Characterizing CNN-based Vessel Detection Algorithm Sensitivity to Optical Sensor Artifacts

1st John G. Warner
Naval Center for Space Technology
US Naval Research Laboratory
Washington DC, United States
john.warner@nrl.navy.mil

2nd Quinton Davidson
Naval Center for Space Technology
US Naval Research Laboratory
Washington DC, United States
quinton.davidson@nrl.navy.mil

3rd Michael Tietz
Naval Center for Space Technology
US Naval Research Laboratory
Washington DC, United States
michael.tietz@nrl.navy.mil

4th William Scharpf
Naval Center for Space Technology
US Naval Research Laboratory
Washington DC, United States
william.scharpf@nrl.navy.mil

5th Charles Keene
Naval Center for Space Technology
US Naval Research Laboratory
Washington DC, United States
charles.keene@nrl.navy.mil

***Abstract*—The number of space-based optical imaging systems have substantially grown in recent years. Many of these commercial based vendors are exploring concepts to move processing applications directly to the space segment or to incorporate advanced algorithms in their ground segment. Typically, Convolutional Neural Network (CNN)-based detection algorithms are considered due to their high reliability in laboratory settings. However, this class of algorithms routinely demonstrates degraded performance when unexpected phenomena are introduced to the data. As more near-real-time applications are considered, CNN-based algorithms will be subjected to sensor calibration artifacts. This paper seeks to characterize the response of CNN-based vessel detection algorithms to common electro-optical sensor calibration artifacts. Several artifacts are explored, including added sensor noise and failed detector artifacts. Each of the explored artifacts uniquely impact the algorithm detection performance. It is found that applying Poisson distributed random noise to the imagery substantially degrades model performance by increasing false alarm rate. Likewise, the application of a uniformly distributed random scale factor to the imagery degrades model performance by lowering positive detection rate. These results may inform plans to implement CNN-based algorithms directly in the space-segment by characterizing typical performance variations due to underlying sensor calibration issues.**

***Index Terms*—CNN, spacecraft, onboard processing, Automatic Target Recognition, maritime**

I. Introduction

The commercial satellite industry is continuing to experience rapid growth. Reductions in launch costs coupled with continued miniaturization of electronics has enabled rapid growth of satellite constellations [1] [2].

The emerging ability to field large constellations is continuing to open up new applications that may be served by space capabilities. One proposed application is the use of large constellations to monitor maritime activity. The vast area of the worlds oceans presents a challenge for detecting illicit maritime activities. However, large constellations of space-based sensors may be able to address this key challenge [3] [4] [5].

Image analysts have already identified challenges with maintaining pace with existing commercial satellite imaging systems. One means to grow analysis capability on par with the large growth of space-based sensors is to increasingly rely on Artificial Intelligence (AI) capabilities [6].

In order to successfully meet the challenge of detecting illicit maritime activities on a global scale via large constellations of space-based sensors, Automatic Target Recognition (ATR) methods, as proposed in Reference [7], would be necessary. ATR would minimize the burden on already taxed image analysts.

One key challenge to scaling ATR to large constellations is the ability to successfully understand operational quality metrics when constellations may have hundreds of sensors. Operators would be challenged to monitor what would be one or two orders of magnitude more sensors than what is typically deployed today.

A further complication is that many ATR methods lack explainability [8]. While a body of work has been devoted to understanding image quality for space-based sensors, see References [9], [10], and [11], substantially less work has been done to understand performance metrics for ATR methods and traceability to underlying sensor root causes, see Reference [12].

This paper seeks to narrow that knowledge gap by examining the relationship between maritime vessel detection ATR algorithm performance and typical factors impacting space-based Electro-Optical (EO) image performance. CNN-based vessel detection algorithms have been trained using vessels from existing commercial satellite imagery. Additional satellite images have been curated using manually applied vessel annotations. This imagery has been degraded using several methods that may occur during the life-cycle of space-

based imaging systems. Baseline performance of the ATR models using pristine imagery is compared to the performance using degraded imagery.

Section II details the methodology used. Results are presented in Section III. Conclusions are offered in Section IV.

II. Methodology

In this experiment, commercial satellite EO imagery is used to train a vessel detection model, as well as measure performance of the models subject to multiple types of image degradation that would be consistent with effects anticipated during a space-based sensor's life-cycle. Here, the imagery, the artifact process, and the ATR methods used are discussed.

A. Imagery Data Description

This experiment relies on EO imagery from the WorldView-2 (WV2) and WorldView-3 (WV3) commercial imaging satellites. References [13] and [14] contain additional details on the imaging systems.

WV2 and WV3 are high resolution imaging satellites that are in Low Earth Orbit (LEO). While these systems offer both panchromatic and multi-spectral image products, only the panchromatic products are used in this experiment. WV2 offers imagery with a Ground Sample Distance (GSD) as low as 50cm, and WV3 offers imagery with a GSD as low as 30cm. Both systems image in the visible spectrum.

The data used here is panchromatic EO data. The imagery has been segmented into approximately 5,000 by 5,000 pixel sub-images. The data have been down-sampled to 8-bit imagery, which is then normalized prior to use with the CNN-based models. The use of 8-bit imagery, rather than the native bit depth, which is typically higher, is used to match the bit depth of imagery used in typical CNN training architectures.

