Appendix

1. Experiments on A Subset of COCO

In order to verify the performance of the proposed method on datasets other than the proposed SIRST dataset, we also conduct a semantic segmentation experiment on StopSign, a subset of the well-known COCO dataset [10] as illustrated in Fig. 1. We choose the mean intersection over union (mIoU) as the evaluation metric and the cross entropy as the loss function. The rest hyper-parameters are the same as the settings for the experiments on the SIRST dataset.

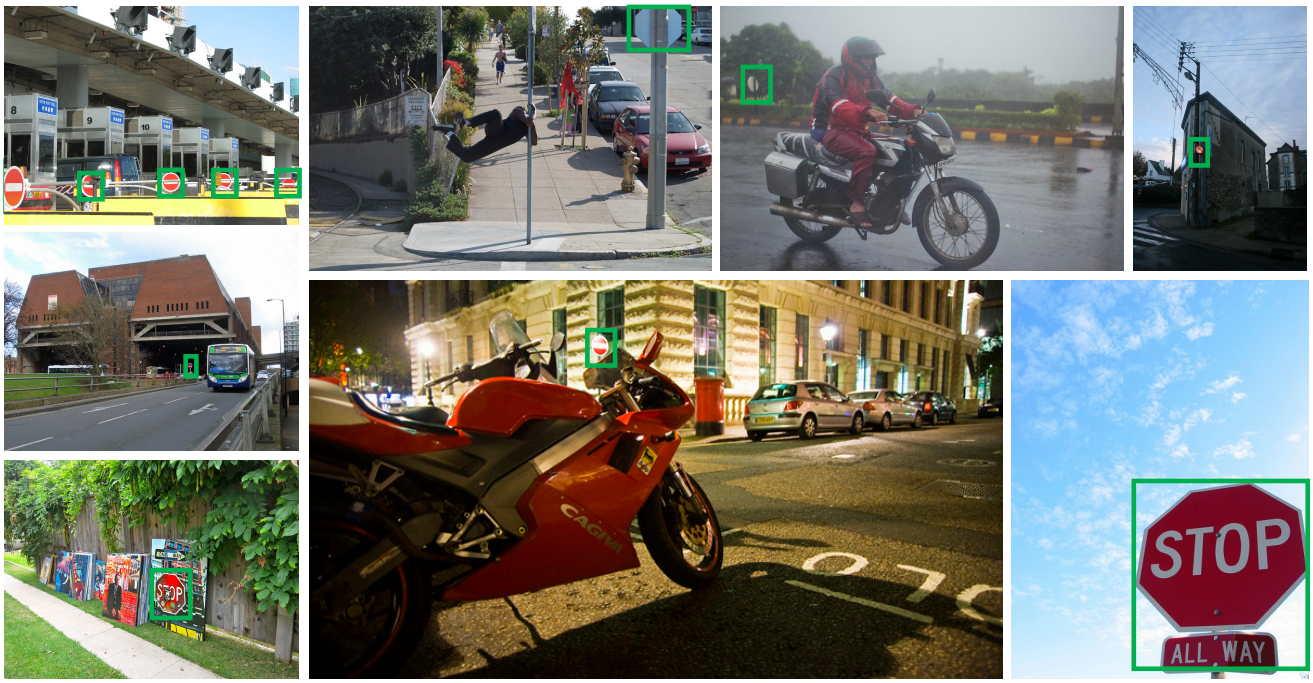

Figure 1: The representative images from the StopSign dataset.

The results are provided in Table 1, from which it can be seen that the proposed network performs best given the same network depth. Considering that the difference between GAU-FPN and ACM-FPN is that the proposed ACM-FPN has an additional bottom-up pathway based on the proposed point-wise channel attentional modulation, we believe that this performance boost stems from the proposed bottom-up modulation. In addition, it can be seen that a deeper network does not necessarily lead to better performance. For example, in both SK-FPN and ACM-FPN, when the block number b in each stage increases, the mIoU decreases a bit. Therefore, instead of blindly increasing the network depth, designing sophisticated attention modules for the cross-layer feature fusion holds great potential for better performance.

