
IDENTIFICATION OF SMALL OBJECTS IN SATELLITE IMAGE BENCHMARKS

Debojyoti Biswas
Computer Science
Texas State University
San Marcos, TX 78666
ubq3@txstate.edu

Jelena Tešić
Computer Science
Texas State University
San Marcos, TX 78666
jtesic@txstate.edu

September 6, 2022

ABSTRACT

Recent increases in aerial image access and volume, increases in computational power, and interest in applications have opened the door to scaling up object detection and domain adaptation research to production. Aerial data sets are very large in size, and each frame of the data set contains a huge number of dense and small objects. Deep learning applications for aerial imagery are behind due to a lack of training data, and researchers have recently turned to domain adaptation (DA) from a labeled data set to an unlabeled data set to alleviate the issue. These factors create two major challenges: the high variety between datasets (e.g. object sizes, class distributions, object feature uniformity, image acquisition, distance, weather conditions), and the size of objects in satellite imagery and subsequent failure of state-of-the-art to capture small objects, local features, and region proposals for densely overlapped objects in satellite image. In this paper, we propose two solutions to these problems: a domain discriminator to better align the local feature space between domains; and a novel pipeline that improves the back-end by spatial pyramid pooling, cross-stage partial network, region proposal network via heatmap-based region proposals, and object localization and identification through a novel image difficulty score that adapts the overall focal loss measure based on the image difficulty. Our proposed model outperformed the state-of-the-art method by 7.4%.

Keywords Object Detection · Small Objects · Domain Adaptation · Aerial Image Data · Deep Learning

1 Introduction

The next frontier in precision agriculture, emergency rescue system, terrestrial and naval traffic monitoring, and industrial surveillance is the integration of automated, reliable object localization from overhead satellite and aerial imagery [1, 2, 3, 4]. Continuous improvements in deep neural network (DNN) models, combined with increased access to computational resources, have enabled the improvement of object detection methods in both aerial and consumer images [5, 6, 7, 8, 9, 10, 11]. For DNNs to make reliable localization, the networks require a large and diverse number of training data. There are only a handful of reliably annotated datasets for overhead imagery, e.g: DOTA2.0 [12], DIOR [13], Visdrone [14], UAVDT[15]. Therefore, collecting and annotating overhead training data that will provide performance comparable to consumer data proves to be a very complex task.

All object detection approaches developed for consumer images in the past decade assume that the number of objects in an image is in the single digits, and the object size is greater than 1% of the image size. The two images on the left in Figure 1 are typical consumer images. Recent work on domain adaptation (DA) for object recognition from consumer images successfully addresses weather, lighting conditions, geological variance, variation in image quality, and cross-camera adaptation by aligning global feature distribution of data from the origin and target domains [17, 18, 19]. However, satellite images are taken from high altitudes, the number of objects per image is usually in the triple digits, and the size of the object is often less than 0.01% of the size of an image. Examples of parts of satellite imagery (comparable in pixel size with consumer imagery) are illustrated by two right images in Figure 1. Typical



Figure 1: left: consumer images in MSCOCO[16] set; right: aerial images in the DIOR[13] set.

object detection approaches developed for consumer images fail on the satellite imagery due to visual variation within the image, variation among the images in the collection, the relative object size w.r.t., the image size, and the density and number of the objects.

Visual variation Visual variation in satellite imagery is high. High latitude allows for a wider field of view, but the capture image has a higher range of spectral feature variations within it. Higher altitudes result in images within the same collection that vary greatly in global spectral features due to changing lighting, weather conditions, and changing geographic areas captured, as illustrated in Figure 7. In this paper, we propose to use object-level domain adaptation (DA) between any training and test data set (even if it is in the same collection) and align the pixel level features between the training and test data sets. We also introduce robust domain-invariant features that enable us to efficiently automate object localization in previously unseen aerial data sets as illustrated in Figure 10.

Object size and density The size and density of objects in the satellite images differ from the consumer images. The size of the satellite image is up to 400 million pixels, and object sizes are often less than 100 pixels. A typical satellite image patch is 1024×1024 or 1.05 million pixels. If an object is 10×10 or 100 pixels, the size of the object is < 0.0001 of the area of the image. Small objects also tend to be densely packed, as illustrated in Figure 7. The success of object detection depends on how reliable the pixel- and object- feature extraction and region proposal network in the DNN architecture is. The increased number of very small densely packed objects in the image increases the chance of losing pixel level feature information during the feature extraction phase. The RPN-based proposal network misses a large number of small objects in the early stage of the processing pipeline and cannot be recovered in the detection stage [20], as illustrated in Figure 7. In this paper, we have developed an object detection network tuned to satellite imagery. We address visual variations in imagery using object-level domain adaptation (DA) of the target data set (test) to align the characteristics of the object with the source data set (training). We address the localization of small objects in multiple parts of the training and testing pipeline. First, we design a strong darknet-style [21] backbone based on spatial pyramid pooling (SPP) and the partial cross-stage network (CSP) [22], followed by a heatmap-based region proposal generator (RPG) to address the challenge of small dense objects in satellite imagery. Next, the heatmap-based proposal box regression completely eliminates the costly nonmaximal suppression step [23] in the RPN post-processing.

Section 2 summarizes related work, and Section 3 introduces the proposed methodology and improvements to the object detection training pipeline. In Section 4 we present the experimental results on real satellite image sets and show that the proposed approach outperforms the available state-of-the-art models in terms of detecting overlapped and small objects, and producing higher mAP across all classes, and we present future work in Section 5.