Data has been curated in order to train CNN-based ATR models. These data are subsections of the collected image that have been labeled as either containing a vessel or typical clutter within a maritime scene. 18,800 examples have been curated to identify whether the image contains a vessel or clutter. For context, Figure 1 provides examples of the typical vessel and clutter image classes.

(a) Vessel (b) Clutter

Fig. 1: Examples of typical vessel and clutter images. A typical commercial, tanker vessel is shown, along with a cloud/water mixture example. Original images ©MAXAR.

Further, independent data have been curated to measure the impact of imagery artifacts. Additional full-image examples have been manually annotated to denote vessels using the Label Studio software [15]. A quantity of 562 images has been labeled to measure model performance against. Of the total number of images, 442 are focused in open-ocean scenes, whereas the rest are focused in vessel-dense littoral areas, such as ports. Figure 2 shows an example of a vessel annotated in an image.

For evaluating ATR algorithms, these large images (about 5k by 5k pixels) are chipped into 224 by 224 samples using a stride length of 112 pixels, making the actual number of tested samples much larger. A more detailed summary of the test dataset can be found in Table I. The vessels in the annotated full-images are independent of the data used to train the ATR models (i.e. there are no overlapping images between the manually annotated data and the examples used to train the CNN-based models). As such, all the data in Table I is used for testing. As indicated by the number of positive chips relative to the total number of chips, this method of chipping in combination with the nature of maritime imagery lends itself to a highly imbalanced dataset. The distribution of vessels in a given image is particularly sparse in most cases that do not involve a congested port area. Even in images that do involve higher vessel-density, the area of land/water tends to be significantly greater than the area covered by vessels. This skewness in the data is taken into consideration when selecting evaluation metrics, see Section II-E.

TABLE I: Description of Test Data

Dataset	Images	Vessels	Chips	Positive Chips
Complete	562	2309	1380027	25828
Open Ocean	442	1365	1090901	20162
Littoral	120	944	289126	5666

B. Modeling Image Calibration Artifacts

A large body of work is dedicated to the calibration of satellite imaging systems [16] [17] [18] [19] [20] [21] [9].

Examples are taken from the literature as to typical imaging payload issues that may occur during on-orbit operations. The following subsections provide detail on each of these typical issues. These issues have been simulated for this experiment by applying the conditions detailed below to the pristine satellite imagery in order to characterize the impacts on ATR algorithm performance.

1) Non-uniformity Correction Error: Digital imaging sensors, such as the typical Charged Coupled Device (CCD) sensor used in satellite imaging payloads, suffer from Fixed Pattern Noise (FPN). FPN is typically present due to small variations in photoelectron conversion for each pixel in the sensor. A Non-uniformity Correction (NUC) process is typically applied as part of payload calibration [9].

In the context of large, imaging satellite constellations, operators may be challenged to monitor NUC calibration performance across a large number of sensors especially as

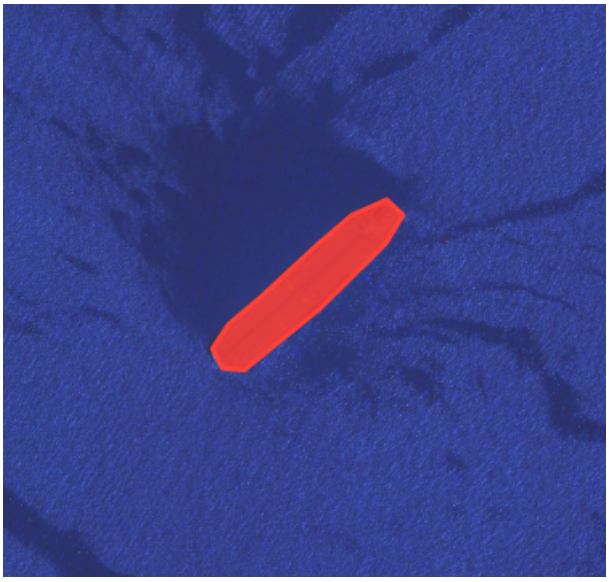


Fig. 2: Image of Typical Annotation of Vessel in Commercial Satellite Imagery. A polygon is used to represent the all pixels associated with the vessel object. Original Image ©MAXAR.

imaging payloads age and may become more difficult to calibrate. Various levels of NUC error are applied to emulate anticipated effects of either poor system calibration or imaging payload life-cycle degradation.

We simulated this bad non-uniformity correction error by multiplying each image pixel with a random scaling factor. For a given input image i, our output bad NUC image was defined in Equation 1.

$$I(x,y) = (\mu + \lambda * \beta(x,y)) * i(x,y) \tag{1}$$

where μ was set statically at 1 and λ was tested for the cases 0.01, 0.1, and 0.3. Here, $\beta(x,y)$ is a random number sampled from a normal Gaussian distribution.

2) Offset Correction Error: Most imaging sensors (CCD and CMOS) have a dark response or current due to current leakage in the photo-sensing diode. This signal does not depend on the light level of the scene, but only depends on the integration time used to collect the image. This is usually assessed by placing a shutter or cover over the imaging system and measuring the dark response for each pixel. Along with the NUC, the offset will provide a mapping between the reported Digital Number (DN) per pixel and the measured light level.

