in-depth performance analysis in Table V using Precision, Recall, and F1-Score for the four proposed models. For the DOTA dataset, we gain the optimal result in all metrics using the final version of the *LGDA* model, with a Precision, Recall, and F1-Score of 46.7, 60.2, and 52.6, respectively. However, for NWPU, we got the optimal value of Precision and F1-Score from the *LGDA* model and recall from the *HeatDA* model.

D. Ablation study

In this section, we first perform an ablation study on each component of our proposed *LGDA* method to demonstrate the effectiveness of each element, as shown in Table IV. Our ablation study is carried out on two target datasets: DOTA and NWPU VHR-10. Next, we perform an ablation study on each important hyper-parameters; the summary of the ablation study is illustrated in Figure 12. Table IV shows that integrating Difficulty Weighted Focal Loss (DWFL) was crucial for our proposed model as we made 3.1% and 7.7% increase in mAP for DOTA and NWPU, respectively. The intermediate version of our proposed model is *HeatDA*, where we investigate the

	Precision				Recall				F1			
Dataset	CenterNet2	SOD	HeatDA	LGDA	CenterNet2	SOD	HeatDA	LGDA	CenterNet2	SOD	HeatDA	LGDA
DOTA	32.1	35.2	37.4	**46.7**	46.9	47.7	50.4	**60.2**	38.1	46.1	42.9	**52.6**
NWPU VHR-10	52.0	59.4	61.8	**71.4**	58.8	63.5	**67.9**	66.8	61.1	61.3	64.7	**69.0**

TABLE V: Comparison of Precision, Recall, and F1 score between our proposed models for the DOTA and NWPU VHR-10 target datasets.

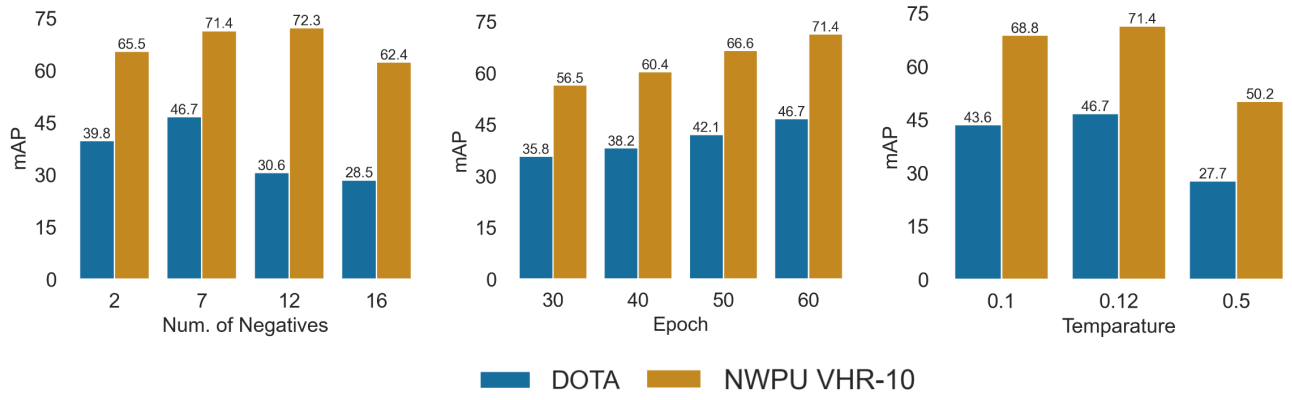


Fig. 12: Ablation Study on Hyper-parameters: (a) mAP vs. Negative Examples, (b) mAP vs. Epochs, and (c) mAP vs.Temperature.

amount of efficiency we can leverage from Transfer Learning, Synthetic images, and DWFL. The Synthetic image generated from GAN networks provides primary pixel-level color and texture information, and we can notice a slight gain of mAP in Table IV for both target datasets. Then we integrate contrastive learning into our DA process and propose two new modules; the first is for aligning local features between the source and target domain, and the latter is for aligning a more abstract view of features with a high-receptive field. The local feature alignment with contrastive learning (LDA) helped us gain over 8% and 12% mAP on DOTA and NWPU VHR-10 by putting a weight of 0.1 on the loss function. On the other side, from the GDA module, we even get better results than the LDA with a weight of 0.01 on contrastive learning. Finally, integrating all small modules, we propose the LGDA method, which achieves 46.7% and 71.4% of mAP on DOTA and NWPU VHR-10 target datasets, respectively.