2 Related Work

Domain adaptation techniques are used to handle the problem of domain change between source and target data sets. In the last couple of years, the Generative Adversarial Network (GAN) has played a very critical role in solving the domain shift problem. Multitask generative adversarial network (MTGAN) upsamples blur small images into clear fine-scale images with the multitask GAN module and improves object detection [24]. The GAN-based approach expands object detection in consumer images to other domains [18, 19]. Adversarial learning has been applied to align the distribution of characteristics between domains and perform progressive domain adaptation to address the problem of large domain gaps [18]. Edge-enhanced superresolution GAN (EESRGAN) improves the detection of small objects using a super-resolution GAN, an edge-enhanced network (EEN) and a detection network [25]. However, previous research has focused on consumer image data sets. Satellite images differ from these because they introduce domain

shifts due to lighting variations, geographic views and weather. Currently, there is no benchmark on domain adaptation for satellite image datasets for object detection task.

State-of-the-art (SOTA) object detectors are single-stage or multistage detectors. Single stage detectors are object detection networks without an RPN module [26, 10, 21, 27] and are mainly based on different scale and aspect ratios of the anchor boxes. Anchor boxes are chosen k -means cluster centroids. Each anchor box is matched to each cell of the feature matrix extracted from the backbone and the feature pyramid network (FPN) [28]. It was shown that the single stage object detection architectures miss a significant amount of small objects in satellite imagery, and they also require a good anchor design for better performance [12, 29]. Although single-stage detectors were able to achieve state-of-the-art results for consumer images, the region proposals from single-stage detectors are over-dominated by the negative examples, which makes the detection layers biased toward a False Positive result during the training. Multistage detectors are often more reliable [30, 31] due to extra effort to improve the images' regions of interest. Aerial images are very high in resolution and often contain a large number of negative/background objects in the image. Multistage detectors use RPN [11, 32] to filter out positive instances from the image with the help of IOU and the non-maximum suppression technique (NMS). With multistage detectors, we have to train an additional network for region generations, though there have been several works for a lighter version of object detection which does not require the RPN network. However, a very recent work uses a heatmap-based region proposal with balanced positive/negative examples per batch [23]. Heatmap-based RPN uses a Gaussian filter which creates a heatmap peak at the center of the object to define proposal regions. The center is used here as an object anchor, and is based solely on location, not box overlap [8]. Therefore, we have only one anchor per object, which eliminates the heavy usage of NMS to filter overlapped proposals, without affecting the quality of the proposal.

As the size of the object decreases, the chances of losing local information in deep layers increase significantly. The first step introduced towards small detection was the use of FPN [28]. Instead of relying on a single scale feature, it was proposed to use different scale features from different stages of the backbone for different scale prediction. Although this method worked well for many large-scale datasets, the performance of the FPN network still depends on a strong backbone network. The backbone network produces bottom-up features from the input image, and the FPN layer upscales the bottom-up features with the combination of lateral connections to create top-bottom features for scale prediction. FPN also helps to strengthen low spatially rich features by combining semantically rich features. Subsequently, several other variants [33, 34] of FPN method were formed for improved performance. The pixel-level appearance features do not contain enough information to localize small objects in an image, and recent research shows that the context-based bidirectional feature fusion of the neighboring pixels helps to localize small objects in an image [35].

The approach has demonstrated a clear advantage in detecting objects with high variance in the background and objects that are loosely packed. In our dataset, objects are very highly packed and surrounding information will add extra noise as input as well as incur more post-processing operations.

3 Methodology

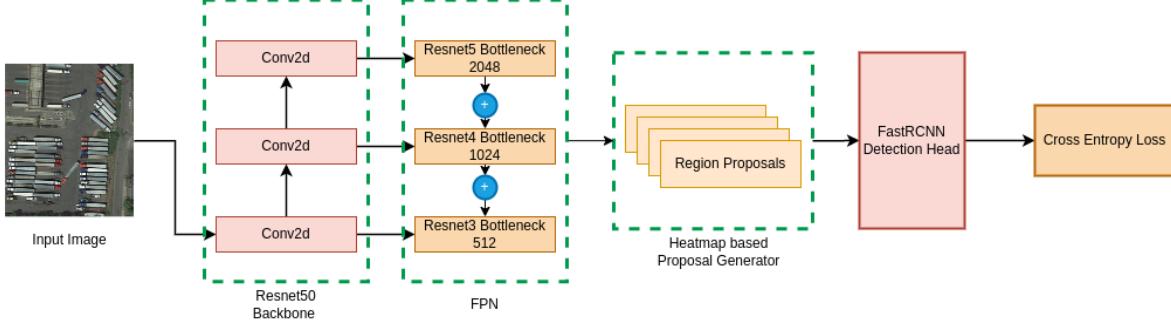


Figure 2: Heatmap based multi-stage Base object detection model.

The objective of our methodology is to design an object detection pipeline for satellite imagery where the domain shift from source (training) to target (test) is not negligible with the goal of (a) better aligning pixel-level features in the training and test sets; (b) create reliable region proposals for satellite imagery; and (c) detect small objects.

First, the *Heat-D* architecture improves pixel level and object feature extraction and introduces the difficulty estimator block that informs a custom loss measure for more robust object detection in satellite images. Second, the *Heat-DDA* architecture improves the detection of objects in satellite datasets as they differ in image capture, i.e. altitude, geographic locations, and weather conditions.

3.1 Base Model

The *base* model used in this paper was recently introduced in [23], and is illustrated in Figure 2. We have used this model as our reference model and adjusted *image load size*, *number of output channels per CNN block* and *IOU in FastRCNN Detection Head* parameters for satellite data set. The base model has three different parts: Backbone, Region Proposal Network (RPN), and Detection Head.

Backbone Backbone combines ResNet50 [23] as feature extractors and ResNet101 [36] as the feature pyramid network (FPN). The residual connection efficiently combines features from previous layers with skip connections [36]. The residual block architecture allows for a deeper model without the vanishing-gradient effect. ResNet50 [23] achieves state-of-the-art performance in the COCO [16] and LVIS [31] data sets. The FPN layer extracts three different scales of features from different layers of the backbone network, and the strides applied on these three scales are 8, 16, and 32, respectively, as illustrated in Figure 2. Also in Figure 2, the Resnet3, Resnet4, and Resnet5 blocks represent strides of 8, 16, and 32 respectively.

