

Satellite Imagery Multiscale Rapid Detection with Windowed Networks

Adam Van Etten
In-Q-Tel CosmiQ Works
avanetten@iqt.org

Abstract

Detecting small objects over large areas remains a significant challenge in satellite imagery analytics. Among the challenges is the sheer number of pixels and geographical extent per image: a single DigitalGlobe satellite image encompasses over 64 km^2 and over 250 million pixels. Another challenge is that objects of interest are often minuscule (~ 10 pixels in extent even for the highest resolution imagery), which complicates traditional computer vision techniques. To address these issues, we propose a pipeline (SIMRDWN) that evaluates satellite images of arbitrarily large size at native resolution at a rate of $\geq 0.2 \text{ km}^2/\text{s}$. Building upon the tensorflow object detection API paper [9], this pipeline offers a unified approach to multiple object detection frameworks that can run inference on images of arbitrary size. The SIMRDWN pipeline includes a modified version of YOLO (known as YOLT [25]), along with the models in [9]: SSD [14], Faster R-CNN [22], and R-FCN [3]. The proposed approach allows comparison of the performance of these four frameworks, and can rapidly detect objects of vastly different scales with relatively little training data over multiple sensors. For objects of very different scales (e.g. airplanes versus airports) we find that using two different detectors at different scales is very effective with negligible runtime cost. We evaluate large test images at native resolution and find mAP scores of 0.2 to 0.8 for vehicle localization, with the YOLT architecture achieving both the highest mAP and fastest inference speed.

1. Introduction

Computer vision techniques have made great strides in the past few years since the introduction of convolutional neural networks [11] in the ImageNet [23] competition. The availability of large, high-quality labeled datasets such as ImageNet [23], PASCAL VOC [4] and MS COCO [13] have helped spur a number of impressive advances in rapid object detection that run in near real-time; four of the best are: Faster R-CNN [22], R-FCN [3], SSD [14], and YOLO [20],[21]. Faster R-CNN and R-FCN typically in-

gests 1000×600 pixel images [22], [3], whereas SSD uses 300×300 or 512×512 pixel input images [14], and YOLO runs on either 416×416 or 544×544 pixel inputs [21]. While the performance of all these frameworks is impressive, none can come remotely close to ingesting the $\sim 16,000 \times 16,000$ input sizes typical of satellite imagery. The speed and accuracy tradeoffs of Faster-RCNN, R-FCN, and SSD were compared in depth in [9]. Missing from these comparisons was the YOLO framework, which has demonstrated competitive scores on the PASCAL VOC dataset, along with high inference speeds. The YOLO authors also showed that this framework is highly transferrable to new domains by demonstrating superior performance to other frameworks (i.e. SSD and Faster R-CNN) on the Picasso Dataset [5] and the People-Art Dataset [1]. In addition an extension of the YOLO framework (dubbed YOLT for You Only Look Twice) [25] showed promise in localizing objects in satellite imagery. The speed, accuracy, and flexibility of YOLO therefore merits a full comparison with the other three frameworks, and motivates this study.

The application of deep learning methods to traditional object detection pipelines is non-trivial for a variety of reasons. The unique aspects of satellite imagery necessitate algorithmic contributions to address challenges related to the spatial extent of foreground target objects, complete rotation invariance, and a large scale search space. Excluding implementation details, algorithms must adjust for:

Small spatial extent In satellite imagery objects of interest are often very small and densely clustered, rather than the large and prominent subjects typical in ImageNet data. In the satellite domain, resolution is typically defined as the ground sample distance (GSD), which describes the physical size of one image pixel. Commercially available imagery varies from 30 cm GSD for the sharpest DigitalGlobe imagery, to 3 – 4 meter GSD for Planet imagery. This means that for small objects such as cars each object will be only ~ 15 pixels in extent even at the highest resolution.

Complete rotation invariance Objects viewed from overhead can have any orientation (e.g. ships can have any

heading between 0 and 360 degrees, whereas trees in ImageNet data are reliably vertical)

Training example frequency There is a relative dearth of training data (though efforts such as SpaceNet¹ are attempting to ameliorate this issue)

Ultra high resolution Input images are enormous (often hundreds of megapixels), so simply downsampling to the input size required by most algorithms (usually a few hundred pixels) is not an option

On the plus side, one can leverage the relatively constant distance from sensor to object, which is well known and is typically ~ 400 km. This coupled with the nadir facing sensor results in consistent pixel to metric ratio of objects.

Section 2 details in further depth the challenges faced by standard algorithms when applied to satellite imagery. The remainder of this work is broken up to describe the proposed contributions as follows. Section 3 describes model architectures. With regard to rotation invariance and small labeled training dataset sizes, Section 4 describes data augmentation and size requirements. Section 5 details the test dataset. Section 6 details the experiment design and our method for splitting, evaluating, and recombining large test images of arbitrary size at native resolution. Finally, the performance of the various algorithms is discussed in detail in Section 7.

