AFRL-RY-WP-TR-2020-0080

METADATA LEARNING OF NON-VISUAL FEATURES: CO-OCCURRENCE OVERLAP FUNCTION FOR RECTANGULAR REGIONS AND GROUND TRUTH DATA

Asif Mehmood
Decision Sciences Branch
Multi-Domain Sensing Autonomy Division

Vasanth Iyer

New College of Florida

APRIL 2020
Final Report

AIR FORCE RESEARCH LABORATORY
SENSORS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7320
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
April 2020	Final	30 December 2019 –30 December 2019

4. TITLE AND SUBTITLE

METADATA LEARNING OF NON-VISUAL FEATURES: CO-OCCURRENCE OVERLAP FUNCTION FOR RECTANGULAR REGIONS AND GROUND TRUTH DATA

5a. CONTRACT NUMBER
N/A

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
N/A

6. AUTHOR(S)

Asif Mehmood (AFRL/RYAT)
Vasanth Iyer (Troy University)

5d. PROJECT NUMBER
N/A

5e. TASK NUMBER
N/A

5f. WORK UNIT NUMBER
N/A

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Decision Sciences Branch (AFRL/RYAT)
Multi-Domain Sensing Autonomy Division
Air Force Research Laboratory
Sensors Directorate
Wright-Patterson Air Force Base, OH 45433-7320
Air Force Materiel Command
United States Air Force

Troy University

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Sensors Directorate
Wright-Patterson Air Force Base, OH 45433-7320
Air Force Materiel Command
United States Air Force

Air Force Office of Scientific Research Public Affairs (AFOSR)
875 North Randolph Street
Suite 325, Room 3112
Arlington, VA 22203

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
AFRL/RYAT

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RY-WP-TR-2020-0080

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
PAO case number 88ABW-2019-6073, Clearance Date 30 December 2019. This is a work of the U.S. Government and is not subject to copyright protection in the United States. Report contains color.

14. ABSTRACT

We report preliminary results on quality ranking from region segmentation on the aerial ground truth data using standard models trained from ImageNet [10] initialization. The results are comparable to an exhaustive search method [4] when the ground truth bounding boxes are not an overt (when the overlap is greater than 0.5 with the ground truth). Increasing the number of background regions helps augmenting the target object and helps find goodness of fit with the same amount of training data.

15. SUBJECT TERMS
non-visual features, YOLO, RCNN. deep learning, metadata, detection, satellite imagery

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Asif Mehmood
Unclassified	Unclassified	Unclassified	SAR	21	19b. TELEPHONE NUMBER *(Include Area Code)* N/A

AFRL Summer Faculty Fellowship Program
2019

Individual Report

Metadata Learning of Non-Visual Features: Co-Occurrence Overlap Function for Rectangular Regions and Ground Truth Data

Submitted By :
Vasanth Iyer Ph.D.,
Assistant Professor, Troy
University, AL

Contents

1 Introduction

To incorporate object locations in a multi-target detection model, we assume that a close duplicate cannot be learned by the model efficiently. So, we use a region-based approach which uses more object location compared to the ground truth locations to localize the targets. The proposed model is able to learn a similarity metric with respect to the ground truth locations which is robust (low false positives) enough for varying images conditions, small aerial target sizes and using few training samples.

We report preliminary results on quality ranking from region segmentation on the aerial ground truth data using standard models trained from ImageNet [10] initialization. The results are comparable to an exhaustive search method [4] when the ground truth bounding boxes are not an overfit (when the overlap is greater than 0.5 with the ground truth). Increasing the number of background regions helps augmenting the target object and helps find goodness of fit with the same amount of training data.

