Mapping Degeneration Meets Label Evolution: Learning Infrared Small Target Detection with Single Point Supervision (*Supplemental Material*)

Xinyi Ying[1], Li Liu[1], Yingqian Wang[1], Ruojing Li[1], Nuo Chen[1], Zaiping Lin[1]✉,
Weidong Sheng[1], Shilin Zhou[1]
[1]National University of Defense Technology
{yingxinyi18, wangyingqian16, liruojing, chennuo97, linzaiping, slzhou}@nudt.edu.cn,
dreamliu2010@gmail.com, shengweidong1111@sohu.com

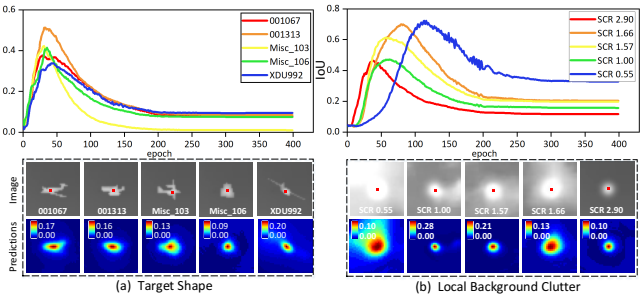

Figure I. IoU and visualize results of mapping degeneration with respect to different characteristics of targets ((a) shape and (b) local background clutter). We visualize the zoom-in target regions of input images with GT point labels (*i.e.,* red dots in images) and corresponding CNN predictions (in the epoch reaching maximum IoU).

Section I investigates the mapping degeneration phenomenon with respect to different characteristics of targets (*i.e.,* shape and local background clutter) for the analyses in Section 5.2.1-*Analyses of Mapping Degeneration*. Section II presents more visual results of labels and network predictions for the analyses in Section 5.2.2-*Effectiveness*. Section III provides additional discussion of the convergence issue for the analyses in Section 5.2.2-*Evolution Frequency*. Section IV includes additional comparison results for the analyses in Section 5.3. Section V provides comparison results with existing weakly-supervised segmentation methods.

I. Analyses of Mapping Degeneration

In this section, we investigate the mapping degeneration phenomenon with respect to different characteristics of targets (*i.e.,* shape, and local background clutter).

Target Shape. We simulate targets [9] with different shapes (*i.e.,* 001067, 001313, Misc_103, Misc_106, XDU992) to investigate the influence of target shape on mapping degeneration. Note that, we try to keep the target size and intensity unchanged when changing the target shape. Quantitative results in Fig. I(a) show that more concentrated shape results in higher maximum IoU. Visualization results show that CNNs can predict a cluster of pixels in a shape-aware manner, but can only recover the main body of targets without fine-grained details (*e.g.,* wings of drones in 001067 and XDU992).

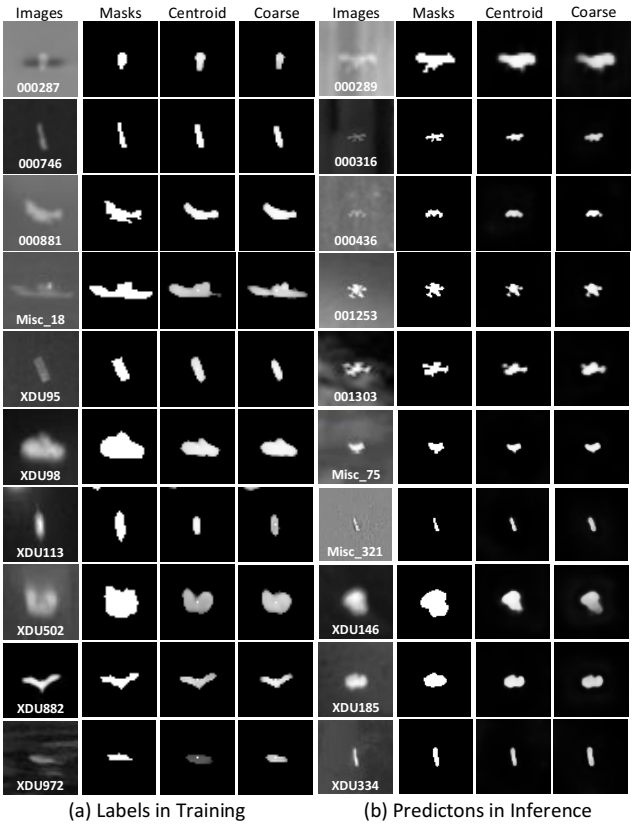

Figure II. Visualizations of regressed labels during training and network predictions during inference with centroid and coarse point supervision.

