Using Temporal Contrast to Detect Small Vessels in Low Resolution Optical Satellite Imagery

WenXin Dong
Stanford University
wxd@stanford.edu

Abstract

Automatic vessel detection in satellite imagery can facilitate the detection and regulation of illegal, unreported and unregulated (IUU) fishing. However, detecting vessels from optical satellite imagery is a challenging task due to the low-resolution of the satellite imagery and the small size of the vessels. Small vessels are easily confounded with sun glare, cloud, and reef. In this project, we aim to improve vanilla Faster R-CNN by making use of temporal contrast. Specifically, we slightly modify the Faster R-CNN model to intake both the target image and a reference image, and produce bounding boxes for the target image. Our temporal contrast method lead to improvement in precision compared to non-temporal baselines.

1. Introduction

Small vessels detection on low resolution imagery is a difficult task because small vessels are 15m - 30m in length and low resolution satellite imagery typically have resolution less or equal to 3m per pixel. This means that small vessels will be only occupy a few pixels, and could be easily confounded with sun glare, cloud, and reef. Some examples are shown in Figure 2.

This is why having temporal contrast of the location can facilitate vessel detection. Small islands and reefs are unlikely to change location overtime, and should be less frequently misclassified as vessels when a second image is provided as reference.

We formulate our problem as the following. We are given a set of images $I^{T1} = \{I_1^{T1}, I_2^{T1}, ..., I_n^{T1}\}$ and the corresponding vessel bounding boxes, $B = \{B_1^{T1}, B_2^{T1}, ..., B_n^{T1}\}$, where each B_i is a list of bounding boxes. Also given is a set of reference images $I^{T2} = \{I_1^{T2}, I_2^{T2}, ..., I_n^{T2}\}$, where each pixel in I_i^{T1} and I_i^{T2} share the exact same geographical coordinates, and I_i^{T1} and I_i^{T2} come from random (but different) timestamps. We want to learn to predict B^{T1} using I^{T1} and I^{T2}.

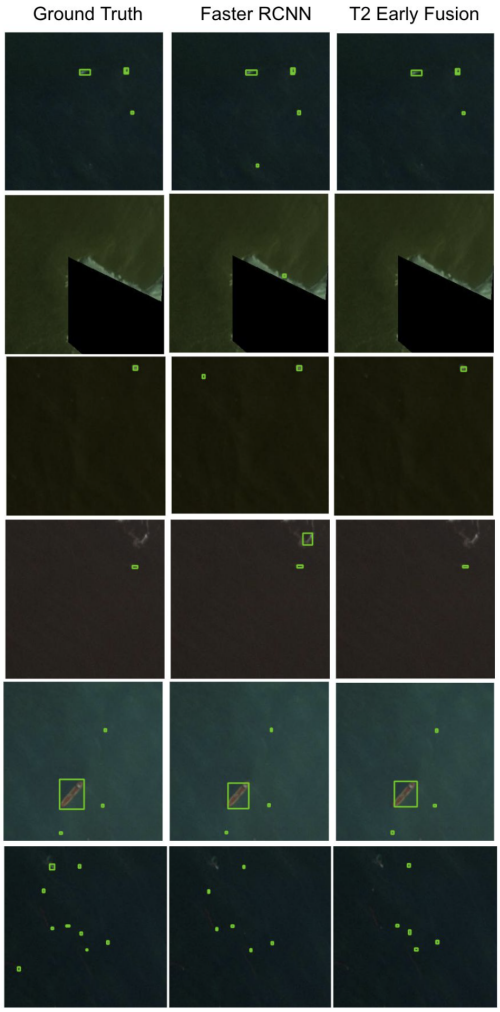

Figure 1. Comparison between Faster R-CNN and our proposed T2 Early Fusion method. T2 Early Fusion can more easily distinguish small vessels from ocean objects such as reef and wave, leading to improvement in precision.