2. Related Work

Many recent papers that apply advanced machine learning techniques to aerial or satellite imagery focus on a slightly different problem than the one we attempt to address. For example, [15] showed good performance on localizing objects in overhead imagery; yet with an inference speed of 10 – 40 seconds per 1280×1280 pixel image chip this approach will not scale to large area inference. Efforts to localize surface-to-air-missile sites [16] with satellite imagery and sliding window classifiers work if one only is interested in a single object size of hundreds of meters. Running a sliding window classifier across the image to search for small objects of interest quickly becomes computationally intractable, however, since multiple window sizes will be required for each object size. For perspective, one must evaluate over one million sliding window cutouts if the target is a 10 meter boat in a DigitalGlobe image.

Efforts such as [17], [26] have shown success in extracting roads from overhead imagery via segmentation techniques. Similarly, [24] extracted rough building footprints via pixel segmentation combined with post-processing techniques; such segmentation approaches are quite different from the rapid object detection approach we propose.

Application of rapid object detection algorithms to the remote sensing sphere is still relatively nascent, as evidenced by the lack of reference to SSD, Faster-RCNN, or YOLO in a recent survey of object detection in remote sensing [2]. While tiling a large image is still necessary, the larger field of view of these frameworks (a few hundred pixels) compared to simple classifiers (as low as 10 pixels) results in a reduction in the number of tiles required by a factor of over 1000. This reduced number of tiles yields a corresponding marked increase in inference speed. In addition, object detection frameworks often have much improved background differentiation since the network encodes contextual information for each object.

The rapid object detection frameworks of YOLO, SDD, Faster-RCNN, R-FCN have significant runtime advantages to other methods detailed above, yet complications remain. For example, small objects in groups, such as flocks of birds present a challenge [20], caused in part by the multiple downsampling layers of the convolutional networks. Further, these multiple downsampling layers result in relatively coarse features for object differentiation; this poses a problem if objects of interest are only a few pixels in extent. For example, consider the default YOLO network architecture, which downsamples by a factor of 32 and returns a 13×13 prediction grid [21]; this means that object differentiation is problematic if object centroids are separated by less than 32 pixels. Faster-RCNN downsample by a factor of 16 by default [22], which in theory permits a higher density of object than the standard YOLO architecture. SSD incorporates features at multiple downsampling layers to improve performance on small objects [14]. R-FCN proposes 300 regions of interest, and then refines positions within that ROI via a $k \times k$ grid, where by default $k = 3$. Another difficulty for object detection algorithms applied to satellite imagery is that algorithms often struggle to generalize objects in new or unusual aspect ratios or configurations [20]. Since objects can have arbitrary heading, this limited range of invariance to rotation is troublesome.

Our response is to leverage rapid object detection algorithms to evaluate satellite imagery with a combination of local image interpolation and a multiscale ensemble of detectors. Along with attempting to address the issues listed above and in Section 1, we spend significant effort comparing how well SSD, Faster-RCNN, RFCN, and YOLO/YOLT perform when applied to satellite imagery.

3. SIMRDWN

In order to address the limitations discussed in Section 2, we implement an object detection framework optimized for overhead imagery: Satellite Imagery Multiscale Rapid Detection with Windowed Networks (SIMRDWN). We extend the Darknet neural network framework [19] and update a number of the C libraries to enable analysis of geospatial

¹<https://aws.amazon.com/public-datasets/spacenet/>

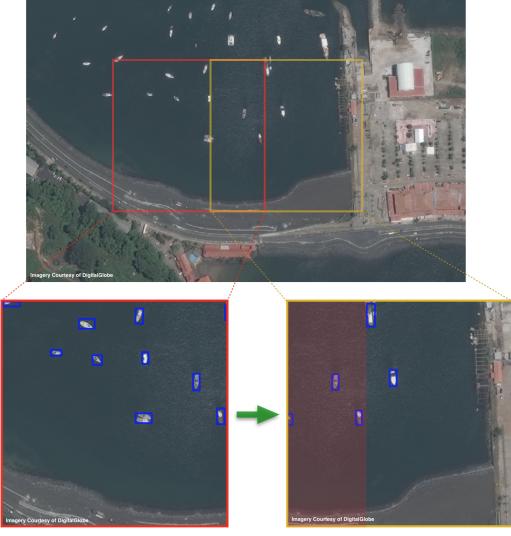


Figure 1: Example of 416 pixel sliding window going from left to right across a large test image. The overlap of the bottom right image is shown in red. Non-maximal suppression of this overlap is necessary to refine detections at the edge of the cutouts where objects may be truncated by the window boundary.

imagery and integrate with external python libraries [25]. We combine this modified Darknet code with the Tensorflow object detection API [9] to create a unified framework. Current rapid object detection frameworks can only infer on images a few hundred pixels in size; since our framework is designed for overhead imagery we implement techniques to analyze test images of arbitrary size.

