Visible and Infrared Image Fusion Framework based on RetinaNet for Marine Environment

Fahimeh Farahnakian[1], Jussi Poikonen[2], Markus Laurinen[2], Dimitrios Makris[3] and Jukka Heikkonen[1]

***Abstract*—Safety and security are critical issues in maritime environment. Automatic and reliable object detection based on multi-sensor data fusion is one of the efficient way for improving these issues in intelligent systems. In this paper, we propose an early fusion framework to achieve a robust object detection. The framework firstly utilizes a fusion strategy to combine both visible and infrared images and generates fused images. The resulting fused images are then processed by a simple dense convolutional neural network based detector, RetinaNet, to predict multiple 2D box hypotheses and the infrared confidences. To evaluate the proposed framework, we collected a real marine dataset using a sensor system onboard a vessel in the Finnish archipelago. This system is used for developing autonomous vessels, and records data in a range of operation and climatic and light conditions. The experimental results show that the proposed fusion framework able to identify the interest of objects surrounding the vessel substantially better compared with the baseline approaches.**

***Index Terms*—Autonomous vehicles, object detection, sensor fusion, convolutional neural networks.**

I. Introduction

Sea transportation is carrying about 90% of the world trade according to the International Maritime Organization (IMO) [1]. With the current growth of maritime traffic, security and safety are vital issues. For this reason, lots of efforts have been deployed to improve the security and safety in the maritime environment over the past few years. To develop a reliable autonomous ships, designing efficient object detection is a critical task [2]. However, object detection in maritime environments is still a challenging and complex task due to varying light, view distances, weather conditions and dynamic nature of the sea caused by waves. In addition, light reflection, camera motion and illumination changes may cause to false detections [3].

One of the main technologies that improve the understanding of the surrounded environment and therefore the robustness of object detection is sensor fusion. As each individual sensor has some weakness, combining the data from different sensors can optimise the situational awareness under all conditions. For example, visible cameras provide high resolution images for the object classification task. Although, infrared cameras can increase nigh-time navigation safety and detect warm objects at night time with high accuracy. Therefore, we believe that multi-sensor data fusion can develop a reliable perception capability for object detection in autonomous vehicles

Various object detection approaches have been proposed for autonomous vehicles in recent years. Most of these approaches utilized a Convolutional Neural Network (CNN) based network. This network is able to learn rich features outperforming hand-crafted features. Generally, CNN-based object detector can be divided into two main groups: two-stage [4] and one-stage detectors [5]. Two-stage object detectors utilize a classifier on a sparse set of candidate object locations [4]. In contrast, one-stage object detectors generate dense sampling of possible object proposal in a faster and simpler fashion. However, two-stage detectors have usually higher accuracy than one-stage ones because they maintain a manageable balance between the foreground and the background. RetinaNet [5] can match the speed of one-stage detectors while achieving similar accuracy comparing with all existing state-of-the-art two-stage detectors. In addition, it has proposed a loss function that acts as a more effective alternative to previous approaches for dealing with class imbalance.

In this paper, we present an early fusion or feature-level fusion framework based on RetinaNet for marine environment. Our framework performs object detection in two phases, In the first phase, the fusion framework combines the information of two source images from the visible and infrared cameras using visual saliency map based on weighted least square [6]. In the second phase, it employs RetinaNet on the fused image for detection of the interest objects around the vessel. The detected object is classified into one of five types of vessel or navigation buoys. The proposed framework is evaluated on a real dataset we collected in the Finnish archipelago in different environmental conditions (i.e. weather conditions, day/night). In addition, different backbone networks are proposed for RetinaNet to find the best model. We also study the impact of fusion on object detection performance. Experimental results show that our proposed framework outperforms the existing state-of-the-art approaches. To the best of our knowledge, currently there are no existing works on using real sensor fusion data for object detection in maritime environment.

The remainder of the paper is organized as follows. Section II discusses some of the most important related works. The proposed fusion framework is introduced in Section III.

