The Effects of Super-Resolution on Object Detection Performance in Satellite Imagery

Jacob Shermeyer and Adam Van Etten
CosmiQ Works, In-Q-Tel
{jshermeyer,avanetten}@iqt.org

Abstract

We explore the application of super-resolution techniques to satellite imagery, and the effects of these techniques on object detection algorithm performance. Specifically, we enhance satellite imagery beyond its native resolution, and test if we can identify various types of vehicles, planes, and boats with greater accuracy than native resolution. Using the Very Deep Super-Resolution (VDSR) framework and a custom Random Forest Super-Resolution (RFSR) framework we generate enhancement levels of $2\times$*,* $4\times$*, and* $8\times$ *over five distinct resolutions ranging from 30 cm to 4.8 meters. Using both native and super-resolved data, we then train several custom detection models using the SIMRDWN object detection framework. SIMRDWN combines a number of popular object detection algorithms (e.g. SSD, YOLO) into a unified framework that is designed to rapidly detect objects in large satellite images. This approach allows us to quantify the effects of super-resolution techniques on object detection performance across multiple classes and resolutions. We also quantify the performance of object detection as a function of native resolution and object pixel size. For our test set we note that performance degrades from mean average precision (mAP) = 0.53 at 30 cm resolution, down to mAP = 0.11 at 4.8 m resolution. Super-resolving native 30 cm imagery to 15 cm yields the greatest benefit; a* $13 - 36\%$ *improvement in mAP. Super-resolution is less beneficial at coarser resolutions, though still provides a small improvement in performance.*

1. Introduction

The interplay between super-resolution techniques and object detection frameworks remains largely unexplored, particularly in the context of satellite or overhead imagery. Intuitively, one might assume that super-resolution methods should increase object detection performance, as an increase in resolution should add more distinguishable features that an object detection algorithm can use for discrimination. Detecting small objects such as vehicles in satellite imagery remains an exceedingly difficult task for multiple reasons [37] and an artificial increase in resolution may help to alleviate some of these issues. Some of the issues present include:

1. Objects such as cars in satellite imagery have a small spatial extent (as low as 10 pixels) and are often densely clustered.
2. All objects exhibit complete rotation invariance and can have any orientation.
3. Training example frequency is low versus other disciplines. Few datasets exist that have appropriate labels for objects within satellite imagery. The most notable are: SpaceNet [38], A Large-scale Dataset for Object DeTection in Aerial Images (DOTA) [40], Cars Overhead With Context (COWC) [27], and xView [18].
4. Most satellite imagine sensors cover a broad area and contain hundreds of megapixels, thereby producing the equivalent of an ultra-high resolution image. For example, the native imagery used in this study was on average $\approx$ 57 times larger than benchmark super-resolution datasets Set5, Set14, BSD100, and Urban100. When working with modern neural network architectures these images must be tiled into smaller chunks for both training and inference.

Although several studies have been conducted using SR as a pre-processing step [1, 11, 12, 33, 42, 3, 10, 5], none have quantified its affect on object detection performance in satellite imagery across multiple resolutions. This study aims to accomplish that task by training multiple custom object detection models to identify vehicles, boats, and planes in both native and super-resolved data. We then test the models performance on the native (ground-truth) imagery and super-resolved imagery of the same Ground Sample Distance (GSD: the distance between pixels measured on the ground). Additionally, this is the first study to demonstrate the output of super-resolved 15 cm GSD satellite imagery. Although no native 15 cm satellite imagery exists

for comparison, this data can be compared against coarser resolutions to test the benefits provided by super-resolution.

The cost-benefit analysis of such a study is enormous. Satellite manufacturers spend the majority of their budget on the design and launch of satellites. For example, the DigitalGlobe WorldView-4 satellite cost an estimated $\$835$ million dollars when one includes spacecraft, insurance, and launch [8]. Ideally, one could couple an effective SR enhancement algorithm with a smaller, cheaper satellite that captures images in coarser resolution. The process of capturing and subsequently enhancing coarser data could drastically reduce launch cost, expand satellite field of view, reduce the number of satellites in orbit, and improve downlink speeds between satellites and ground control stations.

2. Related Work

2.1. Super-Resolution Techniques and Application to Overhead Imagery

Single-image Super-Resolution (SR) is the process for deriving high-resolution (HR) images from a single low-resolution (LR) image. Although super-resolution remains an ill-posed and difficult problem, recent advances in neural networks and machine learning have enabled more robust SR algorithms that exhibit effective performance. These techniques use high-resolution image pairs to learn the most likely HR features to map to the LR image features and create an output SR product.

Over the past five years, convolutional neural network approaches have been used to produce state of the art super-resolution results. Dong et al. [7] was the first to establish a deep learning approach with SRCNN. This has been followed up by several successive approaches, major alterations, and improvements. Very Deep Super Resolution (VDSR) [15] exhibited state of the art performance and was one of the first to modify the SRCNN approach by creating a deeper network with 20 layers to learn a residual image and transform LR images into HR images. Developed concurrently, the Deeply-Recursive CNN (DRCN) [16] introduced a recursive neural network approach to super-resolve imagery. The Deeply Recursive Residual Network (DRRN) [34] builds upon the VDSR and DRCN advancements using a combination of the residual layers approach and recursive learning in a compact network.