Local Background Clutter. We simulate a Gaussian-

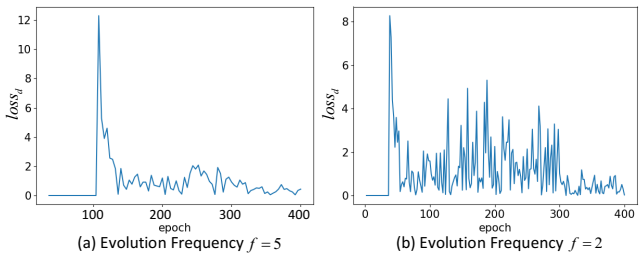

Figure III. $loss_d$ with respect to evolution frequency f=5 and f=2.

based extended target (with intensity 100 & radius 7), and add them to different locations of the background image to investigate the influence of local background clutter on mapping degeneration. We employ SCR of the local neighborhood to quantify the local background clutter. Results in Fig. I(b) show that background clutters significantly change the observed target appearance in size, shape, and contrast, and our method can only predict the high-contrast regions in the input images. Therefore, high-contrast background clutters introduce false alarms, and thus degrade the detection performance.

II. Visual Results of Labels and Network Predictions

In this section, we provides additional visual results of regressed labels during training and network predictions during inference on NUAA-SIRST [3], NUDT-SIRST [9], and IRSTD-1K [18] datasets. It can be observed from Figure II that our LESPS can effectively regress mask labels during training, and can achieve accurate pixel-level SIRST detection in the inference stage.

III. Discussion of Convergence Issue

In this section, we discuss the convergence issue of our label evolution framework (*i.e.,* LESPS). Specifically, we calculate the focal loss between evolved labels and network predictions before and after label update, and use their absolute difference (*i.e.,* $loss_d$) to measure the proximity degree between predictions and labels. Fig. III shows $loss_d$ of each update with evolution frequency f=5 and f=2 (*i.e.,* update every 5 and 2 epochs). It can be observed that $loss_d$ is gradually reduced in general, which reveals that predictions gradually approximate labels and networks can converge steadily. In addition, the training process of $loss_d$ with f=5 is more steady than that of f=2. This is because, a relatively lower update frequency (*i.e.,* f=5) represents more training time before label update, and thus stabilize the training process.

Table I. Average IoU ($\times 10^2$), P_d ($\times 10^2$), F_a($\times 10^6$) values on 3 public datasets [3, 9, 18] of DNA-Net trained with pseudo labels generated by input intensity threshold, LCM-based methods [6, 7, 10] and LESPS under centroid and coarse point supervision. Best results are shown in boldface.

Pseudo Label	Centroid			Coarse		
	IoU	P_d	F_a	IoU	P_d	F_a
Threshold=0.3	4.92	81.78	13.18	5.67	83.12	11.98
Threshold=0.5	13.24	73.08	5.31	15.54	76.03	4.89
Threshold=0.7	14.51	45.50	4.28	15.21	46.88	3.84
RLCM [6]	21.43	89.10	2.67	22.53	90.56	3.69
WSLCM [8]	8.68	86.64	50.10	8.89	84.45	80.24
TLLCM [7]	21.95	90.96	7.72	26.05	**94.15**	4.27
MSLCM [11]	31.43	**93.16**	**2.50**	36.32	92.43	**1.17**
MSPCM [10]	28.89	92.62	3.84	29.79	93.95	2.28
Ours	**57.34**	91.87	20.24	**56.18**	91.49	18.32

Table II. Average IoU($\times 10^2$), P_d($\times 10^2$), F_a($\times 10^6$) values of different methods. "#Params." represents the number of parameters. Best results are shown in boldface.