In the context of a large constellation, the dark level can change and grow with radiation effects and other on orbit aging effects. It would be important to understand these life-cycle related impacts on ATR performance.

For this analysis, the bad offset error is simulated by adding a Poisson distributed random number to each pixel in the image. Given an input image i, the output image can be defined as given in Equation 2.

$$I(x,y) = i(x,y) + P((x,y);\lambda) \tag{2}$$

where the random number is sampled from the Poisson distribution defined as

$$P\left((x,y);\lambda\right) = \frac{e^{-\lambda}\lambda^{x}}{x!} \tag{3}$$

This distribution $P((x,y);\lambda)$ describes the probability of events happening over the image sized (x,y) within the observed interval λ. For our analysis, we tested the effects of bad offset error for λ equaling 10, and 30.

3) Bad Pixels Error: Digital imaging sensors during the course of operations may suffer from inoperable or bad pixels. These pixels may suffer from a combination of poor quantum efficiency, high dark current, excessive read noise, or non-linearity due to a range of faults within the sensor electronics [22].

While sensor calibration efforts strive to remove the impact of bad pixels on image processing, the impacts may nevertheless be realized. Any calibration effort must choose an operational cadence, thus leaving some period of time between when an artifact materializes and when it may be mitigated.

The specific manifestation of bad pixels for any imaging system is dependent on the underlying image sensor technology, the sensor calibration process, and the image formation process.

For purposes of this analysis, a pseudo-random sequence of pixel (x,y) coordinate pairs was generated once. The pixel value for each coordinate pair was set to zero for all images. This approach has the benefit of traceability of ATR results back to the bad pixel artifact, since the same set of bad pixels is used for all images.

4) Failed Region Error: Digital imaging sensors during the course of operations may suffer from larger, accumulated faults that leave entire regions of the image inaccessible for image processing [22].

Again, sensor calibration efforts would strive to remove these failed regions from producing image products; however, operational cadence issues may leave some period of time between when a failed region materializes and when it may be mitigated.

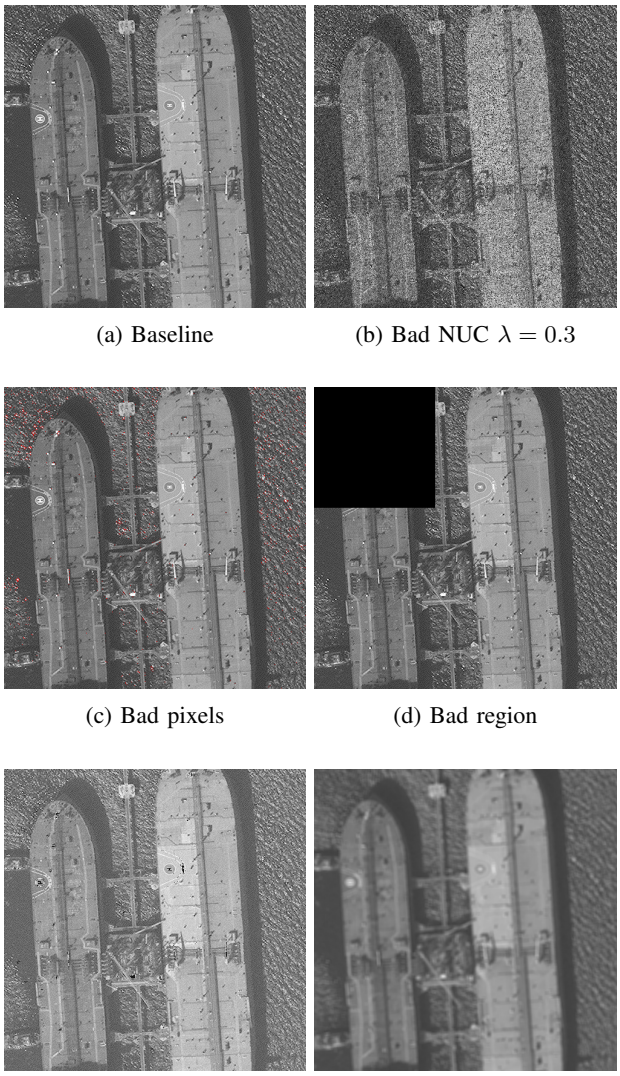
The specific manifestation of a failed region for any imaging system is dependent on the underlying image sensor technology, the sensor calibration process, and the image formation process.

For this analysis, the failed region artifact is emulated by setting the central 25% of pixels to zero.

5) Gaussian Blur: Earth imaging sensors may require focus calibration to ensure resulting imagery is in focus. While specific focus behavior and calibration techniques are directly related to the optical design of the satellite, there is typically a period of several months while the focus is being updated to account for residual settling of optical system components [23].

As such, it is worthwhile to understand the impact of defocused imagery on ATR performance. In order to emulate defocused imagery, a Gaussian blur filter was applied to the baseline imagery. A 7 pixel kernel is used when applying the Gaussian blur filter.