Region Proposal Network (RPN) RPN integrates the heatmap-based region proposal network [8] in the *Base* model object detection pipeline. ResNet Style RPN degrades performance in satellite imagery because there is a large amount of overlap between objects, and there is a significant number of small objects per image [10]. Also, the number of non-maximal suppression computations in every pair of proposals per image is 1000+ times higher for satellite imagery than for consumer imagery. The more efficient way is to use probabilistic region proposals. Probabilistic region proposals are calculated using Gaussian kernels on the different scales of features outputted by the backbone element[8]. The element-wise comparison between the max-pool input and the output of the Gaussian kernel produces a heatmap. The max-pool operation will elevate each pixel in the feature, except for the local maxima, where the value is 1. Each peak in the heatmap corresponds to a keypoint of the object (center). Then, the image features at each keypoint are used to predict the height and width of each object, and the resulting regressed bounding boxes are shown to perform well when objects are close to each other and overlap. As a computational bonus, the proposed RPN enables the detection forward-pass, and we can skip the non-maximal suppression step in our pipeline. We have adjusted *image augmentation size*, *number of proposals per image*, and *heatmap minimum overlap per object* in this block to satellite imagery.

Detection Head Detection head is adapted from the Faster-RCNN detector[11]. The detection head takes as input the filtered region proposals from the RPN module. The RPN provides different scale proposals at different steps, as discussed in the backbone module. The first task in the detection head is to convert each proposal into 7×7 pixel size grids with the same number of channels using the region-of-interest (ROI) pooler. Then, the ROI pooler output is flattened and fed into the fully connected network (FCN) layers. The final stage outputs (N, C) class predictor for C classes and N region proposals, and $(N, 4)$ bounding boxes [11]. We have also adjusted the *number of detection per image* and *IOU at detection block* parameters for the satellite imagery.

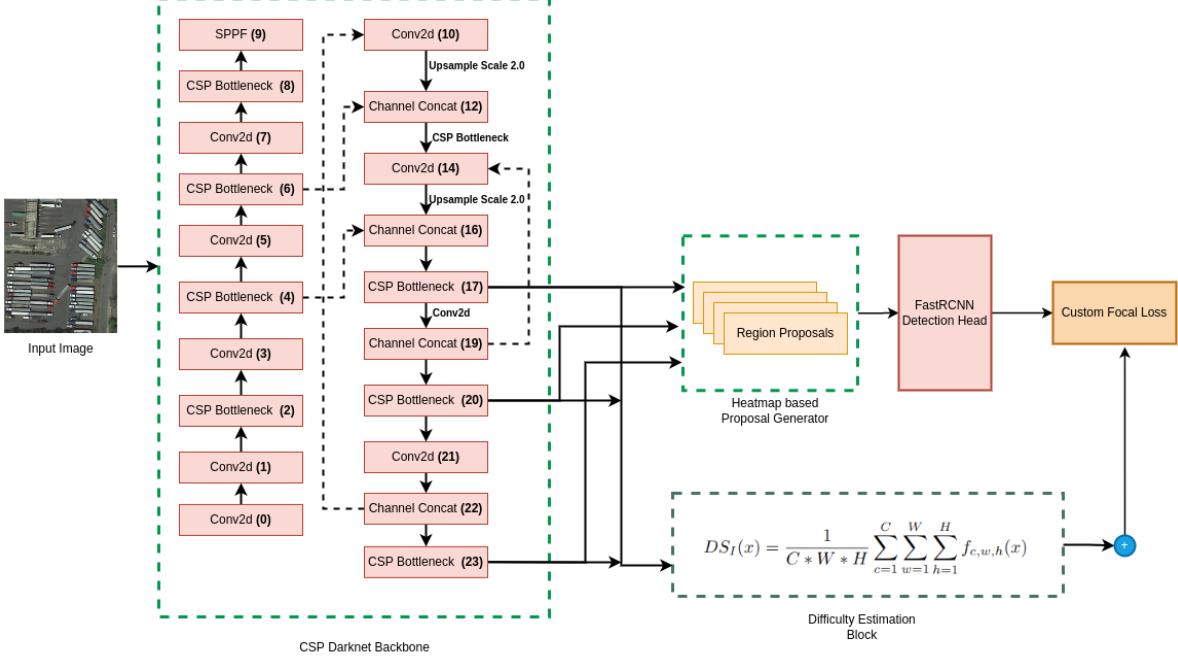


Figure 3: Heat-D object detection pipeline with new difficulty estimator block that informs the new custom focal loss.

3.2 Heat-D Model

We propose *Heat-D Model* as an improvement over *Base Model* by adding the CSP Darknet backbone, adding the *Difficulty Estimation* block, and changing the cross-entropy loss with a modified version of focal loss as illustrated in Figure 3. The *Difficulty Estimation* block calculates the overall difficulty of each image, similar to [37].

Backbone The efficiency of the *RPN* module depends on the effectiveness of the *Backbone* module. 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. The detection head will also fail to identify the object. In *Base Model*, ResNet50 and ResNet101 feature extractors have not been able to extract meaningful features from satellite images that contain many small objects, as illustrated in 11. The deeper layers of CNN architectures, such as ResNet50 and ResNet101, use a larger number of steps. This approach increases semantic information, but loses spatial information in the feature extraction module. On the other hand, the darknet backbone uses the partial split network of stages to preserve better semantic information in the deeper layers of CNN [21]. First, we propose to integrate the partial cross-stage (*CSP darknet*) [22] as a new backbone, since it offers aggregation layers at low and high resolution. In the next step, we propose replacing the max-pooling layer with the spatial pyramid-pooling layer for finer feature extraction. The final configuration proposed for the detection of small objects in satellite imagery is illustrated in Figure 3: the concatenation of layers between layers 6 and 12, layer 4 and 16, layer 14 and 19 and layer 10 and 22 propagates the information from the lower level to the higher level and improves the quality of the characteristics during the extraction of characteristics. The RPN module takes features from Layers 17, 20, and 23 for proposal generation, as illustrated in Figure 3.