We use Faster R-CNN [7] as our baseline model. We experiment with two temporal contrast methods: Early Fu-

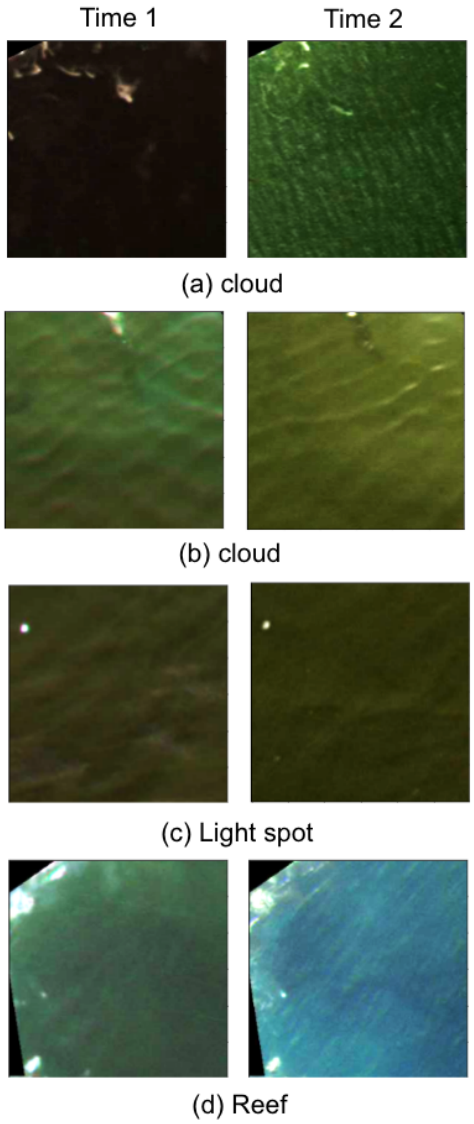

Figure 2. Examples of target reference image pair from the dataset showing objects on the sea surface that could be easily confused with vessels. Having a reference image helps to eliminate confusion.

sion and Late Fusion. In Early Fusion, we input 6-channel images to Faster R-CNN by stacking I_i^{T1} and I_i^{T2} channel-wise. In Late Fusion, we use an aggregated feature map by running the backbone on I_i^{T1} and I_i^{T2} independently. We compare our methods against several baselines and observe improvement in precision compared to non-temporal baseline. In error analysis in section 5, we show that while our fusion models are less likely to confuse vessel-size objects with vessels in general, they are prone to be misguided by noisy temporal features such as shifted wave highlights.

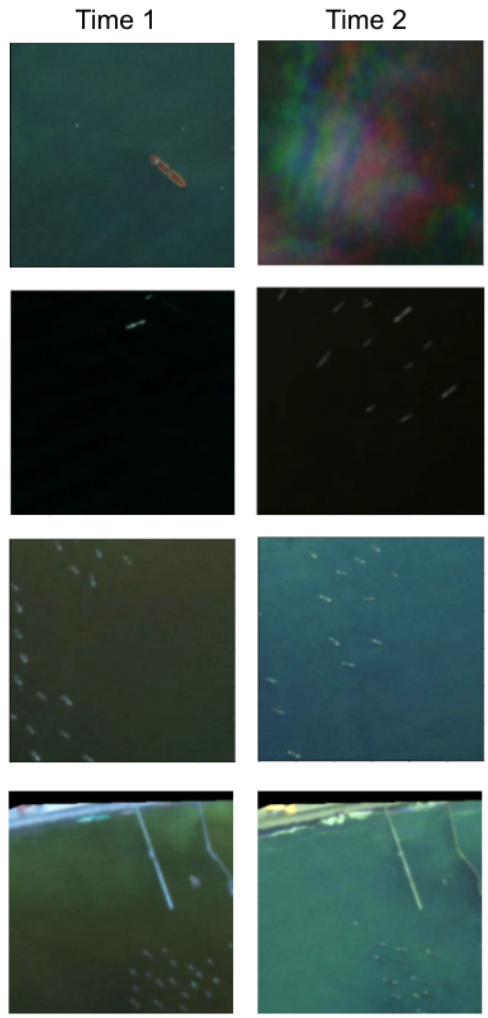

Figure 3. Examples of target reference image pair from the dataset. As shown by some examples, vessels are very likely to shift position overtime.