6) Artifact Examples: Figure 3 shows examples of each artifact on a sample image.



(a) Baseline (b) Bad NUC $\lambda = 0.3$

(c) Bad pixels (d) Bad region

(e) Bad offset $\lambda = 30.0$ (f) Gaussian blur

Fig. 3: Labeled examples of different anomalies on sample image. The visual impact of each of the modeled artifacts can be observed. Note, the example bad pixels image is colorized red for viewing purposes only. Original imagery ©MAXAR

C. Automatic Target Recognition Models

The ATR model used here is based on the Inception V1 Convolutional Neural Network (CNN) architecture [24]. The model was trained using an augmented dataset of 18,800 original images and random initial weights to classify images as either binary ship or not-ship. The dataset consisted of 9400 ship images and 9400 not-ship images and had an 80-20 train test split. The model was trained for 40 epochs using a batch size of 50. The ADAM optimizer was used with an initial learning rate of 0.001 that decreased to 0.0001 after 10 epochs. Cross entropy was used as the loss function. The achieved classification accuracy on the validation dataset was 91.1%.

The Inception V1 model was chosen due to internal heritage. While more modern CNN approaches exist, the purpose of the experiment is to measure changes relative to a baseline performance, so a model with sufficiently reasonable performance is needed. Future work should incorporate additional model architectures.

In order to evaluate the ATR model against ground truth annotations, softmax scores were generated for chips derived from each Worldview image. A chip size of (224,224) pixels and a stride length of 112 pixels was used to process each image. This strategy resulted in roughly 2000 (chip, softmax) pairs for each 5000 by 5000 image. For metrics requiring a binary classification, these softmax scores were thresholded on a value of 0.75. Chips corresponding to a softmax greater than or equal to 0.75 were assigned 1 (corresponding to a vessel detection) and those below were assigned 0 (corresponding to no vessel detected). For calculating metrics, binary masks were created using the ground truth vessel annotations. Areas inside polygon vessel annotations were set as 1 (corresponding to positive vessel ground truth) and everything else to 0 (corresponding to negative vessel ground truth). The (chip, softmax) pairs were then compared to their corresponding chip in the binary mask, with each ground truth chip that intersected a vessel polygon (i.e. containing nonzero pixels) counting as 1 for positive ground truth and the rest as 0 for negative ground truth.

D. Experimental Comparisons

Several cases are examined to characterize the impact of imagery artifacts on CNN-based automatic target recognition performance. Table II defines the cases examined as well as the mnemonic used to identify the cases in the Section III.

TABLE II: Table of Experimental Cases Examined

Case Label	Artifact Applied	Parameter Value
Baseline	None	None
NUCsig01	Bad NUC	$\lambda = 0.01$
NUCsig1	Bad NUC	$\lambda = 0.1$
NUCsig3	Bad NUC	$\lambda = 0.3$
OS10	Offset Correction	$\lambda = 10.0$
OS30	Offset Correction	$\lambda = 30.0$
Blur	Gaussian Blur	(7,7) kernel

E. Performance Metrics

A number of metrics are used to characterize the difference between baseline ATR performance and ATR performance with calibration artifacts applied to the underlying imagery. These metrics are calculated from samples on a per World-View image basis and compared across the various artifact scenarios. The metrics are calculated across various percentiles of the entire population of test images to capture the median performance as well as the performance variance.

The cross entropy function is shown in Equation 4. Here, y is the value of softmax output from the CNN, and $\hat{y}$ is the binary value of the corresponding chip ground truth (one indicating vessel, zero indicating clutter). This function

provides a useful metric as it can characterize how far target values are above the background output level. Lower values in cross entropy denote a distribution that more closely matches the ground truth distribution [25]. These can indicate a higher performance, however due to the skewness of the data evaluated here, they more strongly describe a larger number of non-vessel predictions.

$$\frac{1}{N}\sum -y\log\hat{y} - (1-y)\log(1-\hat{y}) \tag{4}$$

Additional metrics are based on the ATR classification decision given a specific applied threshold value for the softmax output. Here, a threshold of 0.75 is used where values above the threshold are designated as a target (e.g. a value of one), and values below the threshold are designed as clutter (e.g. a value of zero). Detection threshold values are typically a key design parameter. Here, a static value is chosen as a basis of comparison between the baseline and anomaly cases. Since the key metrics are measured relative to the baseline, this value was not subject to further optimization.

A confusion matrix may be assembled to group chip samples into the four outcomes given a binary classification: True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN). These comprise the possible outcomes of whether the ATR output of ship or clutter matches the truth mask of ship or clutter. See Reference [26] for additional discussion on binary classification metrics.

From here, various metrics may be constructed to summarize performance. Again, due to the disproportionate number of negative samples evaluated, the metrics chosen for this analysis avoid TN samples that would otherwise strongly skew them.

The True Positive Rate (TPR), or recall, is given in Equation 5. The TPR denotes what percentage of true vessel samples were correctly identified. A higher number denotes better performance.

$$TPR = \frac{TP}{TP+FN} \tag{5}$$

The precision is given in Equation 6. The precision measures the percentage of samples correctly identified as vessels over all samples identified as vessels. A higher number denotes better performance.