Our findings show that 75% of the proposals from RPN modules in *Base Model* tend to be trivial and frequent objects, such as backgrounds and partial object patches. Since the objective is to detect and identify small and dense objects in satellite imagery, we introduce two new blocks in the pipeline: *Custom Focal Loss* and *Difficulty Estimator Block*, as illustrated in Figure 3.

Difficulty Estimator (DE) Domination of trivial class labels overwhelms the classifier with trivial class bias. The increased number of trivial examples in the region proposals incurs a small amount of loss for every example, which in turn contributes significantly while using cross-entropy as a loss function. The difficulty score (DS) for an FPN level with a resolution of $C \times W \times H$ for the image I is calculated in Eq. 1.

$$DS(I) = \frac{1}{C * W * H} \sum_{c=1}^C \sum_{w=1}^W \sum_{h=1}^H f_{c,w,h}(I) \quad (1)$$

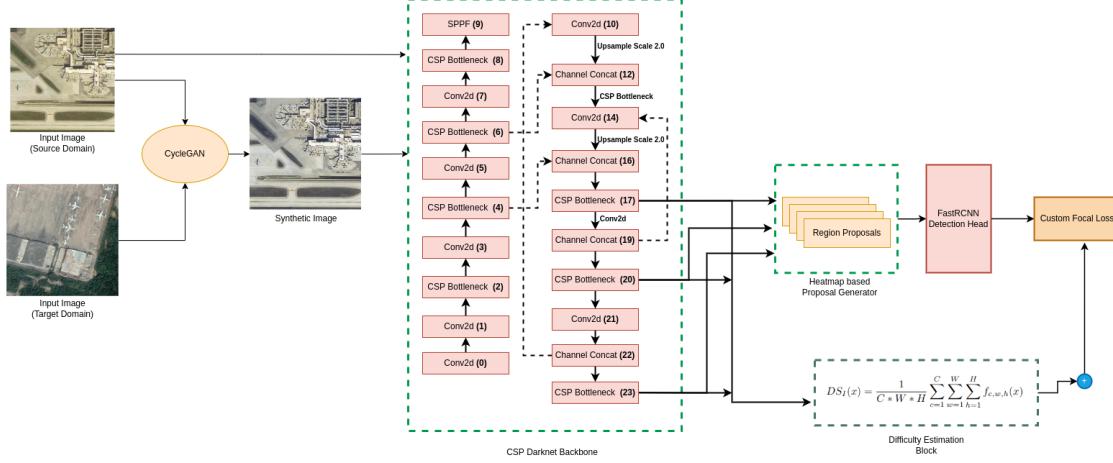


Figure 4: Heat-DDA model with progressive domain adaptation pre-processing block.

Here, C , W , H are feature output channels, feature width and feature height respectively at any FPN level. And, $f_{c,w,h}(I)$ denotes the value from Sigmoid Linear Unit (SiLU) at every pixels in the image I . Using this block we calculate the number of total neurons fired for a single image in the forward pass. We sum up all activation values and divide by the total feature dimension C , W , H to get the Difficulty Score (DS) at a FPN level. We derive this DS from 3 different FPN level and average the values to get the final DS for an image I . The increase in complexity for this DS block is nearly negligible. The Big Oh (O) notation for this block is $O(r)$, where r is the size of the batch in each iteration.

Custom Focal Loss Custom 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 focalization, as illustrated in Figure 4. This adjustment reduces the impact of the class labeling imbalance and the trivial/difficult object identification span for objects in the satellite imagery data set. In the proposed focal loss function, we use difficulty scores calculated for each image by a difficulty estimator block as a weight factor to focus more on complex images with a high diversity of objects and a high variation in pixel-level features. The basic form of the focal loss function is $FL(p_t, y) = \alpha_t * (1 - p_t)^\gamma * CE(p, y)$, where 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 traditional cross-entropy loss function. We propose a new measure, Difficulty Weighted Focal Loss (DWFL) and define it as a product of difficulty score and focal loss for the image $DWFL(x, p, y) = DS(I) * FL(p, y)$, where $DS(I)$ is defined in Eq. 1. The value α is used in the DWFL calculation to control the class imbalance problem in our source and target datasets. Here, the value of the parameter α increases if the frequency of a particular class is very low and decreases if the frequency of a particular class is very high. In this way, we focus more on minor classes. The α_c is calculated as in Eq. 2 for each class, where the modulating factor α'_c depends on the frequency $|C_c|$ of a particular class in the data set and $|C_1 \cup C_2 \cup C_3 \dots|$ is the total number of all instances of all classes in the data set.

$$\forall c \in C, \alpha'_c = -1 * \log \left(\frac{|C_c|}{|C_1 \cup C_2 \cup \dots|} \right) \Rightarrow \alpha_c = \beta * \frac{\alpha'_c - \min(\alpha_c)}{\max(\alpha_c) - \min(\alpha_c)} \quad (2)$$

We use these normalized α_c values from Eq. 2 across different classes $c, c \in C$ to mitigate the imbalance of class labeling. The scaling factor $\beta = 0.6$ was found to be the most appropriate for the satellite imagery. The range of α'_c values in the DIOR data set is 0.2 to 0.79 and the range of α'_c values in the DOTA data set is 0.15 to 0.96, which represents a very tight scaling factor for FL in both data sets. The proposed normalization of α_c in Eq. 2 is more effective and gives a stable loss calculation for a highly unbalanced class count in the data set.