2. Related Work

There are 2 main types of satellite imagery: radar and optical. Synthetic aperture radar (SAR) satellites such as Sentinel 1 have the advantage of being unaffected by cloud coverage and light condition. However, they typically come in low resolution (one pixel $\geq 5 \times 5$m). On the other hand, optical satellite imagery typically come in higher resolution(one pixel $\geq 3 \times 3$m), but capture scenes the same way as human eyes do, hence are sensitive to whether conditions and day-night shifts. In this project, we use PlanetScope's 3m resolution optical imagery for detecting small vessels.

A large body of work exists in object detection using very high resolution ($\leq$ 1m) optical satellite imagery. While very high resolution satellite imagery are great for

detecting small objects, they lack global coverage and are prohibitively expensive. Nevertheless, this body of work is very relevant to low resolution vessel detection methodology-wise. Some of the earlier work use traditional image processing pipelines such as using intensity thresholding [12], and shape and texture analysis [13]. More recent works demonstrate the effectiveness of deep CNN models, including Faster R-CNN [11], YOLO [4,9], and U-Net [5]. Many contributions in this realm come from modifications made to model architectures. One such example is to adjust Faster R-CNN to better detect small vessels and vessels in dense clusters by combining feature maps from earlier layers of the backbone model with the one produced by the last layer [11] . The authors also modified the default aspect ratios of anchor boxes in the region proposal network to common vessel aspect ratios. Similarly, the authors of [5] made modifications to the U-net architecture. While these work are promising, they work with single images and are prone to false positives. We are not aware of any prior work on optical vessel detection that exploits performance gain from using multi-temporal data.

More relevant work come from Change Detection (CD) [1–3, 6, 8, 10]. CD systems produce a difference image using binary segmentation based on a sequence of coregistered images, aiming to distinguish real changes from irrelevant changes. Some method applied in change detection include RNN and LSTM [8] and transformers [1], and many existing DL-based CD methods are based on CNNs [2, 3, 6, 10]. While our task is object detection instead of pixel-level classification, the image fusion techniques proven to be effective in CD could be effective for our task as well. [2, 3, 6] show the effectiveness of concatenating image pairs channel-wise as input to CNN models, [2, 3] shows effectiveness of fusing feature maps from encoder stream. [10] is a particularly relevant work that uses Faster R-CNN for change detection. They showed that using pixel by pixel difference between two images worked significantly better than using the channel-wise concatenation. Inspired by their work, we experiment with difference image as input in our project as well.

3. Dataset

We collected 10 scenes from 5 different locations along the Peruvian shoreline from PlanetScope[1]. Each scene pair come from random but different timestamps within 2022 April 20 and 2016 Jan 1. Each scene is a satellite image of size around 10000 x 4000 pixels, and has 3m per pixel resolution. After preprocessing each scene by cropping off land area and chunking it into a set of $299 \times 299 \times 3$ images, we end up with a total of **592 image pairs**, each image pair containing a target image, I_i^{T1}, and a reference image, I_i^{T2}.

[1]Planet.com

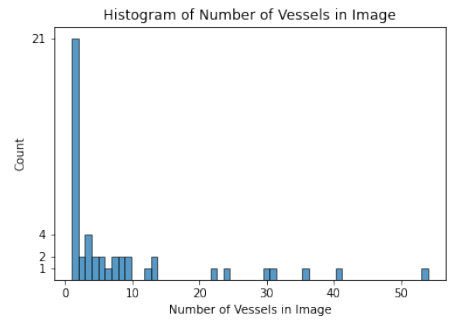

Figure 4. 48 target images in the training set have at least one vessel. Histogram shows the number of vessels per image in these 48 images.