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

The F_1 score is given in Equation 7. The F_1 score measures the harmonic mean of precision and recall and is a measure of accuracy. A F_1 score of 1.0 indicates that both precision and recall are 1.0 while a F_1 score of 0.0 indicates that either precision or recall are 0.0. As such, a higher number indicates better performance.

$$F_1 = \frac{2}{recall^{-1} + precision^{-1}} \tag{7}$$

Last, these metrics are calculated for each of the 562, larger images. As ATR performance can vary widely with specific image content, these metrics are compiled at various percentiles to show both the median performance and the degree of performance variance.

III. Results

Results for the experimental cases run are summarized here. All imagery (1380027 chips, see Table I) was processed using an NVIDIA A30 GPU in 3 hours, 31 minutes. The baseline performance metrics for the Inception V1 based model are given in Table III below. As can be seen, the model performs reasonably well across a broad population of test images. Recall values are lower than typically expected from a binary classifier because ground truth image chips are considered positive if any portion of a vessel intersects them regardless of area. For the purposes of the analysis, metrics are measured relative to the baseline case, so further work to refine this is not necessarily informative to characterizing the impact of anomalies.

Table IV shows the performance metrics for the Inception V1 based model for the subset of imagery in open ocean. As can be seen, the model shows improved precision, F_1 score, and cross entropy when littoral imagery is excluded. These improvements are driven by a high false alarm rate on land chips, indicating the model struggles with identifying littoral features, such as land and port infrastructure, as vessels. This could be expected as the 18,800 image dataset that the ATR algorithm was trained on largely included open ocean and clutter in the not-ship category, so it hadn't seen many examples of these littoral features. Overall anomaly-based trends in algorithm performance remain the same regardless of dataset, so as to avoid further skew from disproportionate false positives, analysis focuses solely on open ocean images.

TABLE III: Baseline Performance Metrics for Inception V1 Based Model for All Imagery Across Various Percentiles of Population

Metric	10th P.	25th P.	Median	75th P.	90th P.
F1	0.0415	0.1793	0.4749	0.8000	0.9427
Precision	0.0376	0.2685	0.8000	1.0000	1.0000
TPR	0.0592	0.2359	0.6667	0.8588	1.0000
Cross Entropy	0.0359	0.0832	0.1636	0.3741	0.6859

TABLE IV: Baseline Performance Metrics for Inception V1 Based Model for Open Ocean Imagery Across Various Percentiles of Population

Metric	10th P.	25th P.	Median	75th P.	90th P.
F1	0.0656	0.2611	0.6259	0.8421	0.9615
Precision	0.2708	0.6667	0.9474	1.0000	1.0000
TPR	0.0364	0.2000	0.6095	0.8397	1.0000
Cross Entropy	0.0321	0.0663	0.1346	0.2487	0.4218

Table V shows the cross entropy results by artifact case given the Inception V1 model at various percentiles across

the open-ocean test imagery. Interestingly, the NUCsig3 case shows improved cross entropy from the baseline across the distribution of imagery. In the OS30 case, the cross entropy is uniformly worse across the distribution of cases from the baseline.

TABLE V: Comparison of Cross Entropy Results for Inception V1 Based Model for Open Ocean Imagery Across Various Percentiles of Population

Case	10th P.	25th P.	Median	75th P.	90th P.
Baseline	0.0321	0.0663	0.1346	0.2487	0.4218
Bad Pixel	0.0321	0.0663	0.1346	0.2487	0.4218
Bad Region	0.0330	0.0677	0.1331	0.2514	0.4137
NUCsig01	0.0308	0.0671	0.1399	0.2599	0.4234
NUCsig1	0.0288	0.0729	0.1462	0.2428	0.3763
NUCsig3	0.0245	0.0472	0.1114	0.2020	0.2916
OS10	0.0591	0.0971	0.1836	0.4637	1.0923
OS30	0.0576	0.0980	0.1667	0.3739	1.1134
Blur	0.0421	0.0784	0.1393	0.4107	0.7073

Figures 4–7 show the performance of each anomaly model for each metric. Each of the figures shows a box and whisker plot, where the mean value is shown as an orange horizontal line, the extents of the box are at first and third quartiles of the population, and the extents of the whiskers are at the 10^{th} and 90^{th} percentiles of the population.

As can be seen in Figure 4, model F_1 score can vary substantially across the applied image artifacts. The median F_1 score in each population sees a decrease from the baseline, and in the case of the NUC and OS artifacts, larger degrees of scaling factor/noise cause a larger shift in the median. At the same time however, the upper and lower percentiles of any anomaly distribution are much more resistant to the performance effects of its anomaly. NUCsig3 and blur see the largest overall decreases in F_1 score, while bad region and NUCsig01 see the smallest. Bad pixel is entirely unaffected.

Figure 5 shows the precision results for each of the modeled cases. Both NUCsig3 and blur saw increases in median precision, although the lower quartile for blur dropped by 0.2. In the case of the NUC artifact, median and lower quartile precision decrease when a small scaling factor is applied (NUCsig01, NUCsig1) but actually increase from baseline when a larger one is applied (NUCsig3). The smaller degrees of scaling cause weakly confident vessel predictions to become weakly confident non-vessel predictions, and the decline in true positives moves precision down. When the larger degree of scaling is applied, all weakly confident detects decline leaving only the very confident true positives. Precision increases in turn.