Method	Annotations per object	#Params.	IoU	P_d	F_a
MaskRCNN+ [C3]	10 points in bbox	88.6M	51.30	**94.38**	82.77
PointRend [C4]	100+ elaborated points	120.3M	56.02	94.30	61.48
Implicit PointRend [C3]	10 points in bbox	700.0M	52.00	94.13	85.79
DNA-Net+LESPS (Ours)	1 coarse point	**4.8M**	**56.18**	91.49	**18.32**

IV. Quantitative and Qualitative Results

IV.1. Comparison to SISRT Detection Methods

Table III provides additional comparative results of six traditional methods (Max-Median [4], WSLCM [8], WSPCM [8], NRAM [16], RIPT [1], MSLSTIPT [13]). It can be observed that CNN-based methods equipped with LESPS can outperform all the traditional methods, and can achieve over 70% IoU and comparable P_d, F_a values of their fully supervised counterparts.

Figure IV provides ROC results of ACM [2], ALCNet [3], DNA-Net [9] equipped with LESPS under centroid point supervision (*i.e.,* ACM Centroid+, ALCNet Centroid+, DNA-Net Centroid+) and their fully supervised counterparts (*i.e.,* ACM, ALCNet, DNA-Net) achieved on NUAA-SIRST [3], NUDT-SIRST [9], and IRSTD-1K [18] datasets. It can be observed that ROC results of ACM Centroid+, ALCNet Centroid+, DNA-Net Centroid+, and ACM, ALCNet, DNA-Net only have minor differences (*i.e.,* less than 5%).

Figure V provides additional qualitative results. It can be observed that CNN-based methods equipped with LESPS can produce outputs with precise target mask and low false alarm rate, and can generalize well to complex scenes.

IV.2. Comparison to Fixed Pseudo Labels

Table I provides additional comparisons to more LCM-based pseudo labels. It can be observed that, compared with LCM-based pseudo labels, DNA-Net with LESPS can achieve the highest IoU values with comparable P_d and reasonable F_a increase.

Table III. IoU ($\times 10^2$), P_d ($\times 10^2$) and F_a($\times 10^6$) values of different methods achieved on NUAA-SIRST [3] NUDT-SIRST [9] and IRSTD-1K [18] datasets. "CNN Full", "CNN Centroid", and "CNN Coarse" represent CNN-based methods under full supervision, centroid and coarse point supervision. "+" represents CNN-based methods equipped with LESPS.