Each pixel in I_i^{T1} and I_i^{T2} share the exact same geographical coordinates. We aim to detect vessels in I_i^{T1} using I_i^{T2} as a reference image. We manually labeled vessel bounding boxes for each I_i^{T1} using VIA[2]. Out of the 592 target images, 48 of them contain at least one vessel. Figure 8 shows the histogram of vessel counts per image.

We chose the shoreline around Peru as our area of interest because of its heterogeneity in vessel clustering (having both dense vessels clusters and sparse single vessels), vessel size (ranging from 15m long to larger than 90m long), and reef distribution (small and large reef/island near shore). Figure 3 shows example images from our dataset.

3.1. Preprocessing

In this section, we describe in detail how we preprocessed our dataset. This process is also shown in Figure 5.

1. Each scene downloaded from PlanetScope has 4 bands (RGB + Near Infrared), of size around 10000 x 4000 pixels. We extract the RBG bands.

2. We create masks for areas of interest (AOI) [3], and use the GDAL library to crop away land, and excessive open water when the scene contains excessive amount of homogeneous ocean surface.

3. We split the AOI into a collection 299 by 299 squares. Since AOI are not exactly aligned with the axis, some squares near the edges of the scene will contain empty pixels. Every empty pixel is set to black. Figure 5 part 3 shows one such example.

4. Method

We compare fusion of T1 and T2 images against baselines that do not use T2 images. Specifically, our methods

[2]https://www.robots.ox.ac.uk/ vgg/software/via/via.html
[3]using http://geojson.io/

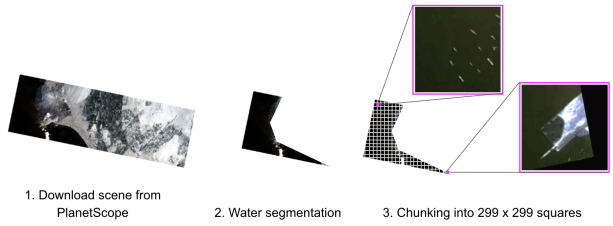

Figure 5. The preprocessing pipeline. We download scene from PlanetScope, mask out land area, and split the resulting scene into squares of size 299 by 299 (while padding empty pixels with 0).

are categorized into Early Fusion and Late Fusion, which we explain below.

All our methodologies are based on Faster R-CNN with ResNet-50 backbone pretrained on COCO. When input is a 6 channel image, we modify the transformation module and the first convolution layer in ResNet-50 to intake 6 channels instead of 3, and copy the weights of the pretrained convolution layer to the 3 additional channels. We train all the parameters Faster R-CNN using SGD with learning rate=0.001, momentum=0.9 and weight decay=0.0005.

We have tried training without using pretrained weights, which led to significant overfitting and validation performance close to random guessing. Therefore, we use pretrained Faster R-CNN for all experimentation.

4.1. Baseline

We compare against three baselines. The first one is vanilla Faster R-CNN (**Faster R-CNN**) which does not make use of reference images at all. We do not make any changes to architecture. The second baseline is random early fusion (**Random Early Fusion**), where we stack the T1 image with a image randomly selected from the set of T2 images into a 6-channel image, as input to Faster R-CNN. The third baseline is identity early fusion (**Identify Early Fusion**), where we stack the T1 image with itself. Random Early Fusion and Identity Early Fusion are non-temporal methods since they do not make of any additional information other than the target image itself.