3.1. Large Image Inference

We partition testing images of arbitrary size into manageable cutouts (416 pixels be default) and run each cutout through the trained models. We refer to this process as windowed networks. Partitioning takes place via a sliding window with user defined bin sizes and overlap (15% by default), see Figure 1. We record the position of each sliding window cutout by naming each cutout according to the schema:

`ImageName|row_column_height_width.ext`
For example:

`panama50cm|1370_1180_416_416.tif`

3.2. Post-Processing

Much of the utility of satellite (or aerial) imagery lies in its inherent ability to map large areas of the globe. Thus, small image chips are far less useful than the large field of view images produced by satellite platforms. The final step in the object detection pipeline therefore seeks to stitch to-

gether the hundreds or thousands of testing chips into one final image strip.

For each cutout the bounding box position predictions returned from the classifier are adjusted according to the `row` and `column` values of that cutout; this provides the global position of each bounding box prediction in the original input image. The 15% overlap ensures all regions will be analyzed, but also results in overlapping detections on the cutout boundaries. We apply non-maximal suppression to the global matrix of bounding box predictions to alleviate such overlapping detections.

3.3. Model Architectures

YOLO We follow the implementation of [25] and utilize a modified Darknet [19] framework to apply the standard YOLO configuration. We use the standard model architecture of YOLOv2 [21], which outputs a 13×13 grid. Each convolutional layer is batch normalized with a leaky rectified linear activation, save the final layer that utilizes a linear activation. The final layer provides predictions of bounding boxes and classes, and has size: $N_f = N_{\text{boxes}} \times (N_{\text{classes}} + 5)$, where N_{boxes} is the number of boxes per grid (5 by default), and N_{classes} is the number of object classes [20]. We train with stochastic gradient descent and maintain many of the hyper parameters of [21]: 5 boxes per grid, an initial learning rate of 10^{-3} , a weight decay of 0.0005, and a momentum of 0.9. We use a batch size of 16 and train for 60,000 iterations.

YOLT To reduce model coarseness and accurately detect dense objects (such as cars), we follow [25] and implement a network architecture that uses 22 layers and downsamples by a factor of 16 rather than the standard $32 \times$ downsampling of YOLO. Thus, a 416×416 pixel input image yields a 26×26 prediction grid. Our architecture is inspired by the 28-layer YOLO network, though this new architecture is optimized for small, densely packed objects. The dense grid is unnecessary for diffuse objects such as airports, but improves performance for high density scenes such as parking lots; the fewer number of layers increases run speed. To improve the fidelity of small objects, we also include a passthrough layer (described in [21], and similar to identity mappings in ResNet [6]) that concatenates the final 52×52 layer onto the last convolutional layer, allowing the detector access to finer grained features of this expanded feature map. We utilize the same hyperparameters as the YOLO implementation.

SSD We follow the SSD implementation of [9]. We experiment with both Inception V2 [10] and MobileNet [8] architectures. For both models we adopt a base learning rate of 0.004 and a decay rate of 0.95. We train

for 30,000 iterations with a batch size of 16, and use the “high-resolution” setting of 600×600 pixel image sizes. These two SSD model architectures are two of the fastest models tested by [9].

Faster-RCNN As with SSD, We follow the implementation of [9] (which closely follows [22]), and adopt the ResNet 101 [7] architecture, (which [9] noted as one of the “sweet spot” models in their comparison of speed/accuracy tradeoffs). We use the “high-resolution” setting of 600×600 pixel image sizes, and use a batch size of 1 with an initial learning rate of 0.0001.

R-FCN As with Faster-RCNN and SSD, we leverage the detailed optimization of [9] for hyperparameter selection. We utilize the ResNet 101 [7] architecture. As with Faster-RCNN, we also explored the ResNet 50 architecture, but found no significant performance increase. We use the same parameters as Faster-RCNN, namely the “high-resolution” setting of 600×600 pixel image sizes, and a batch size of 1.

4. Training Data

Training data is collected from small chips of large images from three sources: DigitalGlobe satellites, Planet satellites, and aerial platforms. Labels are comprised of a bounding box and category identifier for each object (see Figure 2). We initially focus on four categories: airplanes, boats, cars, and airports. For objects of very different scales (e.g. airplanes vs airports) we show in Section 7.2 that using two different detectors at different scales is very effective.