2 Background and Relevance to Prior Work

Most of the current work has been in CNN deep feature extraction in the context of object detection. Out proposed methods extend the model to learn the meta data which enables accurate localization of observed target regions from expected ground truth efficiently. Aerial imagery as it spans a vast spatial area needs to be divided into tiles manually and labeled. The tile labeling is inherently noisy due to the labeling process and ambiguity of aerial visual features, and scenes may either omit class labels or have incorrect class labels. The term noise is used in the context of metadata [7] even though the images are high-resolution pixels. The context here is aerial images or satellite imagery which has visual ambiguity as the targets are very small or sometimes look similar, e.g., a top-level view of a road can look like a river-flow and can be mislabeled when using visual features. The pre-processing step needs to address working with such label noise in metadata [7] for it to have higher classification accuracy. We have experiment a deep learning network for objectness (foreground or background) of a region for video streams and aerial images using the help of pretrained network. Using pre-trained CNN networks [6] such as Cifar10, AlexNet and ResNet50 have generated highly features which has improved the localization accuracy when learning from small aerial datasets. Most of the literature concentrates on the feature extraction part in this report we look deep into the localization and a metric which can be used for aerial image framework.

3 Statement of Problem

When predicting where a target object is located, we predict potential regions inside the image. These regions are generic terms we consider rectangular regions which are greater than 32×32 pixels. We are also given a set of ground-truth annotations which precisely specifies the object locations for out model to train and validate. In our proposed method we pre-process to extract regions from an input image of 650×480 pixels (which are fused with pre-trained models) we like to further provide an "objectiveness" membership function. The membership function gives high scores the potential rectangles which are foreground objects compared to background ones. In machine learning such as task which does not have explicit labels can learned using regression of the four rectangular box coordinates. Further we distinguish the generic term "regions" compared to ground-truth "bounding-boxes" by the way they are generated. Ground-truth are pre assigned bounding-boxes and regions are predicted by out method as the model expects there would be less likelihood of an exact match. The mapping of regions to bounding-boxes are ranked for the top k regions as we expect to have very high number of regions compared to ground-truth bounding boxes per image. We show our approach and how to design an objective functions where the learning is not only based on visual processing but relies on implicit localization features learned by the model's quality metric. Given a good amount of training data, the multiple-layer deep network will learn spatial (localization) and visual features using our proposed overlap function.

4 General Methodology and Procedure to be Followed

Bottom-Up feature map fusion using pre-trained on ImageNet and finding optimal number of spatial-locations using Jaccard similarity. We define an overlap membership function between regions and bounding boxes such that the abjectness' of the input regions can be ranked.

To calculate the best overlap similarly function can now be defined as

$$ABO = \frac{1}{|G^c|} \sum maxOverlap(g_i^c, l_j^c)$$

Where the overlap function is given by Jaccard similarly as shown.

$$Overlap(g_i^c, l_j = \frac{area(g_i^c \cap area(l_j)}{area(g_i^c \cup area(l_j)}$$

There are two separate loss functions one for the object and other for the localization. Currently only one bounding box can be predicted accurately when detecting multiple objects.

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i L_{reg}(t_i^*, t_i^*)$$

Using the similarity measure we apply this to an example satellite data which has cars as shown below. We train the model with only positive samples. On the left we see the generated regions which match the size of a car learned by the model. The model eliminates all the false-positives and the finds the only car in the training example which is shown on the right.

5 Learning Meta-Data to Rank Objectness

The ground truth car training data is as below with no negative samples: In this example we use a small aerial car detection example in large background. To show how the basic model learn localization we use this proof-of-concept on use the training ground-truth as a mask to detect the small object as shown in the figures Tables 1 and Table 2 below. The region proposal [5] algorithm generates possible objects which are shown as "yellow" rectangles of the left. The model is able to mask out the car and it predicted location.

5.1 Testing Data on Satellite Imaginary

We use the custom learned model on testing images which have multiple-targets and the model does not scale well as it has only one bounding box to predict. In the figure it does find cars but does not count all of them and predicts many false positives as well.

5.2 Re-enforced Learning with False-Positives with Object Centring

The current model is robust but does not do well with real data as we have insufficient number of training samples, we false negatives. Shown below is a representation of our tiled images which are now larger than the ground truth objects. Including background help the model to learn similar regions which are non-objects (false

Cars	
Aircraft	

Table 1: Non-Tiled Ground Truth of XVew2 Dataset.