More complex methods followed, such as the Laplacian Pyramid Super-Resolution Network (LapSRN) [17]. Adversarial training has also been employed and the SR Generative Adversarial Network (SRGAN) [19] produces photo-realistic $4\times$ enhanced images. The use of wider and deeper networks has also been proposed. The most notable being Lim et al. [21], which proposed Enhanced Deep Residual Networks (EDSR). Most recently, the Deep Back Projection Network (DBPN) [9] showed state of the art performance for an $8\times$ enhancement by connecting a series of iterative up- and down-sampling stages. Newer block based methods such as the Information Distillation Network (IDN) [14] was developed as a compact network that could gradually extract common features for fast the reconstruction of HR images. In another example, the Residual Dense Network (RDN) [43] uses residual dense blocks to produce strong performance.

Although new and powerful single image SR techniques continue to be developed, these techniques have been infrequently applied to overhead imagery. One of the most notable applications of super resolution to satellite and overhead imagery remains the recent paper by Bosch et al. [2]. The authors analyze several sources of satellite imagery for this research and quantify their success in terms of PSNR for an $8\times$ enhancement using a GAN. In another example, [22] use deep neural networks for simultaneous $4\times$ super-resolution and colorization of satellite imagery. Several papers [41, 25, 35, 20, 28] modify or leverage SRCNN [7] and/or VDSR [15] to successfully super-resolve Jilin-1, SPOT, Pleiades, Sentinel-2, and Landsat imagery.

Ultimately, a few specific papers are direct precursors for this work: In the first, [3] use fine resolution aerial imagery and coarser satellite imagery with a coupled dictionary learning approach to super enhance vehicles and detect them with a simple linear Support Vector Machine model. Their results showed that object detection performance improves when using SR as a pre-processing step versus the native coarser imagery. Xu et al. [42] use sparse dictionary learning to generate synthetic $8\times$ and $16\times$ super-resolved imagery from Landsat and MODIS image pairs. Their results show an increase performance for land-cover change mapping when using the super-resolved imagery. Although these approaches are similar to ours, they fail to use newer neural network based approaches, and are narrower in scope. Finally, [10] super-resolve imagery using DBPN [9] and detect various objects in traditional photography using SSD [23]. They quantify their success in terms of mAP and also add a novel element to this work of designing a loss function to optimize SR for object detection performance. Their results show that end-to-end training of these algorithms gave a performance boost for object detection tasks, and is a promising avenue to explore for future research.

Overall, we hypothesized that SR techniques could improve object detection performance, particularly when using satellite imagery, however no such study has been conducted. To address this question, our study investigates the relationship between object detection performance and resolution, spanning five unique GSD resolutions, with six SR outputs per resolution. Ultimately, we investigate 35 separate resolution profiles for object detection performance.

2.2. Object Detection Techniques

A number of recent papers have applied advanced machine learning techniques to aerial or satellite imagery, yet have focused on a slightly different problem than the one we attempt to address. For example, [24] demonstrated the ability to localize objects in overhead imagery; yet application to larger areas would be problematic, with an inference speed of 10 - 40 seconds per 1280×1280 pixel image chip. Efforts to localize surface to-air-missile sites [26] 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 a large satellite 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. Application of rapid object detection algorithms to the remote sensing sphere is still relatively nascent, as evidenced by the lack of reference to SSD [23], Faster-RCNN [30], R-FCN [6], or YOLO [29] in a recent survey of object detection in remote sensing [4]. 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 (compared to sliding window classifiers) since the network encodes contextual information for each object.

As we seek to study the effect of super-resolution on object detection performance in real-world satellite imagery, and for all of the reasons listed above - rapid object detection frameworks are the logical choice for this study. The premier rapid object detection algorithms (SSD, Faster-RCNN, R-FCN, and a modified version of YOLO called YOLT [36]) were recently incorporated into the unified framework of SIMRDWN [37] that is optimized for ingesting satellite imagery, typically several hundred megapixels in size. The SIMRDWN paper reported the highest performance stemmed from the YOLT algorithm, followed by SSD, with Faster R-CNN and RFCN significantly behind.

3. Dataset

The xView Dataset [18] was chosen for the application of super-resolution techniques and the quantification of object detection performance. Imagery consists of 1,415 km^2 of DigitalGlobe WorldView-3 pan-sharpened RGB imagery at 30 cm native GSD resolution spread across 56 distinct global locations and 6 continents (sans Antarctica). The labeled dataset for object detection contains 1 million object instances across 60 classes annotated with bounding boxes, including various types of buildings, vehicles, planes, trains, and boats. For our purposes, we ultimately discarded classes such as "Building," "Hangar," and "Vehicle Lot" because we found that such objects are better represented by polygonal labels rather than bounding boxes for foundational mapping [38] purposes.

We chose an aggregation schema due to inconsistent labeling within the dataset. Unfortunately, many objects are mislabeled or simply missed by labelers (see Figure 1). This leads to an increase in false positive detection rates and objects being inaccurately tagged as mis-classifications after inference. In addition, many xView classes have a very low number of training examples (e.g. Truck w/Liquid has only 149 examples) that are poorly differentiated from similar classes (e.g. Truck w/Box has 3653 examples and looks very similar to Truck w/Liquid). The question of how many training examples are necessary to disentangle similar classes is beyond the scope of this paper.

Our classes ultimately consist of the following (original xView classes listed in parentheses): Small Aircraft (Fixed-wing Aircraft, Small Aircraft), Large Aircraft (Cargo Plane), Small Vehicle (Passenger Vehicle, Small Car, Pickup Truck, Utility Truck), Bus/Truck (Bus, Truck, Cargo Truck, Truck w/Box, Truck w/Flatbed, Truck w/Liquid, Dump Truck, Haul Truck, Cement Mixer, Truck

(a) (b) (c)

Figure 1: Issues with xView ground truth labels. Red = car, green=truck, orange=bus, yellow=airplane, purple=boat. Note, the incorrectly sized cars in (a), the erroneous "boat" ground truth labels in (b), and the missing cars in (c).