Cars The Cars Overhead with Context (COWC) [18] dataset is a large, high quality set of annotated cars from overhead imagery collected over multiple locales. Data is collected via aerial platforms, but at a nadir view angle such that it resembles satellite imagery. The imagery has a resolution of 15 cm GSD that is approximately double the current best resolution of commercial satellite imagery (30 cm GSD for DigitalGlobe). Accordingly, we convolve the raw imagery with a Gaussian kernel and reduce the image dimensions by half to create the equivalent of 30 cm GSD images. Labels consist of simply a dot at the centroid of each car, and we draw a 3 meter bounding box around each car for training purposes. We reserve the largest geographic region (Utah) for testing, leaving 13,303 labeled training cars. Training images are cut into 416 pixel chips, corresponding to 125 meter window sizes.

Airplanes We labeled eight DigitalGlobe images over airports for a total of 230 objects in the training set. Training images are cut into chips of 125-200 meters depending on resolution.



Figure 2: SIMRDWN Training data examples. *Top left:* boats in DigitalGlobe imagery. *Top right:* airplanes in DigitalGlobe imagery. *Bottom left:* Cars from COWC [18] aerial imagery; the red dot denotes the COWC label and the purple box is our inferred 3 meter bounding box. *Bottom right:* Airport labeled in Planet imagery.

Boats We labeled three DigitalGlobe images taken over coastal regions for a total of 556 boats. Training images are cut into chips of 125-200 meters depending on resolution.

Airports We labeled airports in 37 Planet images for training purposes, each with a single airport per 5000m chip. Obviously, the lower resolution of Planet imagery of 3-4 meter GSD limits the utility of this imagery for vehicle detection.

To address unusual aspect ratios and configurations we augment this training data by rotating training images about the unit circle to ensure that the classifier is agnostic to object heading. We also randomly scale the images in HSV (hue-saturation-value) to increase the robustness of the classifier to varying sensors, atmospheric conditions, and lighting conditions. Even with augmentation, the raw training datasets for airplanes, airports, and watercraft are quite small by computer vision standards, and a larger dataset may improve the inference performance detailed in Section 7.

A potential additional data source is provided by the SpaceNet satellite imagery dataset², which contains a large corpus of labeled building footprints in polygon (not bounding box) format. While bounding boxes are not ideal for precise building footprint estimation, this dataset nevertheless merits future investigation. The impending release of

²<https://spacenetchallenge.github.io/>

Table 1: Train/Test Split

Object Class	Training Examples	Test Examples
Airport*	37	10
Airplane*	230	74
Boat*	556	100
Car†	13,303	19,807

* Internally labeled

† External Dataset

the X-View satellite imagery dataset [12] with 60 object classes and approximately one million labeled object instances will also be of great use for training purposes once available.

5. Test Images

To ensure test robustness and to penalize overtraining on background features, all test images are taken from different geographic regions than training examples. Our dataset for airports is the smallest, with only ten Planet images available for testing. See Table 1 for the train/test split for each category. For airplane testing we label four DigitalGlobe images for a total of 74 airplanes. Our airplane training dataset contains only airliners, though some of the test objects are small personal aircraft -not all classifiers perform well on these objects. Two DigitalGlobe and two Planet coastal images are labeled, yielding 771 test boats. Since we extract test objects from different images than our training set, the sea state is also different in our test images compared to the training images. In addition, two of the four coastal test images are from 3 meter resolution Planet imagery, which further tests the robustness of our models since all training objects are taken from high resolution 0.30 or 0.50 meter DigitalGlobe imagery. The externally labeled cars test dataset is by far the largest; we reserve the largest geographic region of the COWC dataset (Utah) for testing, yielding 19,807 test cars.

6. Experiment Procedure

6.1. Training

Each of the five architectures discussed in Section 3.3 (Faster RCNN Resnet 101, R-FCN Resnet 101, SSD Inception v2, SSD MobileNet, YOLT) are trained on the same data. We create a list of training images and labels for YOLT training, and transform that list into a tfrecord for training the tensorflow models. Models are trained for approximately the same amount of time (as detailed in Section 3.3) of 24-48 hours. We train two separate models for each architecture, one designed for vehicles, and the other for airports (for the rationale behind this approach, see Section

7.1).

6.2. Test Evaluation

For each test image we execute Sections 3.1 and 3.2 to yield bounding box predictions. For comparison of predictions to ground truth we define a true positive as having an intersection over union (IOU) of greater than a given threshold. An IoU of 0.5 is often used as the threshold for a correct detection. We adopt an IoU of 0.5 to indicate a true positive, though we adopt a lower threshold for cars (which typically only 10 pixels in extent) of 0.25. This mimics Equation 5 of ImageNet [23], which sets an IoU threshold of 0.25 for objects 10 pixels in extent.