Methods	Description	NUAA-SIRST [3]			NUDT-SIRST [9]			IRSTD-1K [18]			Average		
		IoU	P_d	F_a	IoU	P_d	F_a	IoU	P_d	F_a	IoU	P_d	F_a
Top-Hat [12]	Filtering	7.14	79.84	1012.00	20.72	78.41	166.70	10.06	75.11	1432.00	12.64	77.79	870.23
Max-Median [4]	Filtering	4.17	69.20	55.33	4.20	58.41	36.89	7.00	65.21	59.73	5.12	64.27	50.65
RLCM [6]	Local Contrast	21.02	80.61	199.15	15.14	66.35	163.00	14.62	65.66	17.95	16.06	68.70	98.77
WSLCM [8]	Local Contrast	1.02	80.99	45846.16	0.85	74.60	52391.63	0.99	70.03	15027.08	0.91	74.82	33759.07
TLLCM [7]	Local Contrast	11.03	79.47	7.27	7.06	62.01	46.12	5.36	63.97	4.93	7.22	65.45	21.42
MSLCM [11]	Local Contrast	11.56	78.33	8.37	6.65	56.83	25.62	5.35	59.93	5.41	7.07	61.20	13.74
MSPCM [10]	Local Contrast	12.38	83.27	17.77	5.86	55.87	115.96	7.33	60.27	15.24	7.23	61.53	55.13
IPI [5]	Low Rank	25.67	85.55	11.47	17.76	74.49	41.23	27.92	81.37	16.18	23.78	80.47	22.96
NRAM [16]	Low Rank	12.16	74.52	13.85	6.93	56.40	19.27	15.25	70.68	16.93	11.45	67.20	16.68
RIPT [1]	Low Rank	11.05	79.08	22.61	29.44	91.85	344.30	14.11	77.55	28.31	18.20	82.83	131.74
PSTNN [17]	Low Rank	22.40	77.95	29.11	14.85	66.13	44.17	24.57	71.99	35.26	20.61	72.02	36.18
MSLSTIPT [13]	Low Rank	10.30	82.13	1131.00	8.34	47.40	888.10	11.43	79.03	1524.00	10.02	69.52	1181.03
MDvsFA [14]	CNN Full	61.77	92.40	64.90	45.38	86.03	200.71	35.40	85.86	99.22	47.52	88.10	121.61
ISNet [18]	CNN Full	72.04	94.68	42.46	71.27	96.93	96.84	60.61	94.28	61.28	67.97	95.30	66.86
UIU-Net [15]	CNN Full	69.90	95.82	51.20	75.91	96.83	18.61	61.11	92.93	26.87	68.97	95.19	32.23
ACM [2]	CNN Full	64.92	90.87	12.76	57.42	91.75	39.733	57.49	91.58	43.86	59.94	91.40	32.12
	CNN Centroid+	49.23	89.35	40.95	42.09	91.11	38.24	41.44	88.89	60.46	44.25	89.78	46.55
	CNN Coarse+	47.81	88.21	40.75	40.64	81.11	49.45	40.37	92.59	64.81	42.94	87.30	51.67
ALCNet [3]	CNN Full	67.91	92.78	37.04	61.78	91.32	36.36	62.03	90.91	42.46	63.91	91.67	38.62
	CNN Centroid+	50.62	92.02	36.84	41.58	92.28	67.01	44.90	90.57	84.68	45.70	91.62	62.84
	CNN Coarse+	51.00	90.87	42.40	44.14	92.80	32.10	46.75	92.26	64.30	47.30	91.98	46.27
DNA-Net [9]	CNN Full	76.86	96.96	22.5	87.42	98.31	24.5	62.73	93.27	21.81	75.67	96.18	22.94
	CNN Centroid+	61.95	92.02	18.17	57.99	94.71	26.45	52.09	88.88	16.09	57.34	91.87	20.24
	CNN Coarse+	61.13	93.16	11.87	58.37	93.76	28.01	49.05	87.54	15.07	56.18	91.49	18.32

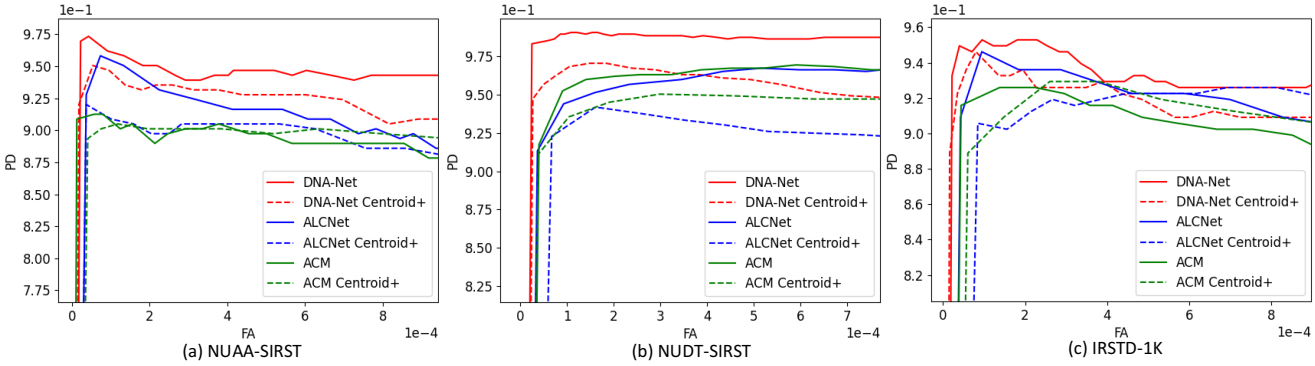

Figure IV. ROC results of different methods achieved on (a) NUAA-SIRST [3], (b) NUDT-SIRST [9], and (c) IRSTD-1K [18] datasets. "Centroid+" represents CNN-based methods equipped with LESPS under centroid point supervision.