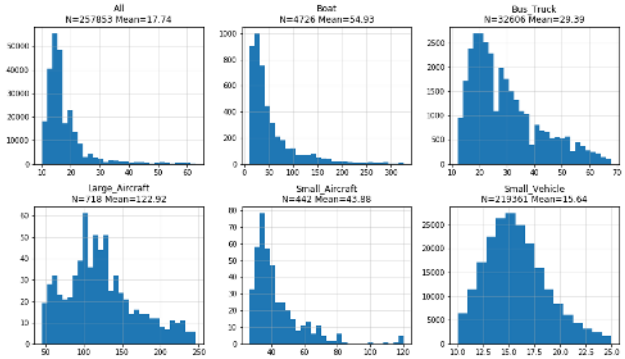


Figure 2: Object size histograms (in pixels), recall that each pixel is 30 cm in extent.

Category	Mean Size (meters)	Counts Train	Test	Total
Boat	16.5	2379	2347	4726
Large Aircraft	36.9	424	294	718
Small Aircraft	13.2	264	178	442
Bus/Truck	8.8	19337	13269	32606
Small Vehicle	4.7	129438	89923	219361

Table 1: Object Counts

Tractor), and Boat (Motorboat, Sailboat, Yacht, Maritime Vessel, Tugboat, Barge, Fishing Vessel, Ferry). See Table 1 for dataset details and Figure 2 for object size histograms.

3.1. Simulation of Optics and Sensors

All data were preprocessed consistently to simulate coarser resolution imagery and test the affects of our SR techniques on a range of resolutions. We intend our results to showcase what can be reasonably accomplished given coarser satellite imagery, rather than simply what is possible given the ideal settings (no blurring, bicubic decimation) under which most SR algorithms are introduced. We attempt to simulate coarser resolution satellite imagery as accurately as possible by simulating the optical point-spread function (PSF) and using a more robust decimation algorithm. This is important because the optics of the telescope greatly impact the appearance of very small objects. The common practice of simply resizing an image by reducing its dimensions by a factor of two will simulate a different sensor containing $1/4$ the number of pixels; yet this approach ignores the different optics present in a properly designed telescope that would be coupled to such a sensor. A properly designed sensor should have pixel size determined by the Nyquist sampling rate: half the size of the mirror resolution determined by the diffraction limit. Given the cost and complexity of launching satellite imaging constellations to orbit, we assume that all imaging satellites will have properly designed sensors. We can use the assumption of Nyquist sampling to determine the PSF of the telescope optics, which can be approximated by a Gaussian of appropriate kernel size:

$$kernel = 0.5 \times GSD_{out} \,/\, GSD_{native} \qquad (1)$$

For our study, data were degraded from the native 30 cm GSD using a variable Gaussian blur kernel to simulate the point-spread function of the satellite depending upon our desired output resolution (Equation 1). We then used inter-area decimation to reduce the dimensions of the blurred imagery to the appropriate output size (e.g. 60 cm imagery will have $1/4$ the number of pixels as 30 cm imagery over the same field of view). We repeat the above procedure to simulate resolutions of 60, 120, 240, and 480 cm. The ground truth data and the outputs from the super-resolution algorithms were randomly split into training (60%) and validation (40%) categories for object detection. The same images are contained in both the training and test sets regardless of resolution to maintain consistency when comparing validation scores.

4. Super-Resolution Techniques

For this study, super-resolution is conducted with two techniques for enhancement levels of $2\times$, $4\times$, $8\times$ over five distinct resolutions ranging from 30 cm to 4.8 meters. We also create 15 cm GSD output imagery using the models trained to super-resolve imagery from 60 cm to 30 cm and 120 cm to 30 cm.

Our first method is a convolutional neural network derived technique called Very Deep Super-Resolution (VDSR) [15]. VDSR has been featured as a baseline for the majority of recent super-resolution research and was one of the first to modify the initially proposed convolutional neural network method SRCNN [7]. This architecture was chosen due its ease of implementation, ability to train for multiple levels of enhancement, use as a standard baseline when introducing new techniques, and favorable performance in the past. We use the standard network parameters as set in the original paper [15] and train for 60 epochs. We chose a patch size of 41 $\times$ 41 pixels and augment by rotations (4) and flipping (2) for eight unique combinations per patch. This process is repeated for each enhancement level (2, 4, and $8\times$), and each is fed into the same network for concurrent training. Average training time for a 2, 4, and $8\times$ enhancement on 200 million pixel example is 55.9 hours on a single Titan Xp GPU. Inference speed on a 544 $\times$ 544 pixel image is very fast ≈ 0.2 seconds on the same hardware, allowing for this method to easily scale to accommodate large satellite images.

The second method is an approach that we have called Random-Forest Super-Resolution (RFSR) and was designed for this work; it requires minimal training time and exhibits high inference speeds. RFSR is an adaptation of other random forest super-resolution techniques such as SRF [32] or SRRF [13] and can process both georeferenced satellite imagery or traditional photography. We chose to include this simpler, less computationally intensive algorithm that does not require GPUs to test its effectiveness against a near state of the art SR solution. The hypothesis is that even a simple technique may improve object detection performance.