*This work is part of the New 3D Analytics Methods for autonomous Ships and Machines projects funded by the Business Finland.

[1] Fahimeh Farahnakian, and Jukka Heikkonen are with Department of Future Technologies, University of Turku, Turku, Finland {fahfar, jukhei}@utu.fi

[2] Jussi Poikonen, and Markus Laurinen are with Rolls-Royce Oy Ab,Turku, Finland {jussi.poikonen, markus.laurinen}@rolls-royce.com

[3] Dimitrios Makris is with Faculty of Science, Engineering annd Computing, Kingston University, London {d.makris}@kingston.ac.uk

Experimental results are presented in Section IV. Finally, the conclusions is presented in Section V.

II. Related Work

Image Fusion methods: the main goal of image fusion is to generate a fused image with complementary information from the same sensor with several imaging parameters or from the multiple sensors. Generally, the multi-sensor fusion methods can be divided into three main groups based on the level of data abstraction used for fusion. (1) Measurement fusion methods first convert the data from each sensor to a common form and then the actual fusion of data is performed in the common representation. (2) Feature level fusion methods extract the relevant feature of each sensor individually and then the obtained features are combined into a single vector as an input of a fusion module. Therefore, the measurement and feature level fusion methods fuse raw sensor data or concatenate feature descriptors. (3) Decision level fusion methods independently perform object detection from each sensor and the outputs of each sensor are fused at the decision level for final classification. The traditional image fusion algorithms can be divided into three main groups depending on the fusion strategies : Multi-Scale Decomposition (MSD)-based methods [6], spatial domain-based methods [7], and Sparse Representation (SR)-based methods [8]. The MSD-based methods usually employ pyramidal transforms, discrete wavelet transform, and discrete wavelet frames. The spatial domain methods usually address the fusion issue via local spatial features such as gradient, spatial frequency and local standard deviation. The SR-based methods measure the activity level of source images in a sparse domain. Recently, Deep Learning (DL) has shown significant success for challenging tasks in sensor fusion such as multi-focus image fusion [9] [10], multi-exposure image fusion [11] and multi-modality imaging [12] [13].

CNN-based Object Detector: CNN [14] is the most popular type of neural network for object detection. Inspired by the success of applying CNN in many challenging object detection problems [14]–[16], our framework employed a CNN-based network for this purpose. SqueezeDet [17] employs a CNN for real-time object detection for autonomous driving. It is based on a small backend network of SqueezeNet [18] in order to detect small-size objects. Region-based Convolutional Neural Networks (R-CNN) [19] led to substantial gains in object detection accuracy. It first identifies region proposals (i.e. regions of interest that are likely to contain objects) and then classifies these regions into object categories or background using a CNN. One disadvantage of R-CNN is that it computes the CNN independently on each region proposal, leading to time-consuming and energy-inefficient computation. In order to improve computational efficiency, Fast R-CNN [4] omits the selective search method for generating object region proposals. In [20], a system based on Fast R-CNN is proposed for detection and classification of on-road objects. The outputs of the system are the rectangular bounding boxes and class information of objects which are useful parameters for motion planning of the self-driving vehicle. The infrared deep learning network is found to be robust to variation in object's view, lighting and climatic conditions. AlexNet [21], ZFNet [22], VGGNet [23], ResNet [24] and GoogLeNet [25] are other popular deep CNNs for object classification and detection.

III. The Proposed Fusion Framework

In this section, we describe the proposed fusion framework for object detection. Fig.1 shows an illustration of the RetinaNet based framework used for object detection. The framework continuously observes the environment through two sensors (visible and infrared cameras). Therefore, the framework considers both sources, visible and infrared images, as inputs. Object detection is performed in two phases. In the first phase, the two input images are fused according to the strategy explained in subsections A. The main goal of this phase is generating the fused images which are more robust to imperfect conditions such as mis-registration. In the second phase, the obtained fused image is processed by an one-stage CNN-based detector to provide a set of detected objects, as described in subsection B. The detector employs the deep RetinaNet architecture is used, considering its efficiency and accuracy. Each detected object is represented by its position and class label that indicates whether it is a vessel (passenger vessel, motorboat, sailboat, docked vessel) or navigation buoy.