(a) Potential region proposals

(b) Identified car object

Table 2: Monotone non-centered aerial imagery.

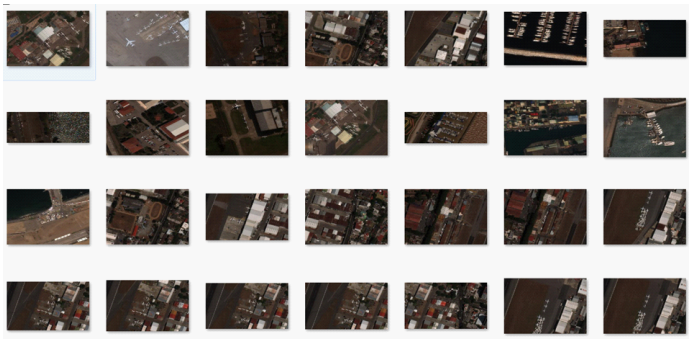


Figure 1: Tiled with centered objects.

positives). Further during tiling of the **XView2** dataset [9] we center the ground truth bounding box so that most or all the target location are at the same locations which are surrounded by false positive samples. The results from the newly generated dataset address the previous issues of false positives as shown in the figures 2 and 3 below. As we use a centering approach multi-target recognition does not work satisfactory with more than two objects and when number of target objects are dense.

5.3 Preliminary Results on Satellite Data-XView2

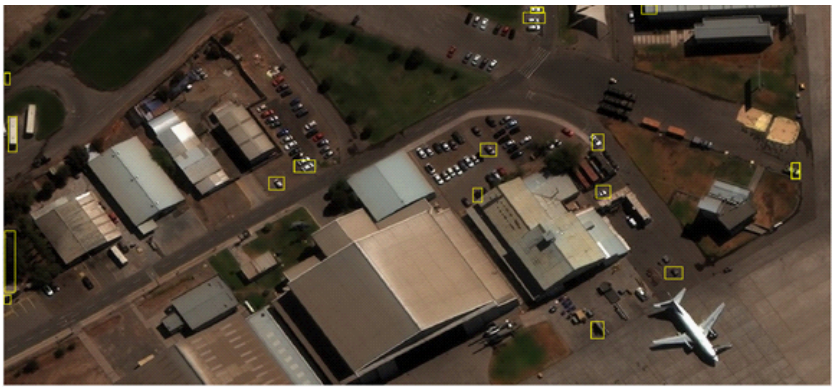


Figure 2: Poor performance with real multi-object test data.

The first analysis of the results with multiple-objects seems to be incomplete from Figure 2. So we constrain our model with fewer target objects and keep them centered. These image augmentations allow the model to predict with high accuracy and avoid false positives. A high precision score of object detection with augmented data is shown in Figure 3. Our proposed model and its loss function have only one

Figure 3: Corrected tiled with a single centered objects from XView2 dataset.

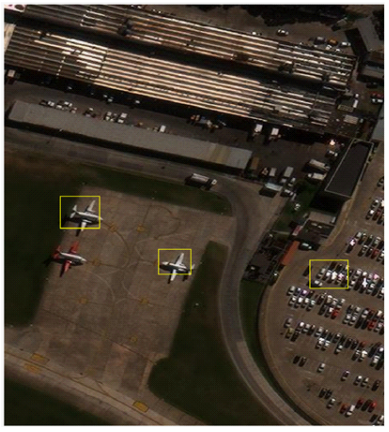


Figure 4: Modified model is not robust to localization over three objects even though the number of objects counted is correct.

detector, and as the objects can be located spatially anywhere, learning the location is a hard problem. Even though the R-CNN model is very precise as we show that it does have false-positives, as shown in Figure 4.