2. Accelerated Implementations

Besides faithfully re-producing these state-of-the-arts models in the toolkit of SIRST, for many non-learning based models, we also implement them with some accelerating schemes without harming the final performance. To elucidate how these schemes help, here are some examples:

Table 1: The mIoU comparison of four networks in various network depths

Network	$b = 1$	$b = 2$	$b = 3$	$b = 4$
FPN [9]	0.894	0.920	0.925	0.928
SK-FPN [8]	0.901	0.932	0.931	/
GAU-FPN [7]	0.918	0.933	0.940	0.944
ACM-FPN (ours)	**0.947**	**0.959**	**0.957**	**0.954**

1. For the local contrast-based methods, given central and neighborhood feature maps, the local contrast map is generally calculated pixel-wisely [2, 14]. However, it can be replaced with a cyclic shift on the whole feature maps to save time. For instance, with this exchanging trick, MPCM can be 15% faster, increasing from 2.67 FPS to 3.07 FPS.

2. For many low-rank based methods [6], the target-background separation is achieved via accelerated proximal gradient (APG) method [1], which is slow. To speed them up, for all low-rank based methods, we implement them with the Inexact Alternating Direction Method (IALM).

3. Again, for low-rank based methods, we add the stopping criteria proposed in [4] as a choice, which can save up to 50 times of the computational time.

3. Implementation details

We implemented all the learning-based methods in MXNet [3] and non-learning based methods in MATLAB. For all learning-based methods, we choose to minimize the Soft-IoU loss function [12] over the training set. To stack images of different sizes into a batch, each image is resized to 512×512 and randomly cropped to 480×480 during training. The detailed hyper-parameter settings of the non-learning methods are listed in Table 2.

Table 2: Detailed hyper-parameter settings of non-learning methods for comparison.

Methods	Hyper-parameter settings
MPCM [14]	$N = 1, 3, ..., 9$
FKRW [11]	$K = 4$, $p = 6$, $\beta = 200$, window size:11×11
SMSL [13]	Patch size: 50×50, $\lambda = \frac{2\times L}{\sqrt{\min(m,n)}}$, $L = 2.0$, threshold factor: $k = 1$
IPI [6]	Patch size: 50×50, stride: 10, λ=$L/\min(\mathrm{m}, \mathrm{n})^{1/2}$,$L = 4.5$, threshold factor: $k = 10$, ε= 10^{-7}
NIPPS [5]	Patch size: 50×50, stride: 10, $\lambda = \frac{L}{\sqrt{\min(m,n)}}$, $L = 2.0$, energy constraint ratio: $r = 0.11$, threshold factor: $k = 10$
RIPT [4]	Patch size: 50×50, stride: 10, $\lambda = \frac{L}{\sqrt{\min(I,J,P)}}$, $L = 0.001$, $h = 0.1$, ϵ=0.01, $\varepsilon = 10^{-7}$, threshold factor:$k = 10$

References

[1] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[2] C. L. Philip Chen, Hong Li, Yantao Wei, Tian Xia, and Yuan Yan Tang. A local contrast method for small infrared target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):574–581, 2014.

[3] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems. In *In Neural Information Processing Systems, Workshop on Machine Learning Systems*, volume abs/1512.01274, 2015.

[4] Yimian Dai and Yiquan Wu. Reweighted infrared patch-tensor model with both nonlocal and local priors for single-frame small target detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(8):3752–3767, 2017.

[5] Yimian Dai, Yiquan Wu, Yu Song, and Jun Guo. Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values. *Infrared Physics & Technology*, 81:182–194, 2017.

[6] Chenqiang Gao, Deyu Meng, Yi Yang, Yongtao Wang, Xiaofang Zhou, and Alexander G. Hauptmann. Infrared patch-image model for small target detection in a single image. *IEEE Transactions on Image Processing*, 22(12):4996–5009, 2013.

[7] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. In *British Machine Vision Conference (BMVC)*, pages 1–13, 2018.

[8] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 510–519, 2019.

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

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, Cham, 2014.

[11] Yao Qin, Lorenzo Bruzzone, Chengqiang Gao, and Biao Li. Infrared small target detection based on facet kernel and random walker. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):7104–7118, 2019.

[12] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing*, pages 234–244, 2016.

[13] Xiaoyang Wang, Zhenming Peng, Dehui Kong, and Yanmin He. Infrared dim and small target detection based on stable multisubspace learning in heterogeneous scene. *IEEE Transactions on Geoscience and Remote Sensing*, 55(10):5481–5493, 2017.

[14] Yantao Wei, Xinge You, and Hong Li. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognition*, 58:216–226, 2016.