4.2. Early Fusion

Our first proposed method is Early Fusion, where we concatenate the target image with a reference image, both of size $(299, 299, 3)$, into a 6-channel image of size $(299, 299, 6)$ as input to Faster R-CNN. When stacking the target image I_i^{T1} with its corresponding reference image I_i^{T2}, we name this method **T2 Early Fusion**. This method corresponds to (a) in Figure 6. We also experimented with using the pixel-wise absolute difference between I_i^{T1} and I_i^{T2}, —$I_i^{T1} - I_i^{T2}$— as the reference images, and name this method **Diff Early Fusion** and it corresponds to (b) in Figure 6. The Diff Early Fusion method directly blacks out static objects that appear in both the target and the reference images such as reefs and port. [4]

4.3. Late Fusion

The second method we experiment is Late Fusion. We run the ResNet-50 backbone on the T1 image and the T2 image separately, and use the difference of the two resulting feature maps as input to the region proposal network. We name this method **Diff Late Fusion** and it corresponds to (c) in Figure 6. We experimented with using a convolution layer to aggregate the feature maps, instead of taking the direct difference. However, that method led to significant overfitting and is thus not discussed in this report.

5. Evaluation

5.1. Metrics

Since our dataset is small, we used 5-fold cross validation. For each fold, we used one of the five scenes as test data. Since none of the scenes overlap geographically, the model has not seen the test locations during training.

Following standard practices, the evaluation metrics we use are precision, recall, F-1, and average precision (AP). We specify Intersection over Union (IoU) and confidence score to be 0.5. For example, $Precision = \frac{TP}{TP+FP}$, where TP + FP = # predicted vessels with confidence score ≥ 0.5. To find matches, detection-results with confidence score ≥ 0.5 are sorted by decreasing confidence and are assigned to ground-truth objects. We have "a match" when the two bounding boxes have an IoU ≥ 0.5 (Intersection over Union greater than 50%). This match is considered a true positive if that ground-truth object has not been already used. AP is the area under the precision-recall curve by changing the confidence score from 0 to 1. Note that metrics are not averaged over images, since vessel counts per image is long-tailed. Instead, they are computed by averaging over bounding boxes in all images.

5.2. Hyperparameter tuning

Since our dataset contains only 592 images and we did not use data augmentation techniques, the model could easily overfit to the training dataset. Thus, we hypothesized that freezing some parameters will help the model generalize. However, as shown in Figure 8, empirical results suggest the opposite. Training all parameters of Faster RCNN was more beneficial than just training a subset of them, for both the baselines and our temporal contrast methods, perhaps due to large input domain shift between satellite im-

[4] In Diff Early Fusion, an implementation detail is we used 0 as the mean value in the transformation module for the last 3 channels, instead of copying the mean values from the first three channels.