Precision-recall curves are computed by evaluating test images over a range of probability thresholds. At each of 30 evenly spaced thresholds between 0.05 and 0.95, we discard all detections below the given threshold. Non-max suppression for each object class is subsequently applied to the remaining bounding boxes; the precision and recall at that threshold is tabulated from the summed true positive, false positive, and false negatives of all test images. Finally, we compute the average precision (AP) for each object class and each model, along with the mean average precision (mAP) for each model.

7. Object Detection Results

7.1. Preliminary Object Detection Results

Initially, we attempt to train a single classifier to recognize all four categories listed above, both vehicles and airports. We note a number of spurious airport detections in this example (see Figure 3), as downsampled runways look similar to highways at the wrong scale.

7.2. Scale Confusion Mitigation

There are multiple ways one could address the false positive issues noted in Figure 3. Recall from Section 4 that for this exploratory work our training set consists of only a few dozen airports, far smaller than usual for deep learning models. Increasing this training set size could greatly improve our model, particularly if the background is highly varied. Another option would involve a post-processing step to remove any detections at the incorrect scale (e.g. an airport with a size of ~ 50 meters). Another option is to simply build dual classifiers, one for each relevant scale.

We opt to utilize the scale information present in satellite imagery and run two different classifiers: one trained for vehicles/buildings, and the other trained only to look for airports in downsampled Planet images a few kilometers in extent. Running a second classifier at a larger scale has a negligible impact on runtime performance, since in a given image there are $\approx 1\%$ as many 2000 meter chips as 200 meter chips.



Figure 3: Poor results of the universal YOLT model applied to DigitalGlobe imagery on two different scales (200m, 1500m). Airplanes are in red. The cyan boxes mark spurious detections of runways, caused in part by confusion from small scale linear structures such as highways.

7.3. Results

For large validation images, we run the classifier at two different scales: 200m, and 5000m. The first scale is designed for vehicles (see Figures 4, 5), and the larger scale is optimized for large infrastructure such as airports (see Supplemental Material). We break the validation image into appropriately sized bins and run each image chip on the appropriate classifier. The myriad results from the many image chips and multiple classifiers are combined into one final image, and overlapping detections are merged via non-maximal suppression. Model performance is shown in Figure 6.

Results for R-FCN and Faster-RCNN do not track with the conclusions of [9] that these models occupy a “sweet spot” in terms of speed and accuracy. Inspection of results indicate that both these models struggle with different object sizes (e.g. boats larger than the typical training example), and are very sensitive to background conditions. In an effort to improve results, we experiment with model hyperparameters, and for each model we explore the following: training runs of [30,000, 100,000, 300,000] iterations, input image size of [416, 600], first stage stride of [8, 16], batch size of [1, 4, 8]. These experiments yield no improvement over the default hyperparameters listed in Section 3.3, so it appears that at least for our dataset Faster RCNN and R-FCN struggle to localize objects of interest.

Airport detection is poor for all models, likely the result of the small training set size, since airports are a large and distinctive feature that do not suffer from many of the complications listed in Section 1. It does appear that the YOLO/YOLT models perform significantly better with this training set, though further research is required to determine



Figure 4: Portion of evaluation image with the YOLT model showing labeled boats. False positives are shown in red, false negatives are yellow, true positives are green, and blue rectangles denote ground truth for all true positive detections.



Figure 5: Portion of evaluation image with the YOLT model showing labeled aircraft. False positives are shown in red, false negatives are yellow, true positives are green, and blue rectangles denote ground truth for all true positive detections. Performance is good despite the atypical look angle and lighting conditions.

if these models are truly more robust or if another mechanism explains the superior performance of YOLO/YOLT to other models. We also note a significant increase in mAP from YOLO to YOLT, which stems from improved localization of cars and boats (which are often tightly packed) where the denser network of YOLT pays dividends.

Table 2 displays object detection performance and speed for each model architecture. We report inference speed in terms of GPU time to run the inference step. Currently, pre-processing (i.e. splitting test images into smaller cutouts) and post-processing (i.e. stitching results back into one global image) is not fully optimized and is performed on the CPU, which increases run time by a factor of 1.5 – 1.75. Inference rates for airports are $\sim 600 \times$ faster than the inference rate for vehicles reported in Table 2, ranging from $60 \text{ km}^2/\text{s}$ (Faster RCNN) to $270 \text{ km}^2/\text{s}$ (YOLT).