Figure 6 shows the cross entropy results for each of the modeled cases. Interestingly, the NUCsig3 case shows an improvement across the test image cases. Both the OS10 and OS30 cases show substantially worse cross entropy performance. Given the skewness of the data in favor of negative samples, this suggests that the OS artifact is largely increasing the number of overall vessel detects while the NUC artifact is largely doing the opposite. This matches the changes in precision and TPR discussed previously.

Figure 7 shows the TPR for each of the modeled cases. OS10 and OS30 both see slight increases in performance here, while NUCsig3 sees a more significance decrease, particularly in the median case. In the cases of bad NUC, it takes a large scaling factor to cause a precipitous decline in TPR, as NUCsig01 and NUCsig1 are very close to the baseline. In combination with precision, these TPR results would suggest that the bad NUC artifact decreases likelihood of a sample chip being detected as a vessel while bad offset increases it. This is in agreement with the trends in cross entropy.

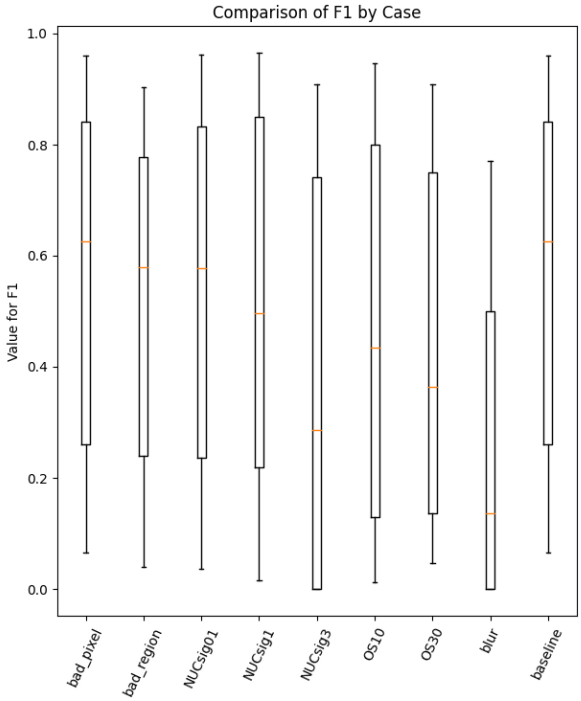


Fig. 4: Box Plot of F_1 Score by Case. Median decreases for every artifact except bad pixel.

IV. Conclusions

In order to characterize the impact of various EO sensor calibration artifacts on the ability to detect vessels in imagery, nine cases were examined across a set of sample imagery. Baseline model performance is shown and found to meet the expected relatively high levels of performance, showing it is a suitable model for examining the impacts of the artifacts identified.

The baseline model performance against pristine imagery was compared to model performance given various artifacts applied to the imagery. A subset of imagery in open-ocean was specifically examined knowing the chosen CNN-based model struggles misidentifying land features in littoral areas as vessels.

Notably, the application of the NUC artifact with a λ value of 0.3 was shown to decrease overall probability of predicting

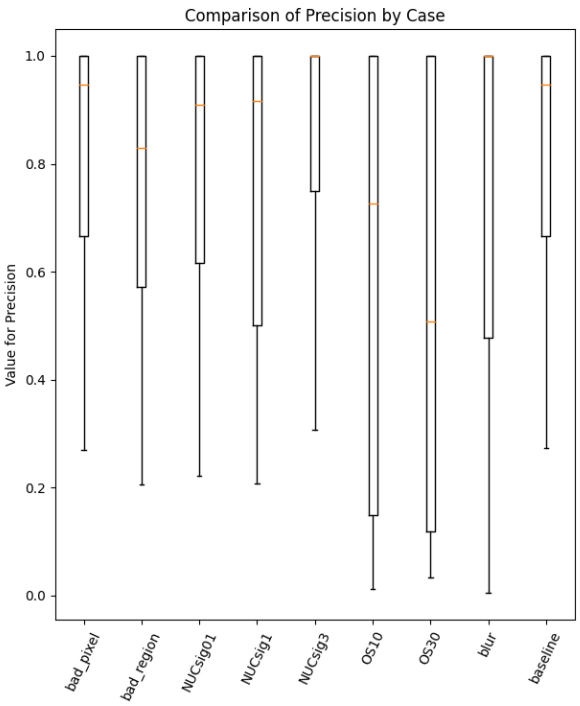


Fig. 5: Box Plot of Precision Metric by Case. Note the decrease in precision for OS cases.