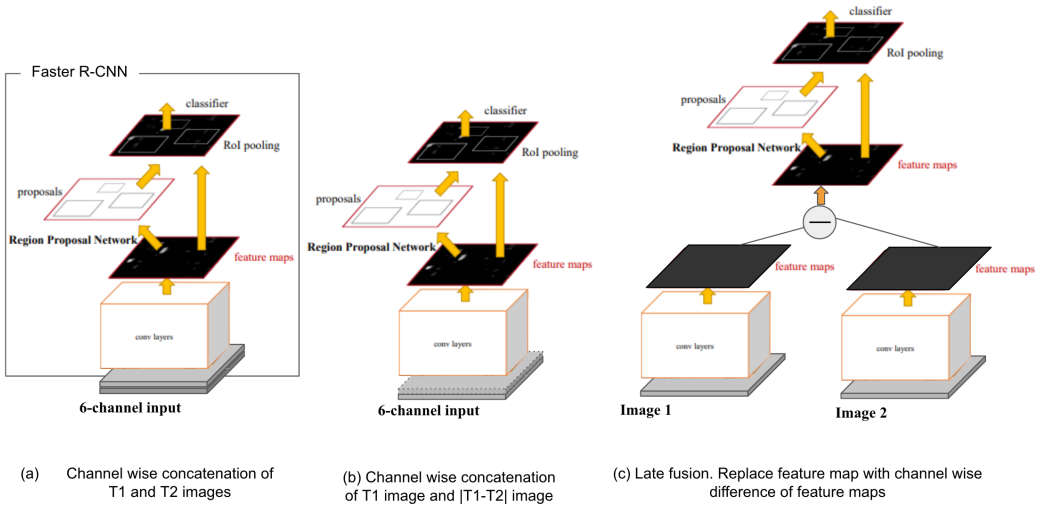

Figure 6. Model architectures. Figure adapted from the original Faster RCNN paper [7]. We use Faster R-CNN with ResNet-50 backbone pretrained on COCO. (a) Early Fusion: we modify the first convolution layer in ResNet-50 to intake 6 channels instead of 3, and copy the weights of the pretrained convolution layer to the 3 additional channels. (b) Difference Image: we input the pixel-wise absolute difference between the two images. (c) Late fusion: We take the absolute difference of the feature maps produced by the backbone model as input feature map to the region proposal network.

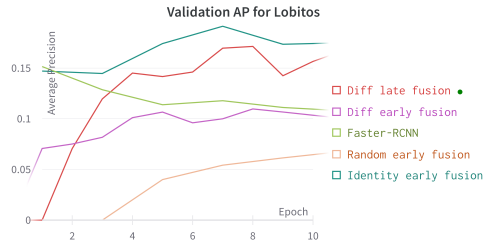

Figure 7. Validation Average Precision for a particular validation split, where the validation scene contained many vessel-like light spots. Identity Early Fusion outperform vanilla Faster RCNN.

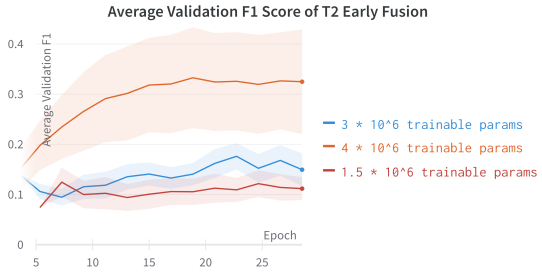

Figure 8. Freezing parameters in pretrained Faster RCNN degrades validation performance. Figure shows average validation F1 score of T2 Early Fusion by freezing the ResNet backbone, freezing Layer 4 in the ResNet backbone, and not freezing the backbone. The trend suggests that not freezing any weights performs best overall. The same experiment was run on Identity Early Fusion, Diff Late Fusion, and vanilla Faster RCNN, and the same trend was observed throughout.

agery and pretraining images. Thus, we have trained all models from end to end.

5.3. Results

As shown in Table 1, we observe strong performance from vanilla Faster RCNN, which achieves the best overall AP, F1, Recall and Small Vessel Recall. Temporal Fusion methods - T2 Early Fusion, Diff Early Fusion, and Diff Late Fusion - achieve better precision at the expanse of recall, when fixing confidence score to 0.5. T2 Early Fusion achieves similar precision but significantly better Recall compared to Diff Early Fusion and Diff Late Fusion, suggesting the benefit of having the raw reference image as input. This is not surprising because difference images are further outside the pretraining input domain compared to the the reference images, so pretrained weights may not help to extract features from different images. Late Fusion underperforms Early Fusion, which may be due to information loss in feature maps. Figure 1 compares vanilla Faster R-CNN against T2 Early Fusion, showing through examples the improvement in Precision T2 Early Fusion brings.

Despite the potential Fusion methods display, vanilla Faster RCNN achieves the best Average Precision and F1 score. Vanilla Faster R-CNN being the overall best performing model is somewhat unintuitive since having additional information should certainly not hurt performance. Our hy-

Method	AP	F1	Precision	Recall	Small Vessel Recall
Faster R-CNN	**0.35**	**0.50**	0.49	**0.65**	**0.53**
Random Early Fusion	0.20	0.31	0.29	0.48	0.41
Identity Early Fusion	0.33	0.47	0.49	0.63	0.52
T2 Early Fusion	0.27	0.41	**0.53**	0.60	0.44
Diff Early Fusion	0.15	0.26	**0.55**	0.30	0.17
Diff Late Fusion	0.23	0.37	**0.55**	0.41	0.22

Table 1. Model performance evaluated at IOU=0.5 and confidence=0.5. All models were trained for 30 epochs except for Diff Late Fusion which was trained for 60 epochs due to smaller convergence. All models were trained under the same hyperparameter configuration, except for batch size due to memory limit. Faster RCNN has trained using batch size = 16 and all other models were trained using batch size = 2. Small Vessel refers to vessels with ground truth bounding box of length between 15 and 30m. Faster R-CNN performs overall the best, but we observe improvement in precision using the early fusion method at the expense of recall. We report the validation result obtained by independently maximizing for each metric.