Our method uses a random forest regressor with a few standard parameters. The number of estimators is set to 100, the maximum depth to 12, and the minimum samples to split an internal node equal to 200. Finally, we use bootstrapping and out-of-bag samples to estimate the error and R^2 scores on randomly selected unseen data during training. These parameters were finely tuned using empirical testing

	VDSR	RFSR
Inference Time (per image)	0.16 seconds	0.7 seconds
Training Time (for 2, 4, 8×)	55.9 hours	10.8 hours

Table 2: Average inference time per 544×544 pixel image and training time for a set of 1,500 images at native 30 cm GSD resolution. RFSR used a 64GB RAM CPU and VDSR used a NVIDIA Titan Xp GPU for inference and training.

to maximize PSNR scores (see Section 6 for details on metrics) while maintaining minimal training time (4 hours or less per level of enhancement on a 64GB RAM CPU). It should be noted that PSNR scores could be mildly improved using a deeper tree with more estimators, at the cost of training time.

Like several other SR techniques, RFSR is trained only using the luminance component from a YCbCr converted image. HR images are degraded to create LR and HR image pairs. The degraded LR image is then shifted by one and then two pixels in each direction versus the HR image and then compressed into a 3-dimensional array. The original up-sampled LR image is then subtracted from the 3-D LR array, and from the HR image for a residual training schema. This normalizes the LR stack and HR image pair and also removes homogeneous areas, emphasizing important edge effects. After training and inference the interpolated LR image is then added back to the models' output image to create the super-resolved output. RFSR can only produce one level of enhancement (2, 4, or 8×) at a time. Average training time for all three enhancements on ∼ 200 million pixel examples is 10.8 hours on a 64GB RAM CPU. Average inference speed on a 544×544 pixel image is 0.7 seconds for this same hardware (Table 2).

5. Object Detection Techniques

As discussed in Section 1, advanced object detection frameworks have only recently been applied to large satellite imagery via the SIMRDWN framework. In the SIMRDWN paper, the authors reported the highest performance stemmed from the YOLT algorithm, followed by SSD, with Faster R-CNN and RFCN significantly behind. Therefore, we opt to utilize the YOLT and SSD models within SIMRDWN for this study. For the YOLT model we adopt the dense 22-layer network of [36] with a momentum of 0.9, and a decay rate of 0.0005. We use a 544 × 544 pixel training input size (corresponding to 164 × 164 meters). Training occurs for 150 epochs. For the SSD model we follow the TensorFlow Object Detection API implementation with the Inception V2 architecture. 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 same 544 × 544 pixel input size as YOLT. For both YOLT and SSD we train models on the "native" imagery (original 30 cm data, the convolved and resized imagery described in Section 3.1), as well as on the outputs of RFSR and VDSR applied to the object detection training set. This approach yields a multitude of models across the myriad architectures, super-resolution techniques, and resolutions (see Figure **??**, thus enabling a detailed study of performance.

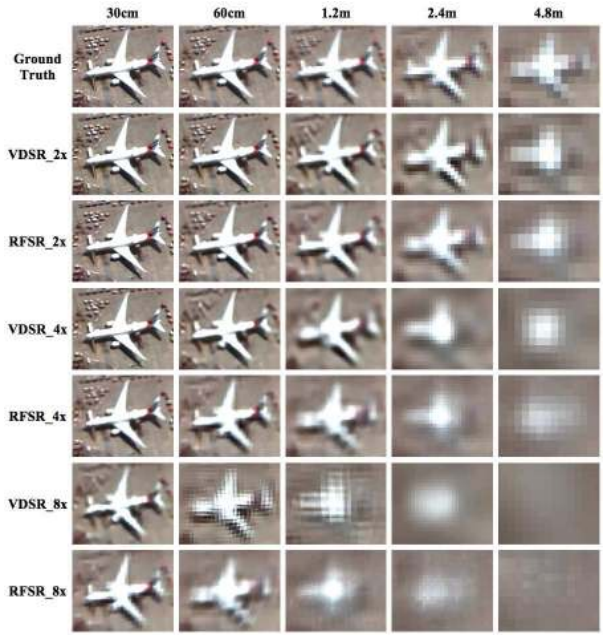


Figure 3: The effects of super-resolution on a plane and neighboring objects. As resolution degrades super-resolution becomes a less tractable solution.

6. Metrics

Overall, super-resolution remains an active field of research with rather limited direct focus on end application. Typical performance metrics include Peak Signal-to-Noise Ratio (PSNR) or the Structural SIMilarity (SSIM) Index (which we report in Section 7.1), however these measures do not quantify the enhancement to object detection performance [39]. Although these images may be more visually appealing as a result of super-resolution, such techniques may have little impact on object detection performance.

For object detection metrics, we compare the ground truth bounding boxes to the predicted bounding boxes for each test image. 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, though we adopt a lower threshold of 0.25 since most of our objects are very small (e.g.: cars are only 10 pixels in extent). This mimics Equation 5 of ImageNet [31], which

Figure 4: Examples of 15 cm GSD super-resolved output from RFSR and VDSR versus the original 30 cm GSD native imagery.

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. One-sigma error bars are computed via bootstrap resampling, using 500 samples for each scenario.

7. Experimental Results

7.1. Super-Resolution Performance

As expected, super-resolution performance was strongest for the VDSR method, although RFSR produces comparable results in some circumstances (Table 3). As in other studies, the metrics degrade as the amount of enhancement increases. Both techniques performed the strongest on the 60 cm imagery, likely because initial bicubic interpolation scores are high and the fact that the image resolution is situated between a coarse and fine scale where the image features are easier to detect and enhance.