Next, we look for the optimal value for the number of negative examples for the hyper-parameters study, as shown in Figure 12(a). We started our experiments with a value of 2, and we can see that the model did not perform well with fewer negative examples in the contrastive loss. We increased the value to 7 to generalize learning and recorded our best performance in the DOTA target dataset. Increasing the negative examples further did not help us learn in target datasets due to the imbalanced nature of the data set. The vehicle class dominates our DOTA data set, as shown in Table I. Increasing the number of negative examples also increases the chances of getting false negative (FN) examples, as shown in Figure 11 in contrastive learning. In highly imbalanced datasets such as DOTA2.0, the FN example makes the model biased toward a particular class, and the model's overall performance degrades significantly. However, there was a slight improvement for the NWPU VHR-10 dataset with negative example 12, as NWPU is much more balanced than DOTA. The adverse effect of bias is evident in Figure 12(a) when trained with 16 negative examples. After careful inspection and to reduce computational expense, we set the number of negative measures equal to 7 for both datasets.

Second, our baseline model was trained for 30 epochs, and the LGDA model was trained for 30 more epochs as we added two more loss functions for contrastive learning. As illustrated in Figure 12(b), our experiments found that training for more epochs does not significantly improve performance. Therefore, all results were recorded with 60 training epochs. Lastly, the temperature value in contrastive loss is susceptible and small changes in value can drastically change the outcome. This is evident in Figure 12(c); placing a 50% penalty on contrastive loss dramatically reduces performance, and using a 12% penalty shows the optimal result on target datasets.

VI. Conclusion and Future Work

Object detection in aerial images is one of the most challenging tasks in computer vision research because many small and overlapped objects exist in the photos. The success of DNN object localization depends on a large amount of annotated training data and a reliable feature extractor module in the pipeline. This paper presents a robust feature extractor that captures balanced low- and high-level features for small objects. Moreover, we offer the heat-map-based region proposal module to better grab small things. The domain gap in satellite images is more significant than in consumer images because of weather conditions, geographic changes, and camera orientations. We perform progressive domain alignment by creating two intermediate domains, w.r.t. source and target datasets. The proposed method *LGDA* performed exceptionally well with more than 60% mAP for several classes such as *storage*

tank, *harbor*, and *tennis court* in the DOTA and NWPU VHR-10 target data sets. We also use contrastive learning to adapt to local and global domains. Careful selection of the training pipeline, the number of negative samples, the down-sampling strategy, and the temperature value can improve the effectiveness of contrastive learning. Finally, we validate our approach in two challenging high-variability target datasets that showed significant performance gain over available state-of-the-art methods. For the DOTA and NWPU VHR-10 target datasets, we outperformed the latest state-of-the-art *MGADA* method by +7.3% and +4.6% mAP, respectively. In future work, we plan to introduce clustering-based pseudo-labeling, debiased instance-level domain adaptation, and unknown class discovery for satellite images.

ACKNOWLEDGMENT

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] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 770-778 (2016)

[2] 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)

[3] Redmon, J. & Farhadi, A. Yolov3: An incremental improvement. *ArXiv Preprint ArXiv:1804.02767*. (2018)

[4] Jui, T. & Rivas, P. Fairness issues, current approaches, and challenges in machine learning models. *International Journal Of Machine Learning And Cybernetics*. pp. 1-31 (2024)

[5] 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)

[6] 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)

[7] 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)