Method	AP	F1	Precision	Recall	Small Vessel Recall
Faster R-CNN	0.12	0.26	0.17	0.56	0.8
Identity Early Fusion	**0.19**	**0.33**	**0.22**	**0.67**	**0.8**

Table 2. Validation result for a particular validation split. The validation scene contains many vessel-like light spots. Identity Early Fusion outperforms vanilla Faster RCNN in all metrics. All other methods performed worse than Faster RCNN and results are not included for brevity.

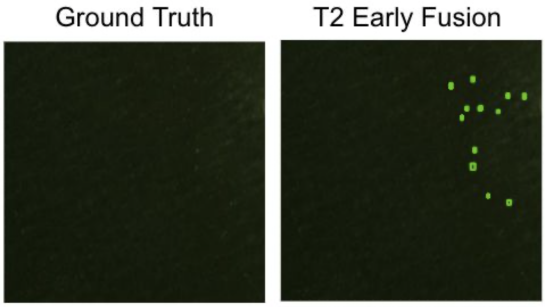

Figure 9. A common pattern exhibited by T2 Early fusion is misclassification of wave noises as vessels. Temporal contrast methods could backfire if model overfits to positional shifting instead of vessel features.

pothesis is that doubling the size of the input domain leads to models overfitting on noisy signals, which counteracted the benefits of having additional training signals, especially when our dataset is small to learn from.

Upon analysing the T2 Early Fusion model, we found a common mistake pattern where the model confuses wave highlights with vessels. Figure 9 shows one such example. One explanation is wave highlights shift overtime, similar to how vessels would shift overtime. Temporal contrast method could backfire if it overfits to noisy temporal features due to doubling input size, instead of vessel features.

Upon closer inspection, we observe some unexpected and interesting performance from the Identity Early Fusion method. As shown in Table 2 and Figure 7, for a particular validation split, where the validation scene contained many light spots that could be easily confused with vessels, Identity Early Fusion outperforms or matches all methods, including vanilla Faster R-CNN, in all metrics. It is interesting since Identity Early Fusion is simply stacking the T1 images with itself, without introducing any additional learning signals. Figure 10 compares vanilla Faster RCNN and Identity Early Fusion through some examples. Even in cases where Identity Early Fusion mistakenly predicts a light spot as a vessel, its confidence score is lower than Faster RCNN's misprediction. We present this result as an interesting observation in addition to the results presented above.

6. Conclusion

We applied Early Fusion and Late Fusion methods on Faster R-CNN to improve small vessel detection in low resolution satellite imagery. We used reference images to aid the detection of vessels in target images. Using transfer learning, Faster R-CNN is a strong baseline, and fusion methods show promising improvement in vessel detection precision. One caveat of Early Fusion methods is that they are prone to overfit to noisy temporal features such as positional shifting due to the doubling input domain size. Finally, we find that Late Fusion underperforms Early Fusion, which may be due to information loss in feature maps.