6 Affects of Hyper Parameters and Object Localization

We have shown that negative samples enhance the accuracy and decrease false positives with additional cost in computation. Similarly, we like to explore other model

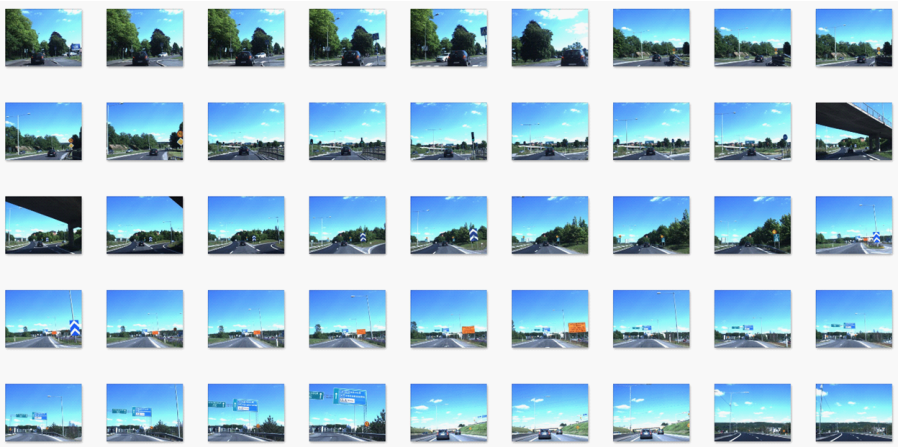


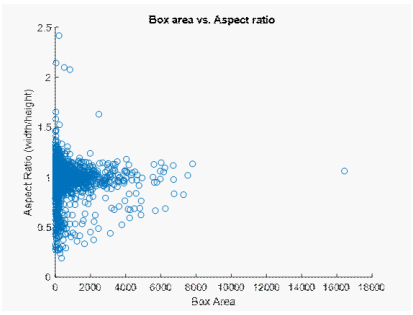
Figure 5: Swedish Sign Dataset.

parameters which affect accuracy. Some of the parameters are the number of anchors, range of object sizes, what is the maximum number of targets in a single image. Due to XView2 dataset is not spatially tiled for an ideal detector we use the Swedish sign dataset [8] which has enough targets with varying sizes. We also validate the results with aerial datasets close XView2 such as Stanford Drone video footages of 60 GB. We define all sets of parameters which can be changed to see the comparison of our model, as shown in Table 6 outline.

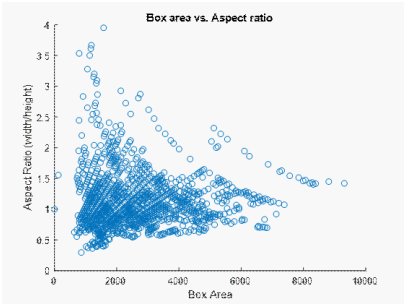
6.1 Intersection of Overlap-IOU

A useful metric such as Jaccard similarity can be applied to localization. Given a set of coordinates used to represent ground truth in a training set. Such that of sets S and T is $|S \in T|/|S \cup T|$, that is, the ratio of the size of the intersection of S and T to the size of their union. We assume that the overlap of the regions is more significant than 0.5. We shall denote the Jaccard similarity of S and T by $SIM(S,T)$.

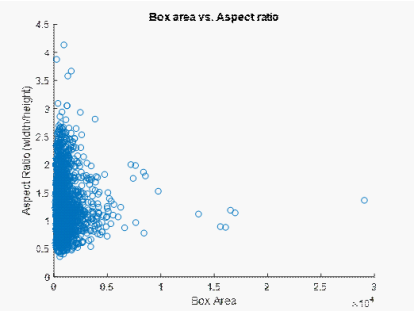
To quantify how a training dataset can be affected, we need to plot the bounding box area compared to their width/height. A plot of the bounding box distribution for Swedish signs is shown in Table 3. From the plot, the aspect ratio (Y-axis) of all the ground truth data dimensions are in the range of 0.5 to 1.0 and 1.5, which is the range the model has to calibrate to accurately. The Swedish signs dataset has fewer variations compared to Stanford drone and XVview2 satellite data and hence performs better.