[8] Liu, J., Xiang, J., Jin, Y., Liu, R., Yan, J. & Wang, L. Boost Precision Agriculture with Unmanned Aerial Vehicle Remote Sensing and Edge Intelligence: A Survey. *Remote Sensing*. **13**, 4387 (2021)

[9] Cheng, G., Li, Q., Wang, G., Xie, X., Min, L. & Han, J. SFRNet: Fine-grained oriented object recognition via separate feature refinement. *IEEE Transactions On Geoscience And Remote Sensing*. (2023)

[10] Yao, Y., Cheng, G., Wang, G., Li, S., Zhou, P., Xie, X. & Han, J. On improving bounding box representations for oriented object detection. *IEEE Transactions On Geoscience And Remote Sensing*. **61** pp. 1-11 (2022)

[11] Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. Microsoft coco: Common objects in context. *European Conference On Computer Vision*. pp. 740-755 (2014)

[12] 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)

[13] Long, X., Deng, K., Wang, G., Zhang, Y., Dang, Q., Gao, Y., Shen, H., Ren, J., Han, S., Ding, E. & Others PP-YOLO: An effective and efficient implementation of object detector. *ArXiv Preprint ArXiv:2007.12099*. (2020)

[14] Mittal, P., Singh, R. & Sharma, A. Deep learning-based object detection in low-altitude UAV datasets: A survey. *Image And Vision Computing*. **104** pp. 104046 (2020)

[15] 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)

[16] Srivastava, S., Narayan, S. & Mittal, S. A survey of deep learning techniques for vehicle detection from UAV images. *Journal Of Systems Architecture*. pp. 102152 (2021)

[17] 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)

[18] 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)

[19] Zhu, P., Wen, L., Bian, X., Ling, H. & Hu, Q. Vision meets drones: A challenge. *ArXiv Preprint ArXiv:1804.07437*. (2018)

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

[21] Zhu, Y., Sun, X., Diao, W., Li, H. & Fu, K. RFA-Net: Reconstructed Feature Alignment Network for Domain Adaptation Object Detection in Remote Sensing Imagery. *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing*. **15** pp. 5689-5703 (2022)

[22] 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)

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

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

[25] Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 779-788 (2016)

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

[27] Liu, Y., Sun, P., Wergeles, N. & Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems With Applications*. **172** pp. 114602 (2021)

[28] Cheng, G., Wang, J., Li, K., Xie, X., Lang, C., Yao, Y. & Han, J. Anchor-free oriented proposal generator for object detection. *IEEE Transactions On Geoscience And Remote Sensing*. **60** pp. 1-11 (2022)

[29] Jiang, D., Li, G., Tan, C., Huang, L., Sun, Y. & Kong, J. Semantic segmentation for multiscale target based on object recognition using the improved Faster-RCNN model. *Future Generation Computer Systems*. **123** pp. 94-104 (2021)

[30] Gupta, A., Dollar, P. & Girshick, R. Lvis: A dataset for large vocabulary instance segmentation. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 5356-5364 (2019)

[31] 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)

[32] He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 2961-2969 (2017)

[33] Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X. & Han, J. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. (2023)

[34] Zhu, L., Lee, F., Cai, J., Yu, H. & Chen, Q. An improved feature pyramid network for object detection. *Neurocomputing*. **483** pp. 127-139 (2022)

[35] Liu, Z., Song, X., Feng, Z., Xu, T., Wu, X. & Kittler, J. Global Context-Aware Feature Extraction and Visible Feature Enhancement for Occlusion-Invariant Pedestrian Detection in Crowded Scenes. *Neural Processing Letters*. pp. 1-15 (2022)

[36] 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)

[37] 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)

[38] Yang, C., Huang, Z. & Wang, N. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 13668-13677 (2022)

[39] Hsu, H., Yao, C., Tsai, Y., Hung, W., Tseng, H., Singh, M. & Yang, M. Progressive domain adaptation for object detection. *Proceedings Of The IEEE/CVF Winter Conference On Applications Of Computer Vision*. pp. 749-757 (2020)

[40] Saito, K., Ushiku, Y., Harada, T. & Saenko, K. Strong-weak distribution alignment for adaptive object detection. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 6956-6965 (2019)