A. *Visual Saliency Map and Weighted Least Square*

Visual Saliency Map and Weighted Least Square (VSM-WLS) [6] is a multi-scale fusion method based on WLS optimization and VSM. The method proposes an Multi-Scale Decomposition (MSD) step using the rolling guidance filter [26] and Gaussian filter to decompose the infrared and visible images into base and detail layers. MSD tries to obtain an effective scale awareness and edge preservation when decomposing images. Decomposed base layers are fused using a weighted average technique to enhance the contrast of the fused image. In addition, a weighted least square optimization is used to fuse the detail layers to enhance the max-absolute fusion rule by considering different characteristics of visible and infrared images. Finally, inverse MSD is applied to the output of both previous steps to construct the final fused image.

B. *RetinaNet*

RetinaNet [5] is a simple dense detector which contains a backbone network and two sub-networks. First, the backbone network computes a convolutional feature map over an entire input image. Then, the first sub-network performs convolutional object classification on the backbone's output and the second sub-network applies convolutional bounding box regression. The backbone network of RetinaNet uses the Feature Pyramid Network (FPN) [27] in order to efficiently constructs a rich, multi-scale feature pyramid from a single resolution input image with a top-down pathway and lateral connections. In our framework, the fused images can be applied as input to a Residual Network (ResNet50 and ResNet101) [24] or Visual Geometry Group (VGG) net [28] encoder, which processes

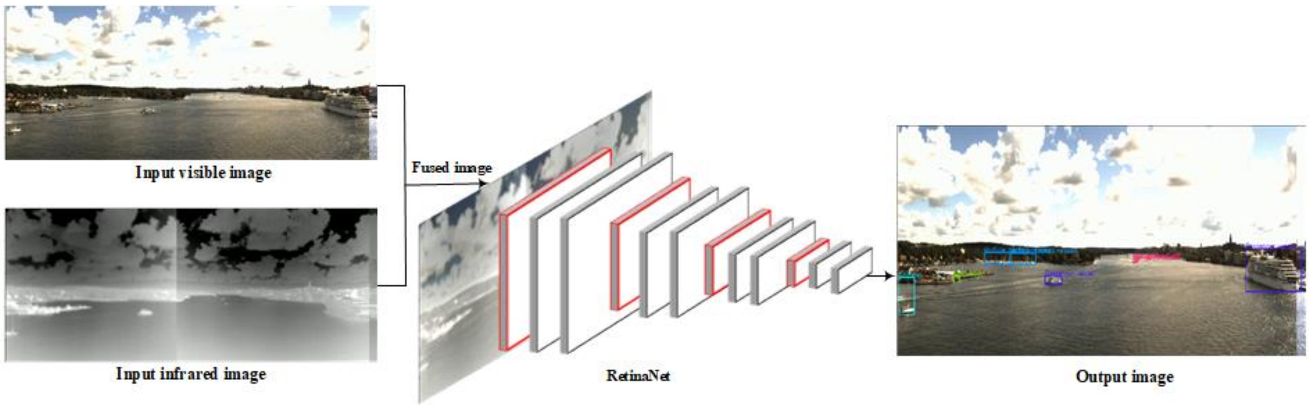


Fig. 1. Proposed RetinaNet based fusion framework. The original input images are of size 3240×944 pixels. They are fused using VSM-WLS in order to provide complementary information for object detection.Then, the fused image is processed by RetinaNet in order to detect and localize objects around the vessel.

the image through convolution kernels and generates deep features. In addition, RetinaNet proposed the focal loss to address the problem of class imbalance and unequal contribution of positive and negative samples as follows:

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma log(p_t) \tag{1}$$

where the parameters α and γ control the balance between negative and positive samples.