V. Comparison to Existing Weakly-Supervised Segmentation Methods

We equip DNA-Net with LESPS, and compare with existing weakly-supervised segmentation methods[1]. Results are shown in Table II. It can be observed that general weakly-supervised segmentation methods require much more annotation effort and computational cost (*i.e.,* 18-146 times of our method) but the performance is comparable or worse. It is demonstrated that different from general methods, point-supervised SISRT detection has its unique characteristics, and needs further exploration.

[1] All models are implemented by their officially public codes.

References

[1] 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 (JSTAR)*, 10(8):3752–3767, 2017.

[2] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric contextual modulation for infrared small target detection. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[3] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Attentional local contrast networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, pages 1–12, 2021.

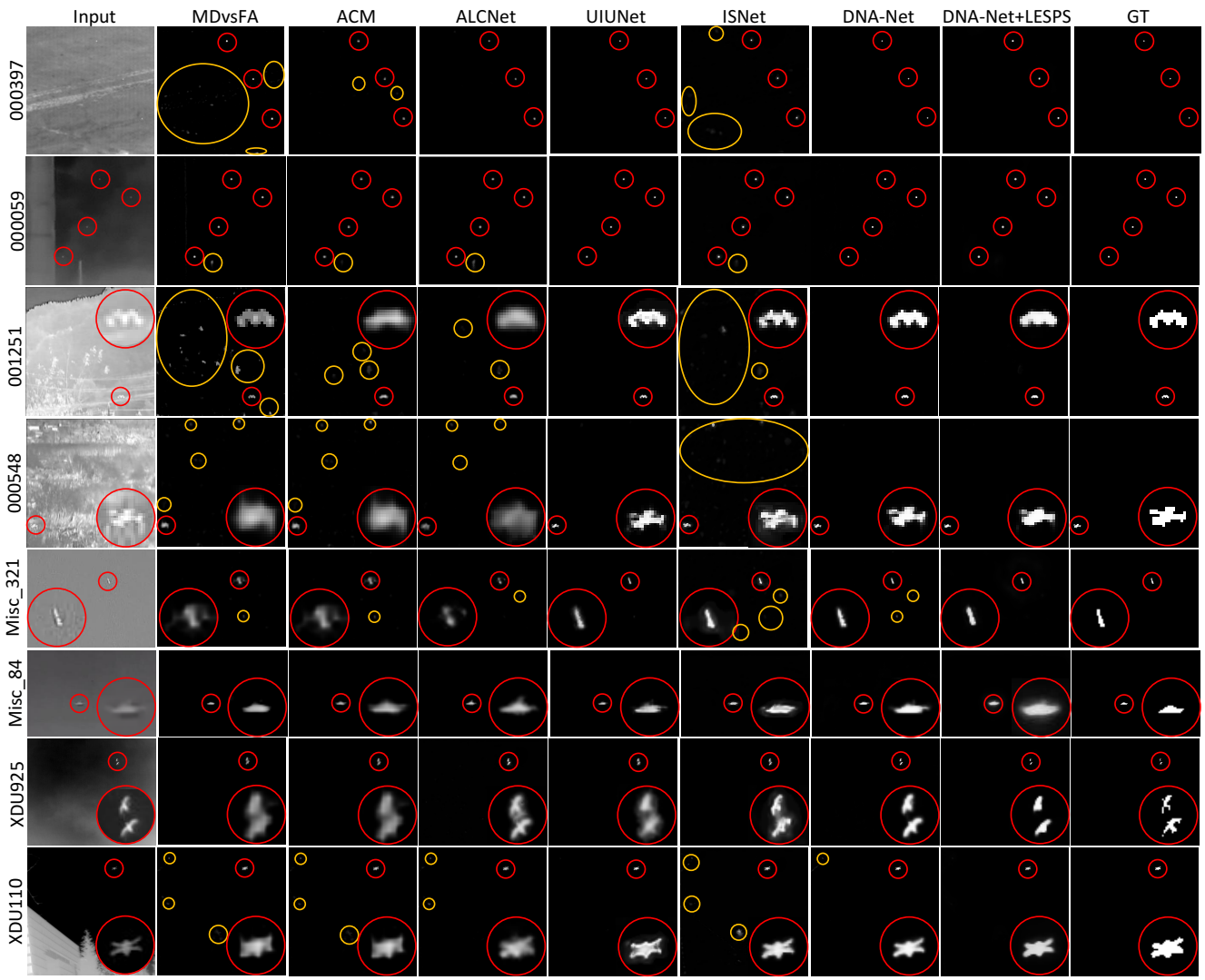

Figure V. Visual detection results of different methods achieved on NUAA-SIRST [3], NUDT-SIRST [9], and IRSTD-1K [18] datasets. Correctly detected targets and false alarms are highlighted by red and orange circles, respectively.

[4] Suyog D Deshpande, Meng Hwa Er, Ronda Venkateswarlu, and Philip Chan. Max-mean and max-median filters for detection of small targets. In *Signal and Data Processing of Small Targets*, volume 3809, pages 74–83. SPIE, 1999.

[5] 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 (TIP)*, 22(12):4996–5009, 2013.

[6] Jinhui Han, Kun Liang, Bo Zhou, Xinying Zhu, Jie Zhao, and Linlin Zhao. Infrared small target detection utilizing the multiscale relative local contrast measure. *IEEE Geoscience and Remote Sensing Letters (GRSL)*, 15(4):612–616, 2018.

[7] Jinhui Han, Saed Moradi, Iman Faramarzi, Chengyin Liu, Honghui Zhang, and Qian Zhao. A local contrast method for infrared small-target detection utilizing a tri-layer window. *IEEE Geoscience and Remote Sensing Letters (GRSL)*, 17(10):1822–1826, 2019.