GSD_{out}	Scale	Bicubic	VDSR	RFSR
30cm	2×	38.68 / 0.8108	**42.39 / 0.8925**	39.79 / 0.8582
30cm	4×	35.86 / 0.6610	**38.79 / 0.7795**	35.85 / 0.7064
30cm	8×	33.82 / 0.5394	**35.69 / 0.6117**	34.32 / 0.5874
60cm	2×	41.26 / 0.9275	**45.08 / 0.9635**	43.03 / 0.9408
60cm	4×	36.98 / 0.8082	**40.50 / 0.8904**	37.41 / 0.8330
60cm	8×	33.99 / 0.6771	**35.44 / 0.7293**	33.78 / 0.6799
1.2m	2×	36.73 / 0.9151	**39.33 / 0.9497**	38.17 / 0.9448
1.2m	4×	32.49 / 0.7738	**35.25 / 0.8633**	33.47 / 0.8332
1.2m	8×	29.41 / 0.6097	**30.58 / 0.6709**	29.84 / 0.6700
2.4m	2×	35.26 / 0.8848	**41.50 / 0.9624**	36.67 / 0.9250
2.4m	4×	31.09 / 0.6898	**33.75 / 0.8117**	32.00 / 0.7659
2.4m	8×	28.46 / 0.5004	**30.78 / 0.6089**	28.87 / 0.5572
4.8m	2×	34.14 / 0.8404	**37.01 / 0.9097**	35.45 / 0.8953
4.8m	4×	30.42 / 0.6079	**33.13 / 0.7527**	31.24 / 0.6934
4.8m	8×	27.98 / 0.4013	**30.22 / 0.5110**	28.39 / 0.4488

Table 3: Average PSNR / SSIM scores for scale 2×, 4×, and 8× across five super-resolution output GSDs. All test imagery is the xView validation dataset (281 images). Bicubic indicates the scores if LR images are just upscaled using bicubic interpolation to match the HR image size.

A few specific examples of super resolution performance are visible in Figure 3, where we test the effects of our algorithm on a large object like a plane. Visually, VDSR and RFSR both perform strongly at 30 cm for both a 2× (60 cm input → 30 cm SR output) and 4× (120 cm input → 30 cm SR output) enhancement, where both the fine details of the plane, and small neighboring objects can be accurately recovered. Recovering the plane at coarser resolutions is extremely difficult, particularly at 4.8 m with an 8× enhancement. In this case the input for the SR algorithm is 38.4 m GSD; at this resolution the satellite is simply insufficiently sensitive to resolve finer objects. Overall, we observe that when the imagery possesses fewer fine features to identify in coarser resolutions, algorithms are unable to hallucinate and recover all object types. A different algorithm such as a GAN may be able to hallucinate visually finer features, however previous studies [2] have shown that these algorithms are unable to exactly recover specific features of various object types.

Finally, in Figure 4 we demonstrate the visual enhancement provided by simulated 15 cm super-resolved output from both VDSR and RFSR. Both methods improve the visual quality by reducing pixelization and enhancing the clarity of features and characters. RFSR appears to produce slightly brighter edge effects than VDSR.

7.2. Object Detection Performance

For each model we compute mean average precision (mAP) performance on a 338-image test set spanning 6 continents (632 sq. km) at each resolution. Example precision recall curves are shown in Figure 6. The YOLT model is clearly superior to SSD, particularly for small objects.

Repeating the computation shown in Figure 6 for all models allows us to determine the degradation of perfor-

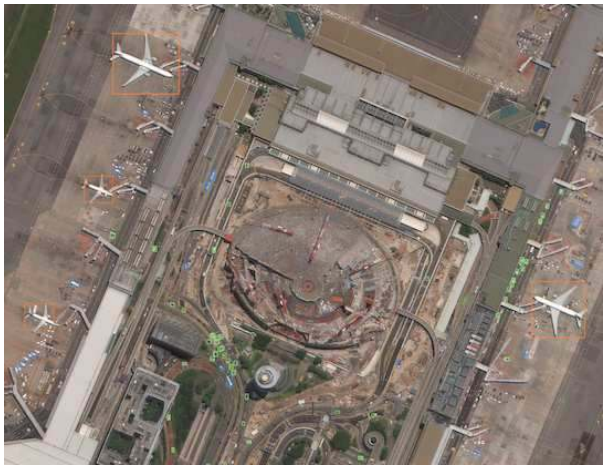
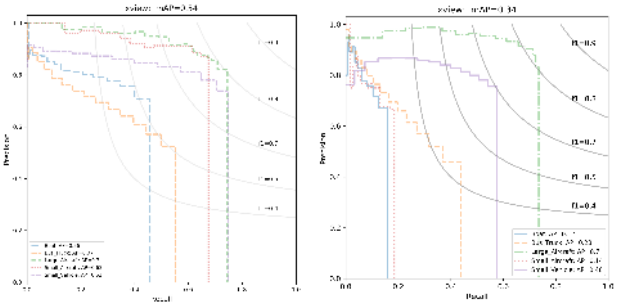


Figure 5: Example output of YOLT model at native 30 cm resolution. Cars are in green, buses/trucks in blue, and airplanes in orange.