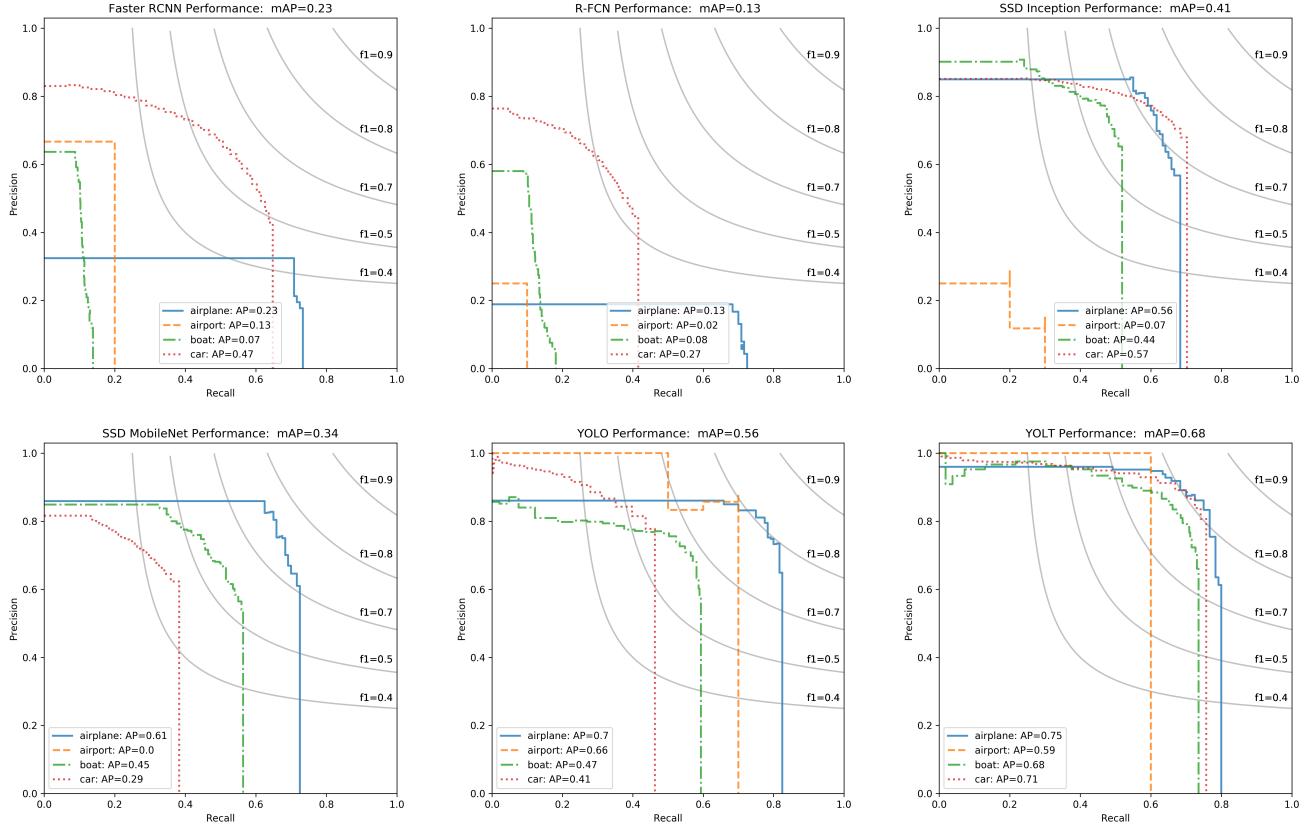


Figure 6: Precision-recall curves for each model

Table 2: Performance vs Speed

Architecture	mAP	Inference Rate (km ² /s)
Faster RCNN ResNet101	0.23	0.09
RFCN ResNet101	0.13	0.17
SSD Inception	0.41	0.22
SSD MobileNet	0.34	0.32
YOLO	0.56	0.42
YOLT	0.68	0.44

7.4. Resolution Performance

We explore the effect of window size (closely related to image resolution) on object detection performance. The YOLT model returns the best AP for cars, though dense regions still pose a challenge for the detector. The YOLT model is trained on native resolution imagery of 416 pixels in extent. In an attempt to improve performance, we train on image cutouts of only 208 pixels, these cutouts are subsequently upsampled to size 416 pixels when ingested by the

network. This simulates higher resolution imagery, though no extra information is provided. This smaller window size decreases inference speed by a factor of four, but markedly improves performance, see Figure 7.

8. Conclusions

Object detection algorithms have made great progress as of late in localizing objects in ImageNet style datasets. Such algorithms are rarely well suited to the object sizes or orientations present in satellite imagery, however, nor are they designed to handle images with hundreds of megapixels.

To address these limitations we implemented a fully convolutional neural network pipeline (SIMRDWN) to rapidly localize vehicles and airports in satellite imagery. This pipeline unifies leading object detection algorithms such as SSD, Faster RCNN, R-FCN, and YOLT into a single framework that rapidly analyzes test images of arbitrary size. We noted poor results from a combined classifier due to confusion between small and large features, such as highways and runways. Training dual classifiers at different scales (one for vehicles, and one for airports), yielded far better results.