(a) Swedish Signs range[0.5-1-1.5]



(b) Stanford Drones range[0.5-1-3.5]



(c) XView2 Satellite images range[0.5-1-3.0]

Table 3: Monotone non-centered aerial imagery.

6.2 Regions

Our observation is that the convolutional feature maps used by region-based detectors, like Fast RCNN, can also be used for generating region proposals. Convolutional Neural Networks can be extended with few more layers to generate region proposals. As before features need to be extracted and on top of these convolutional features, we construct an RPN by adding a few additional convolutional layers. From our previous loss function which has a second component for regression, that simultaneously regress region bounds and objectness scores at each location on a regular grid. The final layer is fine-tuned for the task such as aerial or self-driving cars. These regions are predicted for a regular grid and the objectness score for each location, as shown in Figure 6. RPNs are designed to efficiently predict region proposals with a wide range of scales and aspect ratios. The red rectangles are false positives and the purple are best locations with high object scores.

6.3 Anchors

When using a sliding window over the input object at each location, the network predicts k region proposals. To be able to detect the same object in varying size and scale, we predict $4 \times k$ regions. So the higher the value of k the better the model accuracy at the cost of linear increase in computation. The plot in Table 4 shows how the mean overlap with the ground truth performs with increasing k for the three datasets.

We can observe from plots in Table 4 an interesting training property which differentiates the three datasets. The least mean overlap for Swedish signs dataset is only 0.4 compared to greater than 0.65 for XView2 and Stanford drone datasets confirming their poor performance due to overfitting. We define overfitting it our

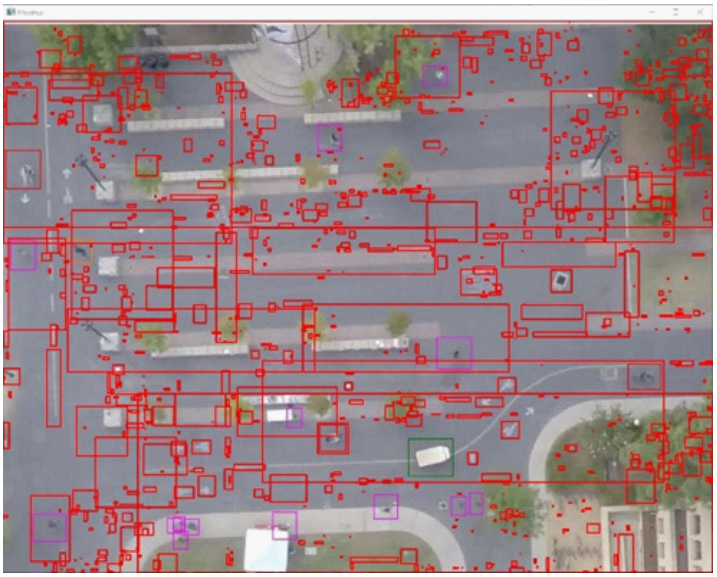
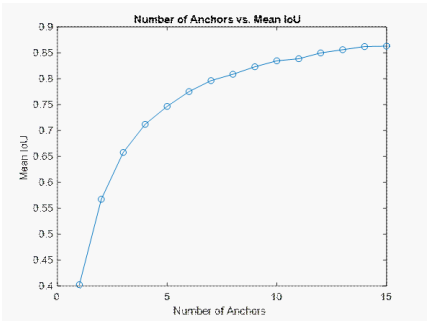
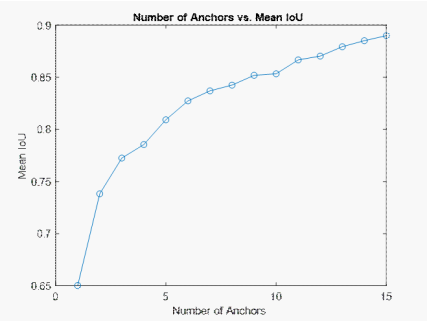


Figure 6: Possible region proposals.