[41] 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)

[42] Song, Y., Liu, Z., Tang, R., Duan, G. & Tan, J. Cross-domain object detection by local to global object-aware feature alignment. *Neural Computing And Applications*. pp. 1-14 (2023)

[43] 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)

[44] Long, M., Cao, Y., Wang, J. & Jordan, M. Learning transferable features with deep adaptation networks. *International Conference On Machine Learning*. pp. 97-105 (2015)

[45] Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T. & Saminger-Platz, S. Central moment discrepancy (cmd) for domain-invariant representation learning. *ArXiv Preprint ArXiv:1702.08811*. (2017)

[46] Hoffman, J., Tzeng, E., Park, T., Zhu, J., Isola, P., Saenko, K., Efros, A. & Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. *International Conference On Machine Learning*. pp. 1989-1998 (2018)

[47] Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T. & Huang, J. Progressive feature alignment for unsupervised domain adaptation. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 627-636 (2019)

[48] 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)

[49] Biswas, D. & Tešić, J. Unsupervised domain adaptation with debiased contrastive learning and support-set guided pseudo labeling for remote sensing images. *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing*. (2024)

[50] Chen, Y., Liu, Z., Zhang, B., Fok, W., Qi, X. & Wu, Y. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **37**, 387-395 (2023)

[51] Li, D., Li, J., Li, H., Niebles, J. & Hoi, S. Align and prompt: Video-and-language pre-training with entity prompts. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 4953-4963 (2022)

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

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

[54] Kang, G., Jiang, L., Wei, Y., Yang, Y. & Hauptmann, A. Contrastive adaptation network for single-and multi-source domain adaptation. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **44**, 1793-1804 (2020)

[55] Kalantidis, Y., Sariyildiz, M., Pion, N., Weinzaepfel, P. & Larlus, D. Hard negative mixing for contrastive learning. *Advances In Neural Information Processing Systems*. **33** pp. 21798-21809 (2020)

[56] Tong, K., Wu, Y. & Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image And Vision Computing*. **97** pp. 103910 (2020)

[57] 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)

[58] 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)

[59] Wang, Y., Bashir, S., Khan, M., Ullah, Q., Wang, R., Song, Y., Guo, Z. & Niu, Y. Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Expert Systems With Applications*. pp. 116793 (2022)

[60] Isola, P., Zhu, J., Zhou, T. & Efros, A. Image-to-Image Translation with Conditional Adversarial Networks. *Computer Vision And Pattern Recognition (CVPR), 2017 IEEE Conference On*. (2017)

[61] 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)

[62] Cheng, G. & Han, J. A survey on object detection in optical remote sensing images. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **117** pp. 11-28 (2016)

[63] Cramer, M. The DGPF-test on digital airborne camera evaluation overview and test design. *Photogrammetrie-Fernerkundung-Geoinformation*. pp. 73-82 (2010)

[64] Hsu, C., Tsai, Y., Lin, Y. & Yang, M. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. pp. 733-748 (2020)

[65] 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)

[66] 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)

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

[68] Li, J., Wei, Y., Liang, X., Dong, J., Xu, T., Feng, J. & Yan, S. Attentive contexts for object detection. *IEEE Transactions On Multimedia*. **19**, 944-954 (2016)

Debojyoti Biswas is a Computer Science Ph.D. student at Texas State University, US. He received his B.Sc. in information and communication engineering from the Noakhali Science and Technology University, Bangladesh. He worked as a Lecturer in the Computer Science department at Leading University, Bangladesh, from 2019 to 2021 before his Ph.D. studies. He is a peer reviewer of several prestigious journals such as IEEE TGRS, IEEE JSTARS, IEEE GRSL, Springer PFGE, Elsevier ESA, etc. His research interests include computer vision, image processing, deep learning, remote sensing object detection, and domain adaptation.

Jelena Tešić, Ph.D. is an Assistant Professor at Texas State University. Prior to 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 July-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.