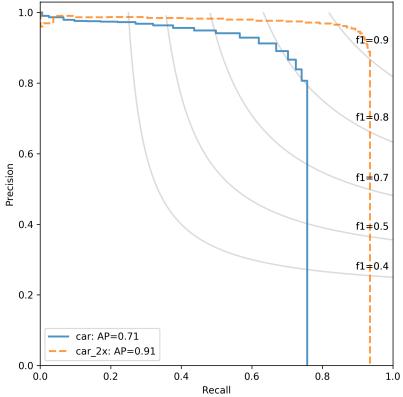


Figure 7: Performance of the YOLT model trained and tested at native resolution (solid), and a model trained and tested at simulated double resolution (dashed).

Our training dataset is quite small by computer vision standards, and mAP scores range from 0.13 (R-FCN) to 0.68 (YOLT) for our test set. While the mAP scores may not be at the level many readers are accustomed to from ImageNet competitions, object detection in satellite imagery is still a relatively nascent field and has unique challenges. In addition, our training dataset for most categories is relatively small for supervised learning methods.

Our test set is derived from different geographic regions than the training set, and the low mAP scores are unsurprising given the small training set size provides relatively little background variation. Nevertheless, the YOLT architecture did perform significantly better than the other rapid object detection frameworks, indicating that it appears better able to disentangle objects from background with small training sets.

Inference speeds for vehicles are high, at $0.09 \text{ km}^2/\text{s}$ (Faster RCNN) to $0.44 \text{ km}^2/\text{s}$ (YOLT). We also demonstrated the ability to train on one sensor (e.g. DigitalGlobe), and apply our model to a different sensor (e.g. Planet). The highest inference speed translates to a runtime of < 6 minutes to localize all vehicles in an area of the size of Washington DC, and < 2 seconds to localize airports over this area. DigitalGlobe’s WorldView3 satellite³ covers a maximum of $680,000 \text{ km}^2$ per day, so at SIMRDWN inference speed a 16 GPU cluster would provide real-time inference on satellite imagery. Results so far are intriguing, and it will be interesting to explore in future works how well the SIMRDWN pipeline performs as further datasets become available and the number of object categories increases.

³<http://worldview3.digitalglobe.com>

References

- [1] H. Cai, Q. Wu, T. Corradi, and P. Hall. The cross-depiction problem: Computer vision algorithms for recognising objects in artwork and in photographs. *CoRR*, abs/1505.00110, 2015.
- [2] G. Cheng and J. Han. A survey on object detection in optical remote sensing images. *CoRR*, abs/1603.06201, 2016.
- [3] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016.
- [4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.
- [5] S. Ginosar, D. Haas, T. Brown, and J. Malik. Detecting people in cubist art. *CoRR*, abs/1409.6235, 2014.
- [6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.
- [9] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR*, abs/1611.10012, 2016.
- [10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.
- [11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
- [12] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord. xview: Objects in context in overhead imagery. *CoRR*, abs/1802.07856, 2018.
- [13] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.
- [14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.
- [15] Y. Long, Y. Gong, Z. Xiao, and Q. Liu. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498, May 2017.
- [16] R. A. Marcum, C. H. Davis, G. J. Scott, and T. W. Nivin. Rapid broad area search and detection of Chinese surface-to-air missile sites using deep convolutional neural networks. *Journal of Applied Remote Sensing*, 11(4):042614, Oct. 2017.
- [17] V. Mnih and G. E. Hinton. Learning to detect roads in high-resolution aerial images. In K. Daniilidis, P. Maragos, and

- N. Paragios, editors, *Computer Vision – ECCV 2010*, pages 210–223, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.
- [18] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. *CoRR*, abs/1609.04453, 2016.
 - [19] J. Redmon. Darknet: Open source neural networks in c. <http://pjreddie.com/darknet/>, 2013–2017.
 - [20] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.
 - [21] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016.
 - [22] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.
 - [23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
 - [24] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios. Building detection in very high resolution multispectral data with deep learning features. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1873–1876, July 2015.
 - [25] A. Van Etten. You Only Look Twice: Rapid Multi-Scale Object Detection In Satellite Imagery. *ArXiv e-prints*, May 2018.
 - [26] Z. Zhang, Q. Liu, and Y. Wang. Road extraction by deep residual u-net. *CoRR*, abs/1711.10684, 2017.

A. Image Appendix



Figure 8: Portion of evaluation image with the YOLT model. False positives are shown in red, false negatives are yellow, true positives are green, and blue rectangles denote ground truth for all true positive detections. This image demonstrates some of the challenges of our test set and the robustness of the model. Our airplane training set only contains airliners, so we only label commercial aircraft in test images, yet the many false negatives in this image are caused by detections of military aircraft.