3.3 Heat-DDA Model

In the second phase of our research, we focus on domain adaptation for the target dataset DOTA2.0. It was assumed here that there is no ground truth annotation for the dataset, and we have access to only the source data set DIOR. The DOTA2.0 and DIOR images were taken from very different geographical locations, and different types of satellite were used to capture images with different orientation. Moreover, the combination of object classes, weather variation, and lighting conditions contributes to a large domain gap between our source and target data set. Performing domain

adaptation in such a large domain gap often provides an unstable result and performs poorly for object detection task. To reduce the domain gap between source and target domains, we have created an intermediate domain which is closer to the target domain. Our research found that applying this type of progressive domain adaptation performs better in most cases. Others [18] had used the idea of progressive domain adaptation in a different way and were able to get a favorable result in the target domain. CycleGAN [38, 39] was used to perform discriminatory domain adaptation between experimental domains. The training images for the source and target domains were fed into the CycleGAN (see Figure 4) domain discriminator, and the goal of the training was to train the model so that the domain discriminator could not differentiate the images for the source and target domains. In other words, the goal was to maximize the loss of the domain discriminator. The conversion of CycleGAN allows us to align pixel-level features (e.g. color and texture) between two domains. Finally, the Heat-DDA model Figure 4 was trained using both the translated domain images and the original source training images, all illustrated in Figure 9. The conversion of intermediate domain helped us to train the final model with target-like domain images without using ground-truth annotations.

4 Experiments

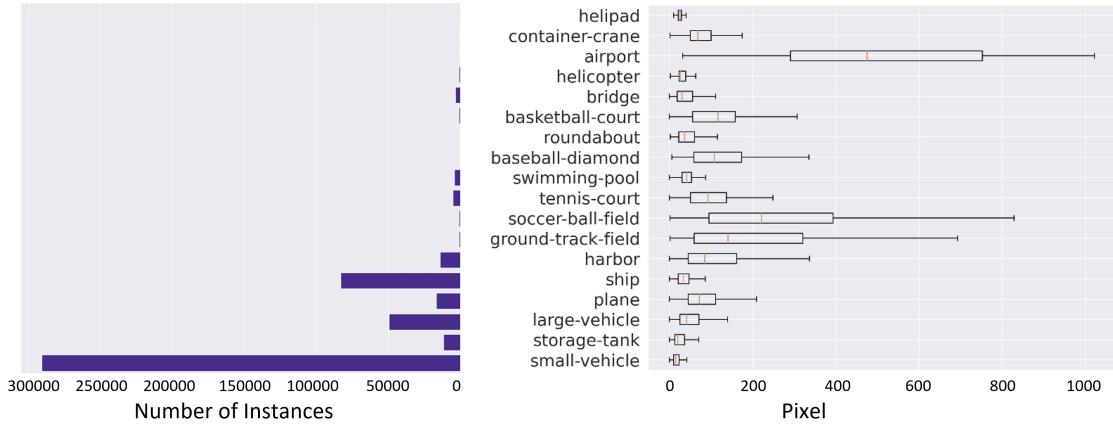


Figure 5: Class instances and object pixel size distribution in the dataset for the DOTA data set.

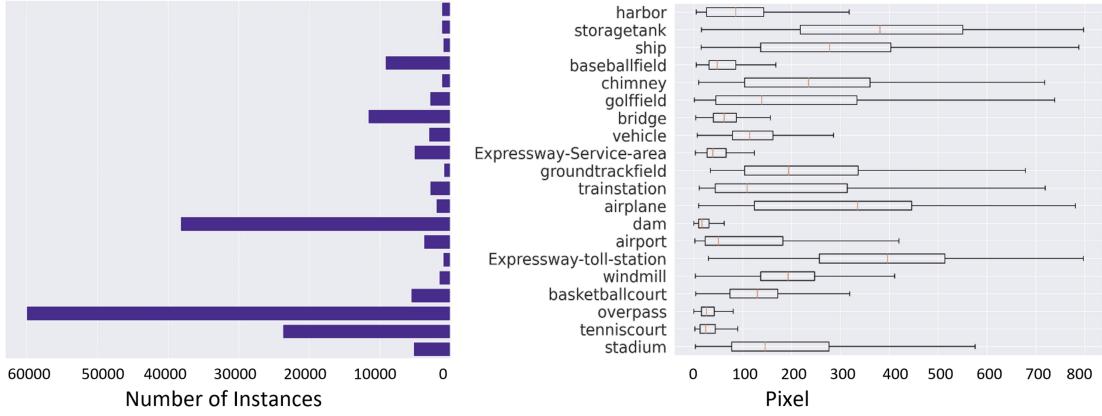


Figure 6: Class instances and object pixel size distribution in the dataset for the DIOR data set

Datasets DIOR dataset consists of 23,463 Google Earth images of areas in 80 countries. The quality of the images varies, and the content was captured during multiple seasons and multiple weather conditions. The data set covers a wide range of spatial resolutions, object size, and object orientation variability, and a diverse class distribution as illustrated in Fig. 6. The spatial resolution of the images is in the range $[0.5m, 30m]$, and the size of the images in the data set is 800×800 pixels. The number of annotated objects in the data set is 192,472, and they are categorized into 20 classes [13]. Labels *ship*, *vehicle*, and *storage* dominate the data set, while labels *expressway toll station*, *train*



Figure 7: *Base* model performance: ground truth examples is on the left-side of the arrow, and *Base* model detection is on the right-side of the arrow and it missed most of the small objects in the dense area.

station, dam, and airport only have few representative examples. The number of images in the training set is 22,450, and in the testing set, we have 1,012 images.