(a) YOLT (b) SSD

Figure 6: Precision-recall curves for native 30 cm imagery for both YOLT and SSD.

mance as a function of resolution, as shown in Figure 7. In this plot we display 1σ bootstrap error bars for each model group. Results for SSD models are significantly worse than YOLT models, with a mAP of 0.30 at native 30 cm resolution. The YOLT model (mAP = 0.53) at this resolution is 77% better than SSD, which aligns fairly well with the findings of [37]. Ultimately, object detection performance decreases by $22 - 27\%$ when resolution degrades from 30 cm to 120 cm, and another $73 - 100\%$ from 120 cm to 480 cm when looking across broad object classes.

We also plot the results of the effects of $2\times$ super-resolution models when using both YOLT and SSD (Figures 8 and 9). When using YOLT, performance improvements are statistically significant only in the finest resolutions (Table 4) with comparable results for both VDSR and RFSR. In Figure 11 we show the change in mAP versus the original 30 cm and 60 cm imagery. The largest performance boosts can be seen when enhancing imagery from 30 cm to 15 cm ($+13\%$ vs 30 cm) and 60 cm to 15 cm ($14 - 20\%$ improvement vs 60 cm). Interestingly, enhancing imagery

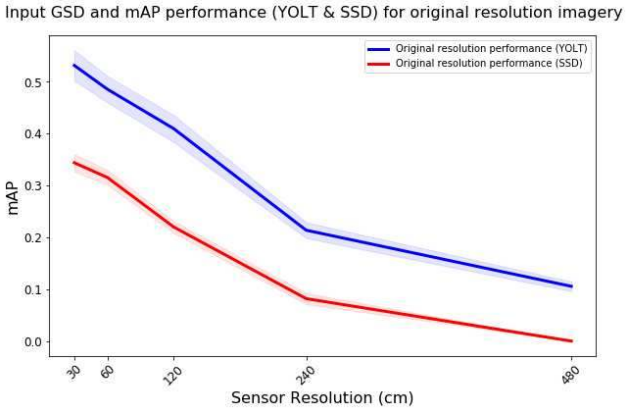


Figure 7: Performance of YOLT and SSD at the native sensor resolution for all object classes.

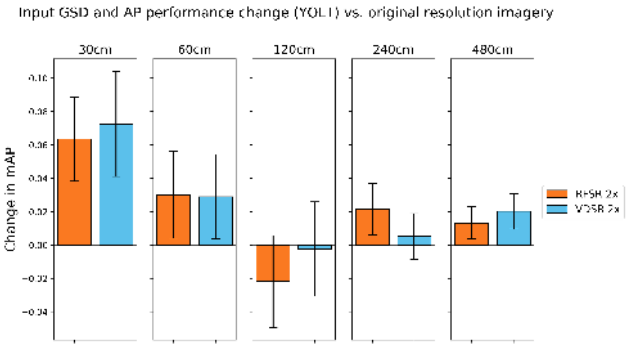


Figure 8: Performance change over original resolution (Figure 7-Blue Line) using YOLT and $2\times$ super-resolved data.

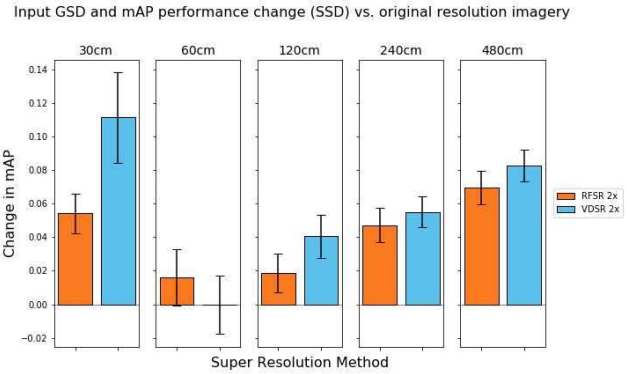


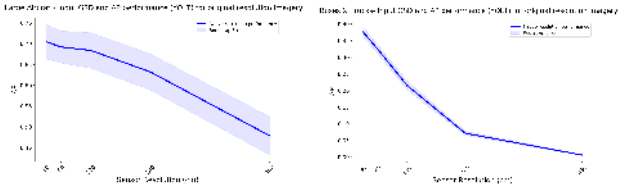
Figure 9: Performance change over original resolution (Figure 7-Red Line) using SSD and $2\times$ super-resolved data.

from 60 cm to 30 cm was much less effective than enhancing imagery from 60 to 15 cm. These findings showcase the value of super-resolution as a pre-processing step in these GSDs. Combined with a state of the art object detection framework, super-resolution has the ability to improve detection rates beyond what is possible with the best commercially available satellite imagery.

Furthermore, although performance is much worse with SSD, super-resolution techniques are much more effective.