IV. Experimental Setup and Results

A. Data Description

Data was collected using a sensor system onboard a vessel in the Finnish archipelago. This sensor system includes visible (visible spectrum) and infrared (thermal) camera arrays, providing output which can be synchronized and stitched to form panoramic images. The individual visible cameras have full HD resolution, while the thermal cameras have VGA resolution. Both camera types have horizontal field of view approximately 35 degrees. Data is collected from this sensor set continuously, providing data from various environmental and geographical scenarios. Registration parameters for image alignment in this sensor set have been determined by manually finding corresponding features in calibration images and by minimizing alignment mismatch. The data shows maritime scenarios with various objects such as ships and other vessels. For the experiments, we selected 5000 and 1750 images for training and testing the network, respectively. Size of each image is 3240×944 pixels. The original training images are augmented via a number of random transformations for training RetinaNet. Random transformations were applied on the images including rotation, cropping, swirl, vertical flip and horizontal flip. This kind of data augmentation has been widely used in previous research. We manually annotated the boundind boxes and the labels for the six object classes (five types of vessels, navigation buoy) on our dataset. Table I shows the distribution of the classes in the training and test datasets for each camera. Note that any far away vessels that could not be visually recognized as ”passenger vessel”, ”motorboat”, ”sailboat” or ”docked vessel”, were placed under the general label ”vessel”.

B. CNN network Hyperparameter

The number of layers depends on the type of backbone in RetinaNet. If the backbone is ResNet50 or ResNet101, the number of layers is 50 and 110, respectively. VGG19 has 19 layers. As there is no guarantee that the deepest network has better performance, various network depths and backbones are tested.

Adam optimizer is used for the stochastic optimization in deep RetinaNet. The other hyperparameters of the network are as follows: anchors are assigned to ground-truth object boxes using an Intersection-over-Union (IoU) threshold of 0.5. We sweep over the number of scale and aspect ratio anchors used at each pyramid level in FPN. We consider cases from a single square anchor at each location to 12 anchors per location spanning 4 sub-octave scales (1,1.2, 1.6) and 3 aspect ratios [0.5, 1, 2, 3]. The learning rate was initialized at 0.00001 with reduction factor of 0.1. The epoch number was 50 and the number of iteration in each epoch was 100. Two losses were computed: the classification loss with focal loss and the regression loss with smooth L1. For the focal loss, the parameters α and γ in Equation (1) are 0.25 and 2 respectively.

C. Evaluation

Effect of fusion: we study how the fusion can effect on the object detection performance. For this purpose, we compare two uni-modal frameworks with our multi-modal framework. The uni-modal framework utilizes only the visible or infrared images to detect the interest objects around of vessel. However, our proposed fusion framework combines the information from two input infrared and visible images using the proposed image fusion methods. In addition, RetinaNet is trained based on three different backbone networks in our experiments: ResNet50, ResNet101 and VGG19 for our experiments. It firstly is pre-trained on ImageNet dataset in order to learn good feature representation. Then, they are fine-tuned on our data.

TABLE I
NUMBER OF EACH OBJECT IN TRAINING AND TESTING MARINE DATASETS FOR EACH CAMERA

	Input images	Passenger vessel	Motorboat	Sailboat	Docked Vessel	Vessel	Navigation buoy	Total
Training dataset	Visible	8481	10849	4753	10500	10250	3500	48333
	Infrared	8481	11349	5006	10750	10500	3500	49586
Test dataset	Visible	1000	3750	3250	214	4750	500	10464
	Infrared	574	3750	3250	214	4750	500	10464

To evaluate the proposed fusion framework, we used the test dataset and measured the Average Precision (%AP) for each class. Table II shows the results of uni-modal and multi-modal frameworks using various backbone networks. The IoU threshold is 0.5 for all classes in all frameworks. The uni-modal framework based on visible images can get the highest accuracy 67.3 and 63.9 AP for two classes "Docked vessel" and "Passenger vessel", respectively. The highest accuracy (64.8) is achieved by the ResNet101 backbone for "Docked vessel" class when the framework performs object detection based on only infrared images. Our fusion framework has higher accuracy for all objects compared with other frameworks. Therefore, the results show that our framework can get the largest gains for "passenger vessels"(68.4% AP) and "Docked vessel" (58.7% AP). Objects in these categories appear larger, so their detection is benefited the most from high-resolution camera data. Based on our results, the proposed RetinaNet-based framework cannot get properly detect small objects which are represented by a low number of pixels (less than 16×16 pixels) in the image. These small objects mostly belong to the vessel and navigation buoy classes.