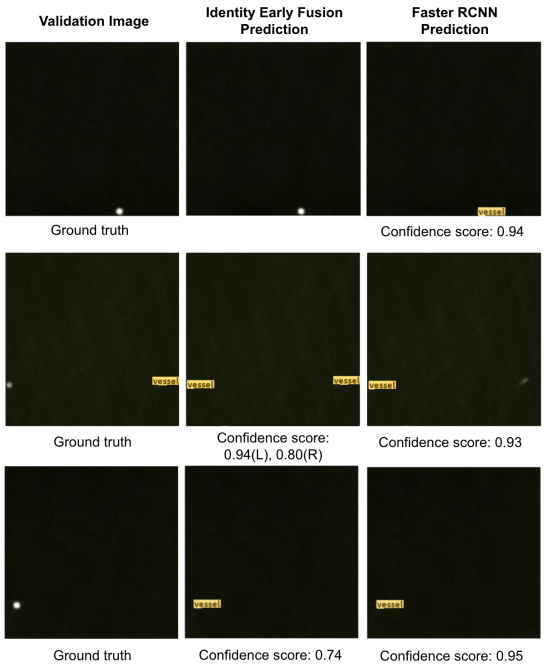

Figure 10. Examples of validation prediction from Faster RCNN and Identity Early Fusion for a particular split, where the validation scene contained many vessel-like light spots. Identity Early Fusion achieved better AP, recall, precision, and F1 score compared to Faster RCNN. Identity Early Fusion demonstrate lower misclassification rate, and lower confidence for wrong predictions.

7. Future Work

Our dataset only contained 592 images and only 48 of those contained vessels. Training a 4 million parameter model on 592 images lead to overfitting. While using pretrained weights helped to mitigate this problem, having a large dataset will certainly help to further mitigate this problem.

We would also like to benchmark our methods against attention-based CNNs, in order to understand whether having local reference is sufficient for vessel detection, or if having global attention is important.

We have experimented with non-linear Late Fusion, where we used a convolutional layer to aggregate the two feature maps, but observed significant overfitting. However, it is possible that non-linear aggregation works better when using a large dataset. We would like to test out this hypothesis once we collect a larger dataset.

8. Contributions and Acknowledgement

I wish to acknowledge Stanford Center of Ocean Solutions for providing me access to satellite imagery on Planet.com, and the course staff for Stanford's course CS 231N: Convolutional Neural Networks for Visual Recognition, for providing me Google Cloud credits.

References

[1] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 3

[2] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067. IEEE, 2018. 3

[3] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks, 2018. 3

[4] Adam Van Etten. You only look twice: Rapid multi-scale object detection in satellite imagery. *CoRR*, abs/1805.09512, 2018. 3

[5] Sagar Karki and Siddhivinayak Kulkarni. Ship detection and segmentation using unet. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–7, 2021. 3

[6] Daifeng Peng, Yongjun Zhang, and Haiyan Guan. End-to-end change detection for high resolution satellite images using improved unet++. *Remote Sensing*, 11(11), 2019. 3

[7] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 1, 5

[8] Marc Rußwurm and Marco Körner. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1496–1504, 2017. 3

[9] Gang Tang, Shibo Liu, Iwao Fujino, Christophe Claramunt, Yanping Wang, and Shaoyang Men. H-YOLO: A Single-Shot Ship Detection Approach Based on Region of Interest Preselected Network. *Remote Sensing*, 12(24):4192, 2020. 3

[10] Qing Wang, Xiaodong Zhang, Guanzhou Chen, Fan Dai, Yuanfu Gong, and Kun Zhu. Change detection based on faster r-cnn for high-resolution remote sensing images. *Remote Sensing Letters*, 9(10):923–932, 2018. 3

[11] Shaoming Zhang, Ruize Wu, Kunyuan Xu, Jianmei Wang, and Weiwei Sun. R-cnn-based ship detection from high resolution remote sensing imagery. *Remote Sensing*, 11(6), 2019. 3

[12] Ying-hai Zhao, Xiu-qing Wu, Ling-yun Wen, and Shou-shi Xu. Ship target detection scheme for optical remote sensing images. *Opto-Electronic Engineering*, 35(8):102–106, 2008. 3

[13] Changren Zhu, Hui Zhou, Runsheng Wang, and Jun Guo. A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features. *IEEE Transactions on geoscience and remote sensing*, 48(9):3446–3456, 2010. 3