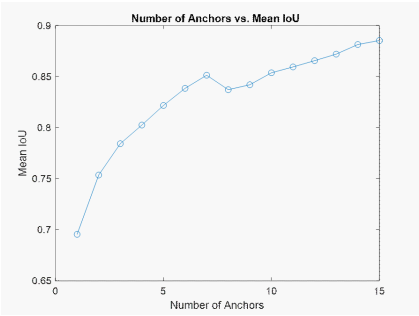
experiments as all objects in the training set are at the same or near locations making the model not to learn further from training data. The mathematical formulas and the distance metric calculations are available in the reference section.



(a) Swedish Signs IOU[0.4]



(b) Stanford Drones range[0.65]



(c) XView2 Satellite images range[0.7]

Table 4: Anchors Vs IOU.

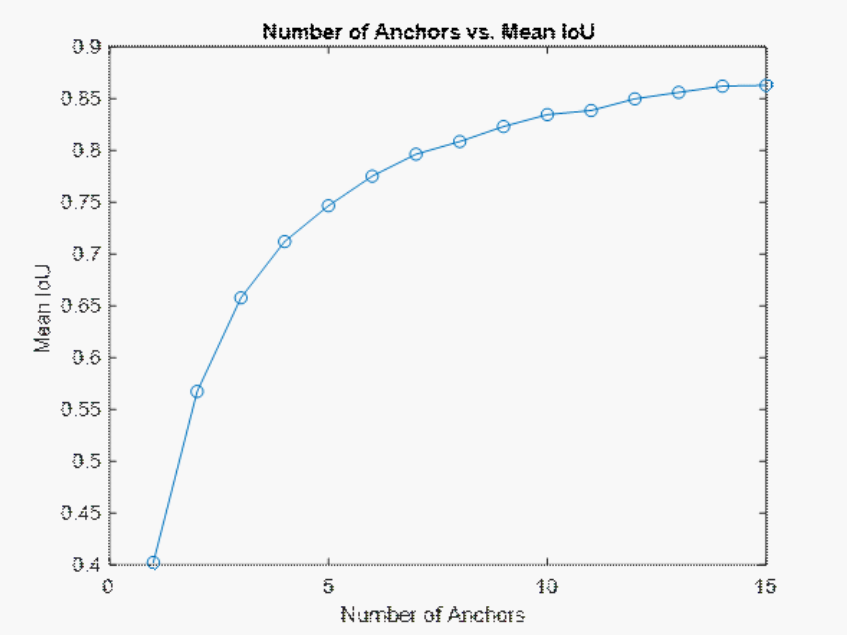
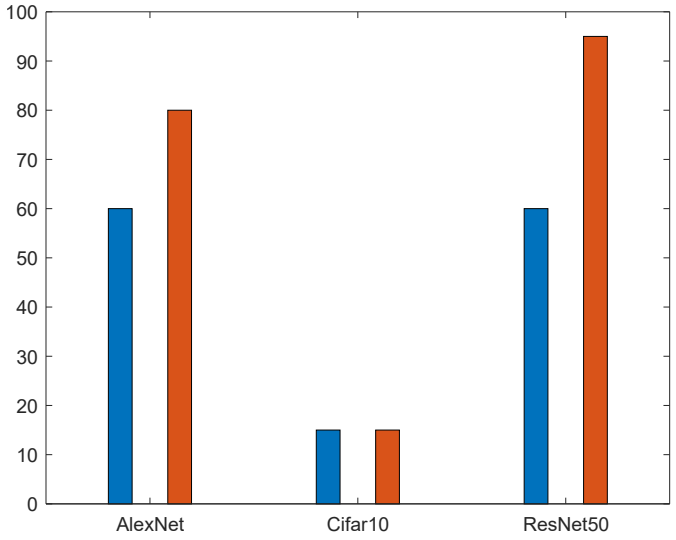


Figure 7: Number of Anchors.