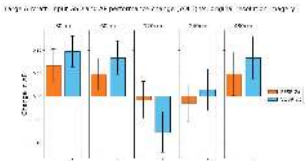
Model	Data	30 cm	60 cm	120 cm	240 cm	480 cm
YOLT	Native	0.53 ± 0.03	0.49 ± 0.03	0.41 ± 0.03	0.21 ± 0.02	0.11 ± 0.01
YOLT	RFSR 2×	0.60 ± 0.03 (+1.7σ)	0.52 ± 0.03 (+0.7σ)	0.39 ± 0.03 (-0.5σ)	0.24 ± 0.02 (+1.1σ)	0.12 ± 0.01 (+0.7σ)
YOLT	VDSR 2×	0.60 ± 0.03 (+1.7σ)	0.52 ± 0.03 (+0.7σ)	0.41 ± 0.03 (+0.0σ)	0.22 ± 0.01 (+0.4σ)	0.13 ± 0.01 (+1.4σ)
YOLT	RFSR 4×		0.56 ± 0.03 (+1.6σ)	0.40 ± 0.03 (-0.2σ)	0.23 ± 0.01 (+0.9σ)	0.12 ± 0.01 (+0.7σ)
YOLT	VDSR 4×		0.59 ± 0.02 (+2.8σ)	0.39 ± 0.03 (-0.5σ)	0.25 ± 0.02 (+1.4σ)	0.10 ± 0.01 (-0.7σ)
SSD	Native	0.30 ± 0.01	0.32 ± 0.01	0.22 ± 0.01	0.08 ± 0.01	0.00 ± 0.00
SSD	RFSR 2×	0.36 ± 0.01 (+4.2σ)	0.33 ± 0.02 (+0.4σ)	0.24 ± 0.01 (+1.4σ)	0.13 ± 0.01 (+3.5σ)	0.07 ± 0.01 (+7.0σ)
SSD	VDSR 2×	0.41 ± 0.03 (+3.5σ)	0.32 ± 0.02 (+0.0σ)	0.26 ± 0.01 (+2.8σ)	0.14 ± 0.01 (+4.2σ)	0.08 ± 0.01 (+8.0σ)

Table 4: Performance for each data type in mAP. For both RFSR and VDSR at each resolution we note the error and statistical difference from the baseline model (e.g. +0.5σ). The native sensor resolution of our original imagery and the input into the super-resolution models is shown on the X-axis. We then compare the super-resolved outputs vs. the original native imagery to test the change in object detection performance.

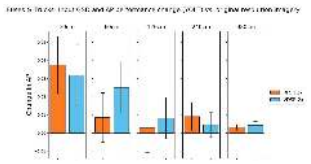


(a) Large Aircraft- Native Resolution Performance

(b) Buses/Trucks- Native Resolution Performance



(c) Large Aircraft - AP Change



(d) Buses/Trucks - AP Change

Figure 10: YOLT performance curves for Large Aircraft (a) and Buses/Trucks (b) at native resolution. Performance change versus these curves when super-enhancing imagery 2× (c and d).

With SSD, for both RFSR and VDSR performance boosts are evident for all resolutions, except for 60 cm to 30 cm. VDSR is generally shown to be slightly superior to RFSR when detecting objects with SSD. For SSD the improvement at 480 cm is statistically quite significant, though this is primarily due to the mAP of 0.0 for native imagery. Performance increases significantly once objects are greater than ≈ 20 pixels in extent. This trend extends across object classes, as shown in the performance curves for individual object classes (See Supplemental Material).

8. Conclusions

In this paper we undertook a rigorous study of the utility provided by super-resolution techniques towards the detection of objects in satellite imagery. We paired two super-resolution techniques (VDSR and RFSR) with advanced object detection methods and searched for objects in a satellite imagery dataset with over 250,000 labeled objects in a diverse set of environments. In order to establish super-resolution effects at multiple sensor resolutions, we degrade this imagery from 30 cm to 60, 120, 240, and 480 cm resolutions. Our baseline tests with both the YOLT and SSD models of the SIMRDWN object detection framework indicate that object detection performance decreases by $22 - 27\%$ when resolution degrades from 30 cm to 120 cm.

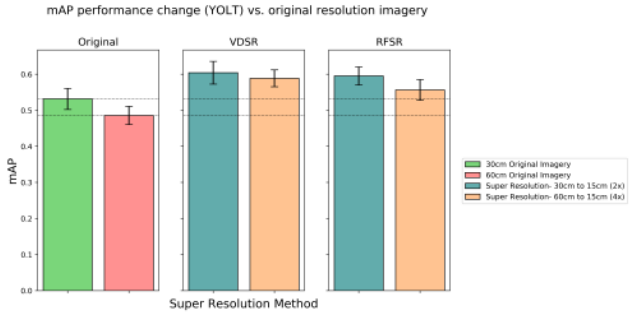


Figure 11: Performance boost of enhancing 30 and 60 cm imagery to 15 cm GSD.

The application of SR techniques as a pre-processing step provides an improvement in object detection performance at most resolutions (Table 4). For both object detection frameworks, the greatest benefit is achieved at the highest resolutions, as super-resolving native 30 cm imagery to 15 cm yields a $13-36\%$ improvement in mAP. Furthermore, when using YOLT, we find that enhancing imagery from 60 cm to 15 cm provides a significant boost in performance over both the native 30 cm imagery ($+13\%$) and native 60 cm imagery ($+20\%$). The performance boost applies to all classes, but is most significant for boats, large aircraft, and buses/trucks. Again with YOLT, in coarser resolutions (120 cm to 480 cm) SR provides little to no boost in performance (-0.02 to +0.04 change in mAP). When using SSD, super-resolving imagery from 30 to 15 cm provides a substantial boost for the identification of small vehicles ($+56\%$), but provides mixed results for other classes (See supplemental material). In coarser resolutions, with SSD, SR techniques provide a greater boost in performance however the performance for most classes is still worse compared to YOLT with native imagery. Overall, given the relative ease of applying SR techniques, the general improvement observed in this study is noteworthy and suggests SR could be a valuable pre-processing step for future object detection applications with satellite imagery.