Qualitative Results: Fig 2 and Fig 3 demonstrate two examples of the detection results from the visible-only framework, infrared-only framework and multi-modal framework. We observe that the proposed fusion framework is better at detection of objects than the uni-modal framework. Detecting very small objects with a few pixels is still challenging as shown in Fig 3 and many of them are not detected by our framework. However, our framework still outperforms others as shown in Table II.

V. CONCLUSION

In this paper, an early fusion framework is proposed in order to detect the interest objects in marine environment. In the proposed framework, the images from both visible and infrared cameras are fused right at the beginning and then an one-stage fast detector, RetinaNet, recognizes and localizes the objects in the fused images. To demonstrate the effectiveness of the proposed framework, we compared it with two uni-modal frameworks applied on only visible or infrared images. We also evaluate the effects of more powerful backbone networks on the performance of RetinaNet in our framework. The experimental results on real marine data show that our multi-modal framework can achieve higher detection accuracy comparison with two another uni-modal frameworks. Our framework is effectively able to detect and classify objects into one of vessel type or navigation buoy in the real marine dataset, as long as their apparent image size is more than 16×16 pixels.

For future work, the effects of object size and distance on the performance of our framework will be studied. As it is very challenging to accurately detect small objects, an improved network structure of RetinaNet will be investigated in the future for this purpose. Further, we will extend our fusion framework by using data from lidar and radar besides RGB and IR cameras to improve the detection results. In addition, more effective fusion schemes based on DL could be further developed to pursue better fusion performance.

ACKNOWLEDGMENT

The implementation is ran on csc taito cluster server.

REFERENCES

[1] Duong Nguyen, Rodolphe Vadaine, Guillaume Hajduch, René Garello, and Ronan Fablet. Multi-task learning for maritime traffic surveillance from AIS data streams. *CoRR*, abs/1806.03972, 2018.

[2] Daniel Socek, Dubravko Culibrk, Oge Marques, Hari Kalva, and Borko Furht. A hybrid color-based foreground object detection method for automated marine surveillance. In Jacques Blanc-Talon, Wilfried Philips, Dan Popescu, and Paul Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems*, pages 340–347, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[3] Roeland T'Jampens, Francisco Hernandez, Florian Vandecasteele, and Steven Verstockt. Automatic detection, tracking and counting of birds in marine video content. *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2016.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[5] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[6] Jinlei Ma, Zhiqiang Zhou, Bo Wang, and Hua Zong. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Physics Technology*, 82, 02 2017.

[7] VPS Naidu and Jitendra R Raol. Pixel-level image fusion using wavelets and principal component analysis. *Defence Science Journal*, 58(3):338, 2008.

[8] Haifeng Li Jian Zou Qiheng Zhang, Yuli Fu. Dictionary learning method for joint sparse representation-based image fusion. *Optical Engineering*, 52:52 – 52 – 12, 2013.

[9] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion*, 36(C):191–207, July 2017.

[10] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144:1–144:12, July 2017.

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):295–307, February 2016.

[12] Hui Li, Xiao-Jun Wu, and Josef Kittler. Infrared and visible image fusion using a deep learning framework. *CoRR*, abs/1804.06992, 2018.

[13] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *CoRR*, abs/1804.08361, 2018.

TABLE II

OBJECT DETECTION RESULTS (AP) OF OUR MULTI-MODAL FUSION FRAMEWORK AND TWO UNI-MODAL FRAMEWORKS IN OUR MARINE TEST DATA.