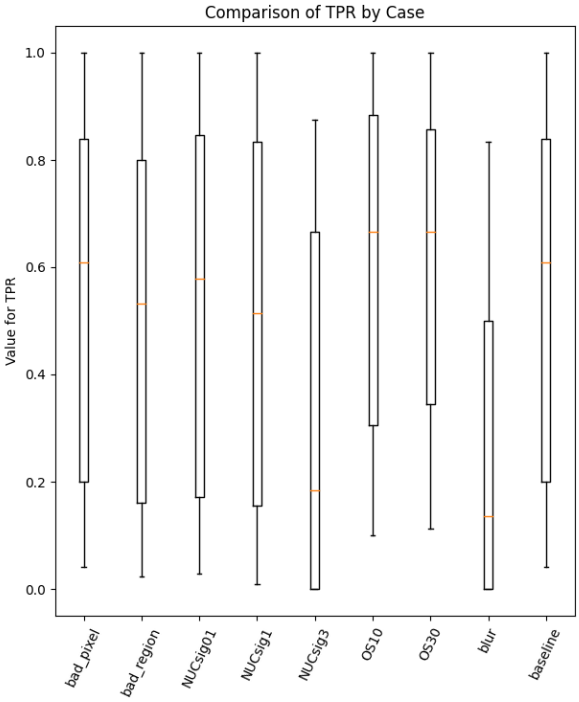


Fig. 7: Box Plot of TPR Metric by case. Note the decrease in TPR for the NUCSig3 and Blur cases.

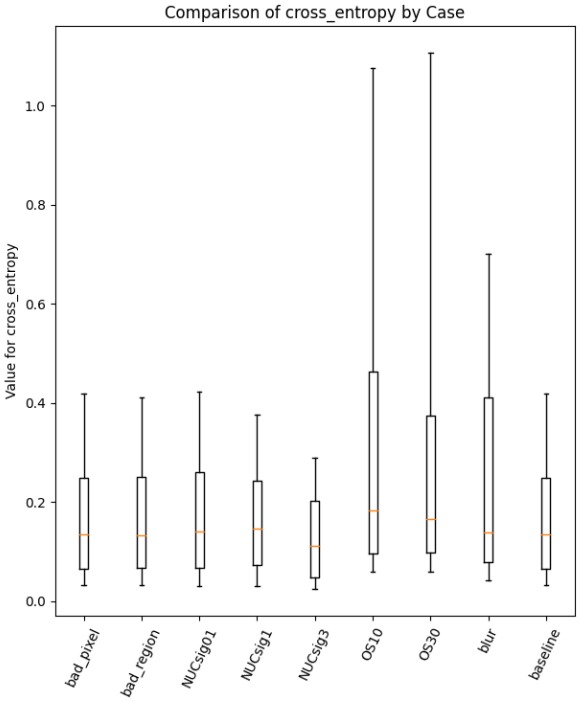


Fig. 6: Box Plot of Cross Entropy Metric by Case. Note the improvement in the NUCSig3 case, while the OS artifact cases show lower performance.

a vessel as measured by the cross entropy, F_1 score, and precision. The application of the OS artifact with a λ value of 10.0 was shown to likewise increase probability of predicting a vessel across the same range of metrics. Otherwise the NUC and OS artifact effects varied as the λ parameter varied.

The Gaussian blur artifact was shown to decrease model performance for the vast majority of cases. This makes sense as filtering out high spatial frequency information from the imagery likely removes features, like ship edges, that would otherwise be used in vessel classification.

Both the bad pixel and bad region artifacts show a minimal difference from the baseline model performance. The lack of deviation for bad region artifacts can likely be attributed to the data examined in this paper. For bad region anomaly injection, a region in the center of the image is dropped to 0 intensity. Given that low clutter-density open water chips produce low softmax scores for the model evaluated, the output softmax map should be unaffected by this anomaly if vessels are not specifically located within the center region. On a per image basis, bad region anomaly degrades performance in the scenario with vessels present. However, the dataset examined has vessels distributed across the entire areas of its images, so as a collection, performance minimally decreases. In the case of bad pixel, it's likely that the density of dropped pixels across the image was not enough to meaningfully alter the features being extracted by the chosen ATR algorithm. Repeating this experiment while varying the percentage of pixels being ran-

domly dropped could be interesting in identifying a threshold at which performance reasonably depreciates.

The application of the NUC artifact with λ 0.3 increased precision at the cost of lowered recall. Overall the total number of positive detects decreased with the ones remaining being high confidence classifications, a similar effect to raising the softmax threshold used for the binary classification. Further work should seek to examine how this effect changes with adjustments to the static threshold for binary classification.

This raises the question as to whether applying the NUC artifact can in some circumstance improve CNN-based model performance. Future work should seek to better characterize this relationship. It may be feasible to either generically improve performance, or at least improve clutter rejection using similar methods to the applied NUC artifact.

Future work may also seek to understand whether these typical artifacts may be mitigated by accounting for them during CNN training.

Last, additional future work should also examine whether these effects remain consistent across CNN architectures, or whether each specific architecture is impacted differently. This same experiment may be repeated with multiple model architectures to determine whether these effects vary as a function of the model architecture.

Overall, each imagery artifact examined showed varied impacts upon CNN-based model performance. The demonstrated range of impacts was substantial. As such, the impacts of calibration artifacts on satellite mission performance should be considered prior to operations.

V. Acknowledgment

This work was funded via the US Naval Research Laboratory Base Program funding. The authors would like to acknowledge their funding sponsors as well as the view made possible by standing atop the shoulders of those who came before them.