[8] Jinhui Han, Saed Moradi, Iman Faramarzi, Honghui Zhang, Qian Zhao, Xiaojian Zhang, and Nan Li. Infrared small target detection based on the weighted strengthened local contrast measure. *IEEE Geoscience and Remote Sensing Letters (GRSL)*, 18(9):1670–1674, 2020.

[9] Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing (TIP)*, 2022.

[10] Saed Moradi, Payman Moallem, and Mohamad Farzan Sabahi. A false-alarm aware methodology to develop robust and efficient multi-scale infrared small target detection algorithm. *Infrared Physics & Technology*, 89:387–397, 2018.

[11] Saed Moradi, Payman Moallem, and Mohamad Farzan Sabahi. A false-alarm aware methodology to develop robust and efficient multi-scale infrared small target detection algorithm. *Infrared Physics & Technology*, 89:387–397, 2018.

[12] Jeanfrancois Rivest and Roger Fortin. Detection of dim targets in digital infrared imagery by morphological image processing. *Optical Engineering*, 35(7):1886–1893, 1996.

[13] Yang Sun, Jungang Yang, and Wei An. Infrared dim and small target detection via multiple subspace learning and spatial-temporal patch-tensor model. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 59(5):3737–3752, 2020.

[14] Huan Wang, Luping Zhou, and Lei Wang. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8509–8518, 2019.

[15] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. Uiu-net: u-net in u-net for infrared small object detection. *IEEE Transactions on Image Processing (TIP)*, 32:364–376, 2022.

[16] Landan Zhang, Lingbing Peng, Tianfang Zhang, Siying Cao, and Zhenming Peng. Infrared small target detection via non-convex rank approximation minimization joint l2, 1 norm. *Remote Sensing*, 10(11):1821, 2018.

[17] Landan Zhang and Zhenming Peng. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sensing*, 11(4):382, 2019.

[18] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. Isnet: Shape matters for infrared small target detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 877–886, 2022.