References

[1] E. Bilgazyev, B. Efraty, S. Shah, and I. Kakadiaris. Sparse Representation-Based Super Resolution for Face Recognition At a Distance. In *Procedings of the British Machine Vision Conference 2011*, pages 52.1–52.11, Dundee, 2011. British Machine Vision Association. 1

[2] M. Bosch, C. M. Gifford, and P. A. Rodriguez. Super-Resolution for Overhead Imagery Using DenseNets and Adversarial Learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1414–1422, Mar. 2018. 2, 6

[3] L. Cao, C. Wang, and J. Li. Vehicle detection from highway satellite images via transfer learning. *Information Sciences*, 366:177–187, Oct. 2016. 1, 2

[4] G. Cheng and J. Han. A survey on object detection in optical remote sensing images. *CoRR*, abs/1603.06201, 2016. 3

[5] D. Dai, Y. Wang, Y. Chen, and L. Van Gool. Is Image Super-resolution Helpful for Other Vision Tasks? *arXiv:1509.07009 [cs]*, Sept. 2015. arXiv: 1509.07009. 1

[6] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016. 3

[7] C. Dong, C. C. Loy, K. He, and X. Tang. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, Feb. 2016. 2, 4

[8] Eric Shear. ULA preparing to launch WorldView-4 satellite from Vandenberg, Sept. 2016. 2

[9] M. Haris, G. Shakhnarovich, and N. Ukita. Deep back-projection networks for super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[10] M. Haris, G. Shakhnarovich, and N. Ukita. Task-Driven Super Resolution: Object Detection in Low-resolution Images. *arXiv:1803.11316 [cs]*, Mar. 2018. arXiv: 1803.11316. 1, 2

[11] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 1

[12] P. H. Hennings-Yeomans, B. V. Kumar, and S. Baker. Robust low-resolution face identification and verification using high-resolution features. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 33–36. IEEE, 2009. 1

[13] J. Huang and W. Siu. Practical application of random forests for super-resolution imaging. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2161–2164, May 2015. 4

[14] Z. Hui, X. Wang, and X. Gao. Fast and accurate single image super-resolution via information distillation network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[15] J. Kim, J. K. Lee, and K. M. Lee. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, Las Vegas, NV, USA, June 2016. IEEE. 2, 4

[16] J. Kim, J. K. Lee, and K. M. Lee. Deeply-Recursive Convolutional Network for Image Super-Resolution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, Las Vegas, NV, USA, June 2016. IEEE. 2

[17] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5835–5843, Honolulu, HI, July 2017. IEEE. 2

[18] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord. xView: Objects in Context in Overhead Imagery. *arXiv:1802.07856 [cs]*, Feb. 2018. arXiv: 1802.07856. 1, 3

[19] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, Honolulu, HI, July 2017. IEEE. 2

[20] L. Liebel and M. Krner. Single-Image Super Resolution for Multispectral Remote Sensing Data Using Convolutional Neural Networks. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B3:883–890, June 2016. 2

[21] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, Honolulu, HI, USA, July 2017. IEEE. 2

[22] H. Liu, Z. Fu, J. Han, L. Shao, and H. Liu. Single satellite imagery simultaneous super-resolution and colorization using multi-task deep neural networks. *Journal of Visual Communication and Image Representation*, 53:20–30, May 2018. 2

[23] 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. 2, 3

[24] 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. 3

[25] Y. Luo, L. Zhou, S. Wang, and Z. Wang. Video Satellite Imagery Super Resolution via Convolutional Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2398–2402, Dec. 2017. 2

[26] 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. 3

[27] 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. *Computer Vision ECCV 2016*, 9907:785–800, 2016. 1

[28] D. Pouliot, R. Latifovic, J. Pasher, and J. Duffe. Landsat Super-Resolution Enhancement Using Convolution Neural Networks and Sentinel-2 for Training. *Remote Sensing*, 10(3):394, Mar. 2018. 2

[29] 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. 3

[30] 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. 3

[31] 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. 5

[32] S. Schulter, C. Leistner, and H. Bischof. Fast and accurate image upscaling with super-resolution forests. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3791–3799, Boston, MA, USA, June 2015. IEEE. 4

[33] S. Shekhar, V. M. Patel, and R. Chellappa. Synthesis-based recognition of low resolution faces. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–6. IEEE, 2011. 1

[34] Y. Tai, J. Yang, and X. Liu. Image Super-Resolution via Deep Recursive Residual Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2798, Honolulu, HI, July 2017. IEEE. 2

[35] C. Tuna, G. Unal, and E. Sertel. Single-frame super resolution of remote-sensing images by convolutional neural networks. *International Journal of Remote Sensing*, 39(8):2463–2479, Apr. 2018. 2

[36] A. Van Etten. You Only Look Twice: Rapid Multi-Scale Object Detection In Satellite Imagery. *ArXiv e-prints*, May 2018. 3, 5

[37] A. Van Etten. Satellite Imagery Multiscale Rapid Detection with Windowed Networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, page In Press, Jan. 2019. arXiv: 1809.09978. 1, 3, 7

[38] A. Van Etten, D. Lindenbaum, and T. M. Bacastow. SpaceNet: A Remote Sensing Dataset and Challenge Series. *arXiv:1807.01232 [cs]*, July 2018. arXiv: 1807.01232. 1, 3

[39] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, Apr. 2004. 5

[40] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[41] A. Xiao, Z. Wang, L. Wang, Y. Ren, A. Xiao, Z. Wang, L. Wang, and Y. Ren. Super-Resolution for Jilin-1 Satellite Video Imagery via a Convolutional Network. *Sensors*, 18(4):1194, Apr. 2018. 2

[42] Y. Xu, L. Lin, D. Meng, Y. Xu, L. Lin, and D. Meng. Learning-Based Sub-Pixel Change Detection Using Coarse Resolution Satellite Imagery. *Remote Sensing*, 9(7):709, July 2017. 1, 2

[43] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2