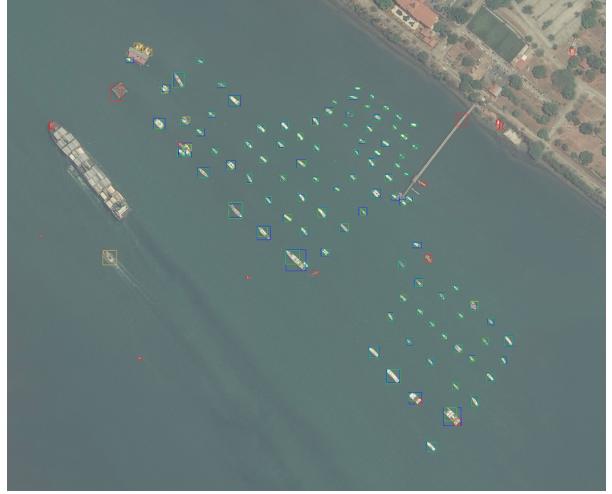


Figure 10: Evaluation image with the R-FCN model. False positives are shown in red, false negatives are yellow, true positives are green, and blue rectangles denote ground truth for all true positive detections.



Figure 9: Evaluation image with the SSD Inception v2 model. False positives are shown in red, false negatives are yellow, true positives are green, and blue rectangles denote ground truth for all true positive detections.

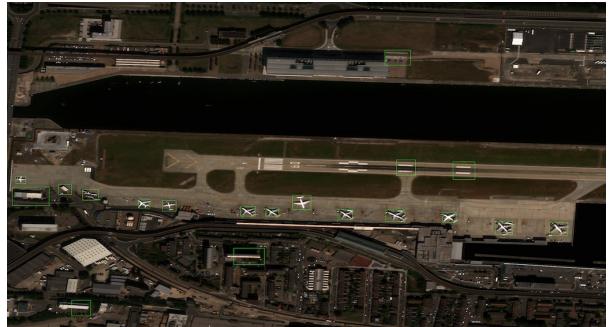


Figure 11: Raw detections from the Faster RCNN model at a detection threshold of 0.5. Airplanes are shown in green; the false positive rate is high.

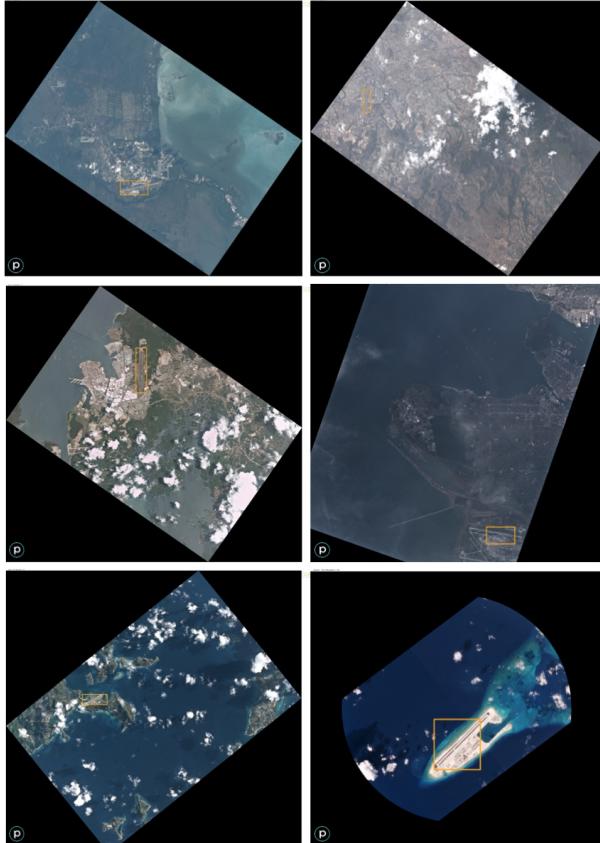


Figure 12: Successful detections of airports and airstrips (orange) in Planet images with the YOLT model over both maritime backgrounds and complex urban backgrounds. Note that clouds are present in most images. The middle-right image demonstrates robustness to low contrast images. Each image takes between 1–3 seconds to analyze, depending on size



Figure 13: Car detection performance on a 600×600 m aerial image at 30 cm GSD over Salt Lake City with the YOLT model trained at 2x resolution. F1 = 0.97 for this test image.



Figure 14: SIMRDWN classifier applied to a SpaceNet DigitalGlobe 50 cm GSD image containing airplanes (blue), boats (red), and runways (orange). In this image we note the following F1 scores: airplanes = 0.83, boats = 0.84, airports = 1.0.