References

[1] A. C. Boley and M. Byers, "Satellite mega-constellations create risks in low earth orbit, the atmosphere and on earth," *Scientific Reports*, vol. 11, no. 1, p. 10642, 2021.

[2] J. R. Behrens and B. Lal, "Exploring trends in the global small satellite ecosystem," *New Space*, vol. 7, no. 3, pp. 126–136, 2019.

[3] S. C. Boraz, "Maritime domain awareness: Myths and realities," *Naval War College Review*, vol. 62, no. 3, pp. 137–146, 2009.

[4] N. P. Bannister and D. L. Neyland, "Maritime domain awareness with commercially accessible electro-optical sensors in space," *International Journal of Remote Sensing*, vol. 36, no. 1, pp. 211–243, 2015.

[5] D. M. Kocak and P. Browning, "Real-time ais tracking from space expands opportunities for global ocean observing and maritime domain awareness," in *OCEANS 2015-MTS/IEEE Washington*. IEEE, 2015, pp. 1–7.

[6] R. Sharp, "Remarks delivered by the director of the national geospatial-intelligence agency," Delivered at the GEOINT 2019 Symposium, June 2019, https://www.nga.mil/news/GEOINT_2019_Symposium.html.

[7] S. Sabogal, A. George, and G. Crum, "Recon: A reconfigurable cnn acceleration framework for hybrid semantic segmentation on hybrid socs for space applications," in *2019 IEEE Space Computing Conference (SCC)*. IEEE, 2019, pp. 41–52.

[8] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1424, 2021.

[9] R. D. Fiete, *Modeling the imaging chain of digital cameras*. SPIE press Bellingham, Washington, 2010.

[10] M. Crespi and L. De Vendictis, "A procedure for high resolution satellite imagery quality assessment," *Sensors*, vol. 9, no. 05, pp. 3289–3313, 2009.

[11] S. Saunier, G. Karakas, I. Yalcin, F. Done, R. Mannan, C. Albinet, P. Goryl, and S. Kocaman, "Skysat data quality assessment within the edap framework," *Remote Sensing*, vol. 14, no. 7, p. 1646, 2022.

[12] N. Drenkow, N. Sani, I. Shpitser, and M. Unberath, "A systematic review of robustness in deep learning for computer vision: Mind the gap?" *arXiv preprint arXiv:2112.00639*, 2021.

[13] N. T. Anderson and G. B. Marchisio, "Worldview-2 and the evolution of the digitalglobe remote sensing satellite constellation: introductory paper for the special session on worldview-2," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII*, vol. 8390. SPIE, 2012, pp. 166–180.

[14] N. Longbotham, F. Pacifici, S. Malitz, W. Baugh, and G. Camps-Valls, "Measuring the spatial and spectral performance of worldview-3," in *Hyperspectral Imaging and Sounding of the Environment*. Optica Publishing Group, 2015, pp. HW3B–2.

[15] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label Studio: Data labeling software," 2020-2022, open source software available from https://github.com/heartexlabs/label-studio. [Online]. Available: https://github.com/heartexlabs/label-studio

[16] M. Dinguirard and P. N. Slater, "Calibration of space-multispectral imaging sensors: A review," *Remote Sensing of Environment*, vol. 68, no. 3, pp. 194–205, 1999.

[17] P. N. Slater, S. F. Biggar, J. M. Palmer, and K. J. Thome, "Unified approach to pre-and in-flight satellite-sensor absolute radiometric calibration," in *Advanced and Next-Generation Satellites*, vol. 2583. SPIE, 1995, pp. 130–141.

[18] M. Dinguirard and P. Henry, "Calibration of the spot hrv cameras," *Remote Sens. Environ.*, 1993.

[19] G. Blanchet, L. Lebegue, S. Fourest, C. Latry, F. Porez-Nadal, S. Lacherade, and C. THIEBAUT, "Pleiades-hr innovative techniques for radiometric image quality commissioning," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 39, p. B1, 2012.

[20] K. S. Krause, "Worldview-1 pre and post-launch radiometric calibration and early on-orbit characterization," in *Earth Observing Systems XIII*, vol. 7081. SPIE, 2008, pp. 338–348.

[21] M. A. Kuester, M. Ochoa, A. Dayer, J. Levin, D. Aaron, D. L. Helder, L. Leigh, J. Czapla-Meyers, N. Anderson, B. Bader *et al.*, "Absolute radiometric calibration of the digitalglobe fleet and updates on the new worldview-3 sensor suite," in *Report of JACIE Civil Commercial Imagery Evaluation Workshop of DigitalGlobe Inc.; DigitalGlobe Inc.: Westminster, CO, USA*, 2017.

[22] R. M. Smith, D. Hale, and P. Wizinowich, "Bad pixel mapping," in *High Energy, Optical, and Infrared Detectors for Astronomy VI*, vol. 9154. SPIE, 2014, pp. 372–387.

[23] D. Mulawa, "On-orbit geometric calibration of the orbview-3 high resolution imaging satellite," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*, vol. 35, pp. 1–6, 2004.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[25] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[26] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, rOC Analysis in Pattern Recognition. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786550500303X