Farmework	Input image	Backbone	Passenger vessel	Motorboat	Sailboat	Docked Vessel	Vessel	Navigation buoy
Uni-modal	Visible	Resnet50	54.3	41.1	38.9	56.3	21.2	11.2
		Resnet101	65.2	56.1	45.8	67.3	27.9	17.9
		VGG19	63.9	53.5	42.6	64.9	22.1	12.1
Uni-modal	Infraded	Resnet50	48.7	40.9	32.9	58.1	22.1	12.1
		Resnet101	63.7	59.2	43.9	64.8	24.6	14.6
		VGG19	61.8	53.7	40.6	63.7	24.8	14.8
Multi-modal	Visible+ Infraded	Resnet50	56.9	54.9	44.9	67.5	19.6	10.1
		Resnet101	**68.4**	**58.7**	**52.2**	**67.8**	**27.3**	**16.6**
		VGG19	66.5	56.1	51.3	69.6	26.5	16.1

[14] Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009.

[15] 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.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[17] B. Wu, F. Iandola, P. H. Jin, and K. Keutzer. Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 446–454, July 2017.

[18] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and $<$1mb model size. *CoRR*, abs/1602.07360, 2016.

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):142–158, January 2016.

[20] G. Prabhakar, B. Kailath, S. Natarajan, and R. Kumar. Obstacle detection and classification using deep learning for tracking in high-speed autonomous driving. In *2017 IEEE Region 10 Symposium (TENSYMP)*, pages 1–6, July 2017.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[22] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

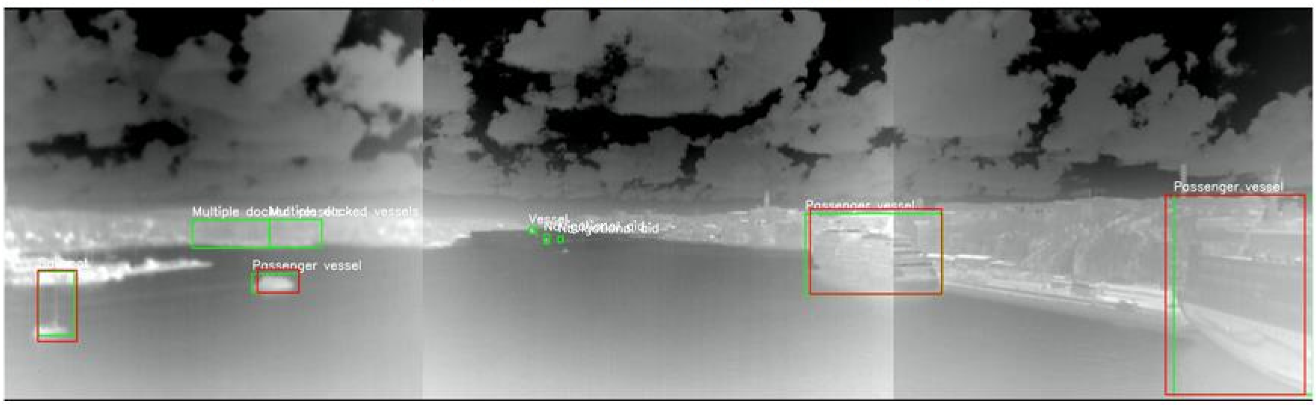
[25] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[26] Qi Zhang, Xiaoyong Shen, Li Xu, and Jiaya Jia. Rolling guidance filter. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 815–830, Cham, 2014. Springer International Publishing.