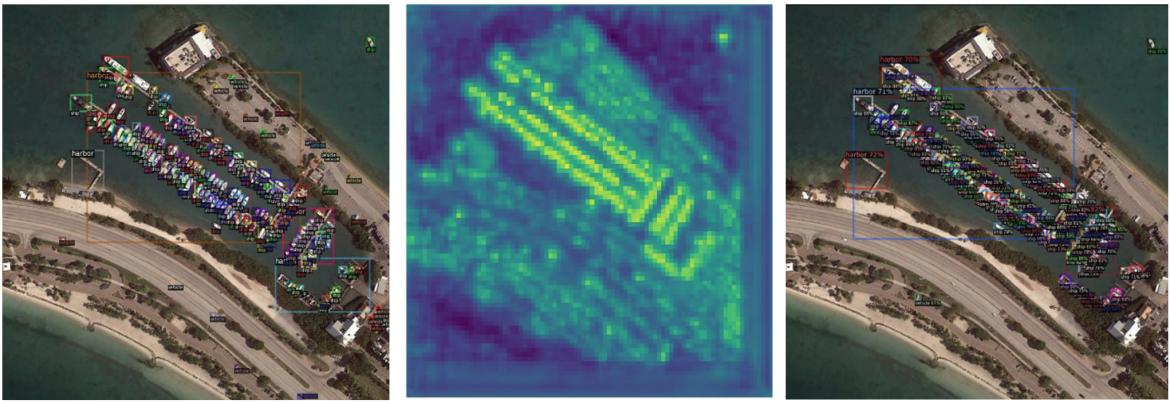


Figure 8: heatmap-based region proposal pipeline results for DIOR data

The **DOTA** data set consists of 2,430 overhead image images collected from Google Earth, the Gaofen-2 (GF-2) satellite, the Jilin-1 (JL-1) satellite and airborne images (taken by CycloMedia in Rotterdam) [12]. It is also known as the DOTA2.0 dataset, but we refer to it as DOTA in this paper for simplicity. The image sizes of Google Earth in the collection range from 800×800 to 4000×4000 pixels, the image size of the GF-2 satellite is $29,200 \times 27,620$ pixels, and the image size of CycloMedia airborne is $7,360 \times 4,912$ pixels. The GSD range in the DOTA dataset is 0.1 to 0.87 m, and the average number of objects per image is 220. The data set contains 1,793,658 annotated objects and is grouped into 18 classes. Most objects have a total size less than 50 pixels, and objects classes *small vehicle*, *ship*, *plane*, and *large vehicle* are densely packed in the images, as illustrated in Figure 5. In the experiment, we split the large images into subimages of size 1024×1024 pixels with an overlap of 200 pixels. The final version of our training set has 12,700 images and the validation set has 4,543 images.

To evaluate and prove the trustworthiness of our domain-adapted *Heat-DDA* model, we have created a new mapped dataset **DOTA2DIOR** (see Figure 1) using the DOTA data set. This data set is used to calculate performance metrics in the source-trained *Heat-DDA* model. From the DOTA dataset, we only kept the classes that are common in the DIOR dataset. Table 1 shows the number of matched classes and the number of objects per class in the mapped data set.

Setup *Base* model, the *Heat-D* model (Figure 3), and the *Heat-DDA* model (Figure 4) were evaluated on three different datasets: DIOR, DOTA2.0, and the target assigned to the source, DOTA2DIOR. After training our *Heat-DDA*, we have evaluated our model for both the source and target domains.

$$P = TP / (TP + FP), \quad R = TP / (TP + FN), \quad F1 = \frac{2 * P * R}{P + R}, \quad (3)$$



Figure 9: Domain translation from DIOR to DOTA using CycleGANN [38]: (a) source domain, (b) translated domain and (c) target domain.

$$AP = \sum_{k=0}^{k=n-1} [R(k) - R(k+1)] * P(k), \quad mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP(k) \quad (4)$$

Performance Measures Precision P , Recall R , $F1$ and mean precision mAP scores are computed in Eq. 3 and Eq. 4. True positives TP are outcomes that the model predicted correctly, false positives FP are outcomes model missed, and true negatives TN are outcomes model erroneously predicted. All of these metrics were calculated on the basis of an IOU of 0.5: 0.95 and the number of proposals per image was set to 256. Precision P measures the fraction of relevant occurrences among recovered instances, and recall R is the fraction of objects that the model correctly identified among all relevant instances. The $F1$ score provides a single measure on how well the model performs when given a class imbalance dataset, and mAP is computed as shown in Eq. 4, 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 (Eq. 4).

class label	vehicle	ship	airplane	harbor	storage tank	tennis court	bridge field	baseball field	track field	basketball court	airport	Total
# obj.	96,783	28,270	6,055	5,711	5,417	1,662	1040	516	417	358	154	146,383

Table 1: DOTA2DIOR dataset.

Figure 8 shows the example performance of the *Heat-DDA* model on a highly dense object image from DIOR. Figure 8(b) shows the strong objectness characteristic of the model using a heatmap proposal. From the proposal heatmap, it is evident that our backbone was able to preserve most of the pixel-wise and object features. As a result the RPN module was able to capture almost every object residing in the image. Finally, Figure 8(c) makes the scale prediction and correctly classifies each object with a high confidence score. On the other hand, Figure 9 shows the effect of pixel-level feature alignment between the source and target domains. Before feature alignment, see Figure 11(a), the center of the source domain was not aligned with the target domain, resulting in a decrease in performance. However, the CycleGAN translated images were able to align Figure 11(b) of the pixel-level features with a uniform center in the feature space. The result of feature alignment is evident in Figure 10: we can compare the performance of the *Base* model Figure 11 with the domain adapted *Heat-DDA* model on the same set of images. The *Base* model is not capable of detecting almost 60% of objects on target images 7, but our proposed *Heat-DDA* model is capable of generating the SOTA detection result for small and medium-sized objects.

Precision of three models in three different data sets are illustrated in Figure 12(a). The average precision of the *Base* model in all 18 classes is 49.7% in the DIOR training set. The average precision of the target-trained *Base* model in the target-only DOTA validation data set is 17.8%. The average precision of the DOTA2DIOR data is also low and shows that the source-trained model *Base* cannot perform well without domain adaptation. *Heat-D* model shows a reasonable improvement with the integration of the improved backbone and the difficulty scoring module in Figure 12 in terms of the precision and recall. Figure 12(a) shows an improvement in AP of 1.2% for DOTA and 1.4% for the DIOR data set. The improvement of the *Heat-D* model is significant for classes with small objects and classes that are



Figure 10: *Heat-DDA* model performance: ground truth examples is on the left-side of the arrow, and *Heat-DDA* is on the right-side of the arrow. *Heat-DDA* model shows improved detection performance when compared to *Base* model detection in Figure 7.