7 Aerial Target Fine Tuning using Pre-Trained Models

The XView2 dataset object detection is challenging as there are many small objects with a monotone background. As these are regular classes of the object but at a very fine scale we use pre-trained networks with average scale to learn the desired scale. We use Matlab built-in feature to extract some of the initial layers from existing ImageNet and Cifar10 to fine-tune over the task of identifying small targets.

Even though using pre-trained models help our fine-tuning to learn richer features, most of the standard architectures use objects of much larger size and scale. So we show the comparative performance of the three datasets by small targets of pixels $\leq 32 \times 32$. Our benchmark, as shown in plot in Table 5 for various architectures, show that deeper layer performs better as it can learn the target features with less number of hyper-parameters.



Original Dataset with $[\leq 32 \times 32]$ (blue) and Targets with $[\geq 32 \times 32](red)$.

Table 5: Pre-Trained Models Vs Target Sizes.

Datasets	IOU	# regions	# anchors	Min. Object Pixels	Accuracy
XView2	-	-	-		
Stanford	-	-	-		
Swedish Signs	-	-	-		

Table 6: Comparison of datasets to train object localization.

8 Summary

Bench marking across many known datasets with XView helps one to build the following properties:

- Reduce the image sizes for GPU computation
- Improve fusing with pre-trained models
- Enable regions with negative sample learning
- Improve localization accuracy

Some of the model parameters are listed below which can be used to optimize the new model efficiently. In summary we see that many parameters affect the deep learning model which are already dealt in the literature. Some of the keypoints are that the XView dataset needs to be adapted to learn localization by further calibrating to the features of a know dataset like Swedish road sign which has the

following characteristics: (i) lowest variance in aspect ratio and at the (ii) same time has many background objects which acts like augmentation of the same target. (iii) Pre-trained cannot be used for targets smaller than 32 pixels, so one needs to resort to custom training, which is highly computationally inefficient.

9 MATLAB Code

```
%% Estimate IOU for XView2 Detector

%%
%% Download CIFAR-10

% Load training data.

clear all;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

echo on;
%evalc('EstimateIOUForStanfordAerial_images.m');
%%%% Load Stored Tables %%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
data = load('F:\ATRC2019\ImageStoreStanfordDroneCarsCartBiker.mat');
imageStore = data.ImageStoreStanfordDroneCars;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Use Minimum size
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
array_aerialDataset = data.ImageStoreStanfordDroneCars;
shuffledIdx = randperm(height(array_aerialDataset));
idx = floor(0.1 * height(array_aerialDataset));
trainingData = data.ImageStoreStanfordDroneCars(shuffledIdx(1:idx),:);
testData = data.ImageStoreStanfordDroneCars(shuffledIdx(idx+1:end),:);


% Display dataset summary
summary(trainingData)
```

```
30
31 % Combine all the ground truth boxes into one array.
32 allBoxes = vertcat(trainingData.Annotations{:});
33 % Plot the box area versus box aspect ratio.
34 aspectRatio = allBoxes(:,3) ./ allBoxes(:,4);
35 area = prod(allBoxes(:,3:4),2);
36
37 figure
38 scatter(area,aspectRatio)
39 xlabel("Box Area")
40 ylabel("Aspect Ratio (width/height)");
41 title("Box area vs. Aspect ratio")
42
43 % Cluster using K-Medoids.
44 numAnchors = 15;
45 [clusterAssignments, anchorBoxes, sumd] = kmedoids(allBoxes
     (:,3:4),numAnchors,'Distance',@iouDistanceMetric);
46
47 % Display estimated anchor boxes. The box format is the [
     width height].
48 numAnchors
49 % Display clustering results.
50 figure
51 gscatter(area,aspectRatio,clusterAssignments);
52 title("K-Mediods with "+numAnchors+" clusters")
53 xlabel("Box Area")
54 ylabel("Aspect Ratio (width/height)");
55 grid
56
57 % Count number of boxes per cluster. Exclude the cluster
     center while
58 % counting.
59 counts = accumarray(clusterAssignments, ones(length(
     clusterAssignments),1),[],@(x)sum(x)-1);
60
61 % Compute mean IoU.
62 meanIoU = mean(1 - sumd./(counts));
63
64
65 maxNumAnchors = 15;
```

```
66 for k = 1:maxNumAnchors
67
68     % Estimate anchors using clustering.
69     [clusterAssignments, anchorBoxes, sumd] = kmedoids(
           allBoxes(:,3:4),k,'Distance',@iouDistanceMetric);
70
71     % Compute mean IoU.
72     counts = accumarray(clusterAssignments, ones(length(
           clusterAssignments),1),[],@(x)sum(x)-1);
73     meanIoU(k) = mean(1 - sumd./(counts));
74 end
75
76 figure
77 plot(1:maxNumAnchors, meanIoU,'-o')
78 ylabel("Mean IoU")
79 xlabel("Number of Anchors")
80 title("Number of Anchors vs. Mean IoU")
81
82
83 function boxWidthHeight = prefixXYCoordinates(boxWidthHeight
       );
84 n = size(boxWidthHeight,1);
85 boxWidthHeight = [ones(n,2) boxWidthHeight];
86 end
87 function dist = iouDistanceMetric(boxWidthHeight,
       allBoxWidthHeight);
88 boxWidthHeight = prefixXYCoordinates(boxWidthHeight);
89 allBoxWidthHeight = prefixXYCoordinates(allBoxWidthHeight);
90
91 dist = 1 - bboxOverlapRatio(allBoxWidthHeight,
       boxWidthHeight);
92 end
```