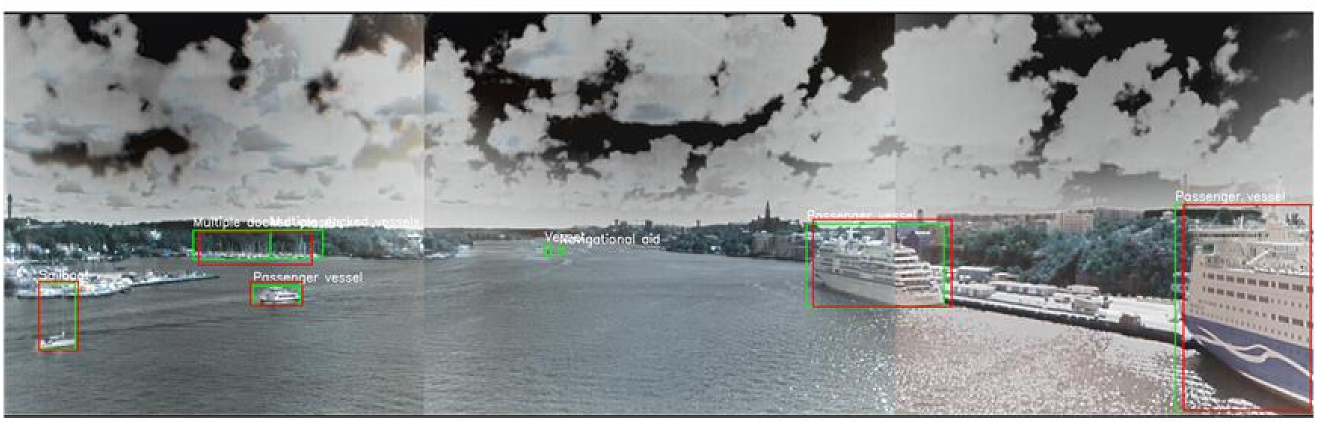
[27] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944, 2017.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

(a) Uni-modal framework based on visible images



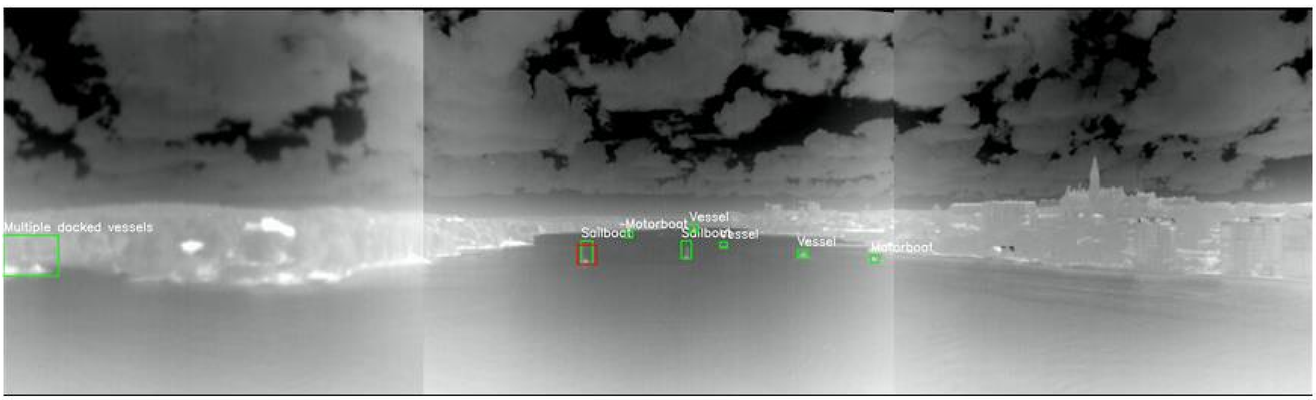
(b) Uni-modal framework based on infrared images



(c) Multi-modal framework based on visible and infrared images

Fig. 2. Qualitative results of the proposed framework on an example image from our dataset. The ground truth bounding boxes are shown in green. Red bounding boxes are the predicted bounding box for each ground truth bounding box.

(a) Uni-modal framework based on visible images



(b) Uni-modal framework based on infrared images

(c) Multi-modal framework based on visible and infrared images

Fig. 3. Qualitative results on example image from our dataset with more small objects. The ground truth bounding boxes are shown in green. Red bounding boxes are the predicted bounding box to each ground truth bounding box.