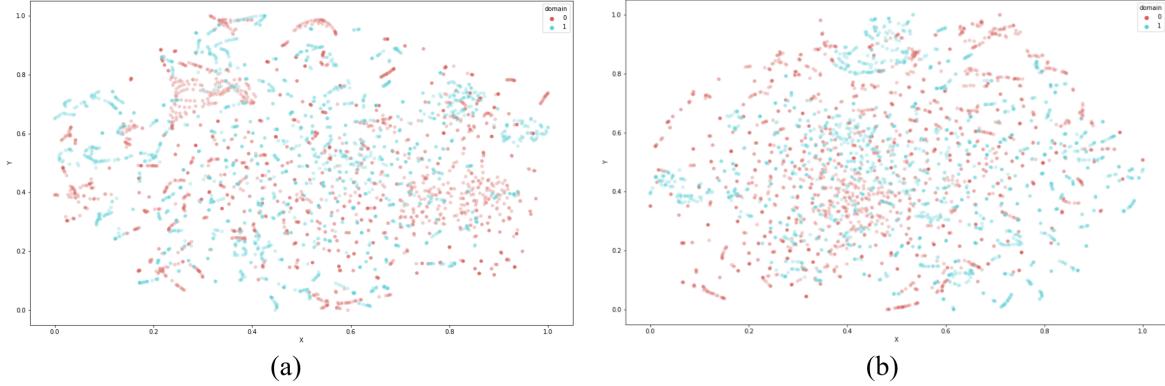


Figure 11: Feature space alignment where *red* dots are the source domain and *cyan* dots are the target domain): (a) 2D feature distribution before alignment; (b) 2D feature distribution after alignment of the same random 20,000 2D feature vectors.

difficult to distinguish. After adaptation of the domain with alignment of pixel-level characteristics, our *Heat-DDA* model shows a 7.4% increase in precision for the DOTA2DIOR data in Figure 12(a). Note that the performance on the source data set did not decrease increased by 1% after adding DA images during training. **Recall** of three models in three different data sets are illustrated in Figure 12(b). The recall with *Base* model for the source and target data sets is respectively 54.3% and 34.3%. The *Heat-DDA* model with domain adaptation achieved a significant improvement in all data sets. The recall measure for the DOTA2DIOR data set was 6.4% higher than the *Base* model (see Figure 12(b))

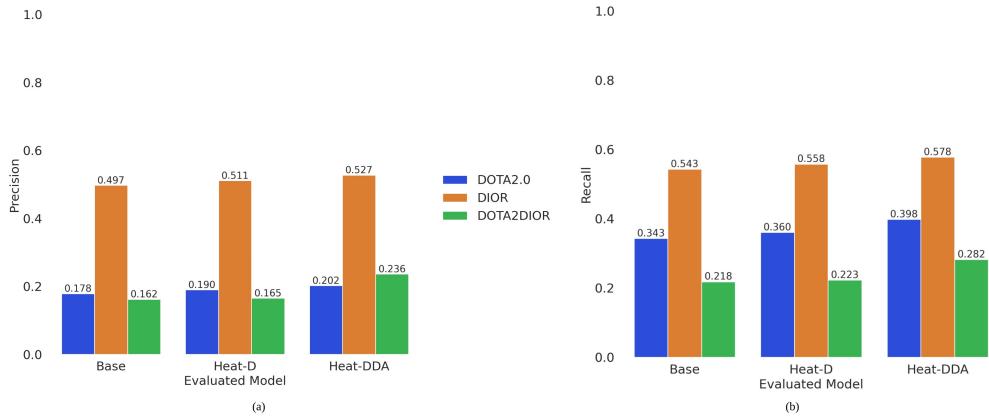


Figure 12: (a) Precision($\text{IOU}=0.50:0.95$) and (b) Recall($\text{IOU}=0.50:0.95$) comparison from different models vs different datasets.

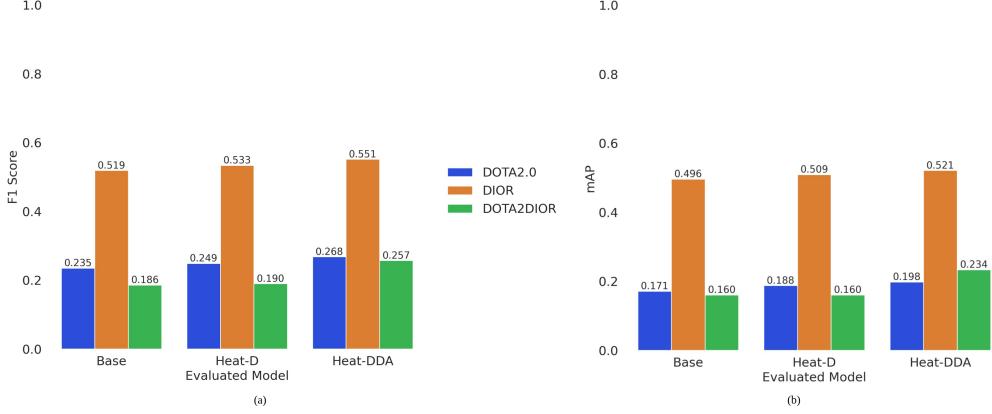


Figure 13: (a) F1 Score(IOU=0.50:0.95) and (b) mAP(IOU=0.50:0.95) comparison from different models vs different Datasets.

and there was an increase of 5.5% and 3.5% in the data set of source and target only, respectively. The **F1 Score** of three models in three different data sets are illustrated in Figure 13(a). The F1 score found in the *Heat-DDA* model for the DOTA2DIOR data set increased by 7.1% compared to the *Base* model. For the other data sets, the improvement of our proposed models was remarkably significant and comparable to other state-of-the-art findings. The **mAP** in the classes for three datasets and models is presented in 13(b). The mAP for source data when trained on source data is 49.6% for the *Base* model and 52.1% for *Heat-DDA* model, a 2.5% increase even after adding translated images for training, as illustrated in 13(b). The mAP for target test data when trained on the target training dataset is 17.1% for the *Base* model and 19.8% for the *Heat-DDA* model, a 2.7% increase 13(b). The best outcome for *Heat-DDA* model was achieved from DOTA2DIOR dataset when trained using *source dataset* and *domain adapted target dataset*, and Figure 13 shows 7.4% increase in mAP compared to the *Base* model.