10 Acknowledgments

This work was supported by AFRL SFFP 2017,2018,2019, and the generous GPU compute time from Microsoft Azure for Research. We want to thank Dr. Mehmood for his suggestions and interesting discussions. We believe the new aerial datasets provided by Sensor Directorate will promote the development of real-time processing of satellite imagery and robust localization algorithms in the sensor domain.

Bibliography

[1] A. Aldroubi, C. Cabrelli, U. Molter, and Sui Tang, Dynamical sampling, *Applied and Computational Harmonic Analysis*, doi:10.1016/j.acha.2015.08.014, 2016

[2] A. Aldroubi, C. Cabrelli, A. F. Cakmak, U. Molter, and A. Petrosyan, Iterative actions of normal operators, Submitted. Available at http://arxiv.org/abs/1602.04527.

[3] K. Groechenig, *Foundations of time-frequency analysis*, Birkhäuser Boston, 2001.

[4] J.R.R. Uijlings and K.E.A. van de Sande and T. Gevers and A.W.M. Smeulders.Selective Search for Object Recognition, Available at http://www.huppelen.nl/publications/selectiveSearchDraft.pdf

[5] Ren, Shaoqing and He, Kaiming and Girshick, Ross and Sun, Jian Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks Available at https://doi.org/10.1109/TPAMI.2016.2577031

[6] Vasanth Iyer; Alexander Aved; Todd B. Howlett; Jeffrey T. Carlo; Asif Mehmood; Niki Pissinou; S. S. Iyengar. Fast multi-modal reuse: co-occurrence pre-trained deep learning models Available at https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10996/2519546/Fast-multi-modal-reuse–co-occurrence-pre-trained-deep/10.1117/12.2519546.full

[7] Vasanth Iyer; Alexander Aved; Todd B. Howlett; Jeffrey T. Carlo; Bernard Abayowa. Autoencoder versus pre-trained CNN networks: deep-features applied to accelerate computationally expensive object detection in real-time video streams Available at http://spie.org/Publications/Proceedings/Paper/10.1117/12.2326848

[8] Fredrik Larsson and Michael Felsberg Using Fourier Descriptors and Spatial Models Available at doi:10.1007/978-3-642-21227-7_23

[9] xView: Objects in Context in Overhead Imagery. http://xviewdataset.org/#dataset

[10] Deng, J. and Dong, W. and Socher, R. and Li, L.-J. and Li, K. and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database