5 Conclusion and Future Work

Object detection in aerial images is one of the most challenging tasks in computer vision research because there are many small and overlapped objects in the images. The success of DNN object localization depends on the performance of the large number of related objects annotated in the training data and on a reliable feature extractor module in the pipeline. In this paper, we introduce a strong feature extractor that captures balanced low-level and high-level features for small objects. Next, we introduce the heatmap-based region proposal module to better capture small objects. Finally, we introduce two new modules in the satellite object detection pipeline, the difficulty scoring module, which informs the image difficulty to the custom focal loss module and balances the aerial object detector against trivial background classes. The proposed *Heat-DDA* method performed extremely well with more than 60% mAP for several classes such as *ship*, *airplane*, and *storage tank* in the DIOR dataset. We also show how to successfully address satellite imagery scarcity and introduce local feature alignment across labeled source and unlabeled target domain datasets for domain-adapted object detection, where our proposed model outperformed the baseline model in the DIOR2DOTA dataset by 7%, 6.4%, & 7.4% for precision, recall and mAP metrics, respectively. Next, we plan, for instance, global feature alignment at the instance level, extension to satellite data, and unknown class discovery.

References

- [1] João Valente, Bilal Sari, Lammert Kooistra, Henk Kramer, and Sander Mücher. Automated crop plant counting from very high-resolution aerial imagery. *Precision Agriculture*, 21(6):1366–1384, 2020.
- [2] Scott Workman and Nathan Jacobs. Dynamic traffic modeling from overhead imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12315–12324, 2020.
- [3] Hafiz Suliman Munawar, Fahim Ullah, Siddra Qayyum, Sara Imran Khan, and Mohammad Mojtabaei. Uavs in disaster management: Application of integrated aerial imagery and convolutional neural network for flood detection. *Sustainability*, 13(14):7547, 2021.
- [4] Jia Liu, Jianjian Xiang, Yongjun Jin, Renhua Liu, Jining Yan, and Lizhe Wang. Boost precision agriculture with unmanned aerial vehicle remote sensing and edge intelligence: A survey. *Remote Sensing*, 13(21):4387, 2021.

- [5] Payal Mittal, Raman Singh, and Akashdeep Sharma. Deep learning-based object detection in low-altitude uav datasets: A survey. *Image and Vision computing*, 104:104046, 2020.
- [6] Xiang Long, Kaipeng Deng, Guanzhong Wang, Yang Zhang, Qingqing Dang, Yuan Gao, Hui Shen, Jianguo Ren, Shumin Han, Errui Ding, et al. Pp-yolo: An effective and efficient implementation of object detector. *arXiv preprint arXiv:2007.12099*, 2020.
- [7] Srishti Srivastava, Sarthak Narayan, and Sparsh Mittal. A survey of deep learning techniques for vehicle detection from uav images. *Journal of Systems Architecture*, page 102152, 2021.
- [8] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.
- [9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [10] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi 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*, pages 2778–2788, 2021.
- [11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- [12] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.
- [13] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020.
- [14] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018.
- [15] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018.
- [16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [17] Ziwei Deng, Quan Kong, Naoto Akira, and Tomoaki Yoshinaga. Hierarchical contrastive adaptation for cross-domain object detection. *Machine Vision and Applications*, 33(4):1–13, 2022.
- [18] Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 749–757, 2020.
- [19] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019.
- [20] Kang Tong, Yiquan Wu, and Fei Zhou. Recent advances in small object detection based on deep learning: A review. *Image and Vision Computing*, 97:103910, 2020.
- [21] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.
- [22] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau 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*, pages 390–391, 2020.
- [23] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021.
- [24] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 206–221, 2018.
- [25] Jakaria Rabbi, Nilanjan Ray, Matthias Schubert, Subir Chowdhury, and Dennis Chao. Small-object detection in remote sensing images with end-to-end edge-enhanced gan and object detector network. *Remote Sensing*, 12(9):1432, 2020.

- [26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [29] Yang Liu, Peng Sun, Nickolas Wergeles, and Yi Shang. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications*, 172:114602, 2021.
- [30] Du Jiang, Gongfa Li, Chong Tan, Li Huang, Ying Sun, and Jianyi Kong. Semantic segmentation for multiscale target based on object recognition using the improved faster-rcnn model. *Future Generation Computer Systems*, 123:94–104, 2021.
- [31] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.
- [32] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [33] Zuoxin Li and Fuqiang Zhou. Fssd: feature fusion single shot multibox detector. *arXiv preprint arXiv:1712.00960*, 2017.
- [34] Lisha Cui, Rui Ma, Pei Lv, Xiaoheng Jiang, Zhimin Gao, Bing Zhou, and Mingliang Xu. Mdssd: multi-scale deconvolutional single shot detector for small objects. *arXiv preprint arXiv:1805.07009*, 2018.
- [35] Jiaxu Leng, Yihui Ren, Wen Jiang, Xiaoding Sun, and Ye Wang. Realize your surroundings: Exploiting context information for small object detection. *Neurocomputing*, 433:287–299, 2021.
- [36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [37] Peng Sun, Guang Chen, and Yi Shang. Adaptive saliency biased loss for object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10):7154–7165, 2020.
- [38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.
- [39] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.