Mapping Degeneration Meets Label Evolution: Learning Infrared Small Target Detection with Single Point Supervision

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

Abstract

Training a convolutional neural network (CNN) to detect infrared small targets in a fully supervised manner has gained remarkable research interests in recent years, but is highly labor expensive since a large number of per-pixel annotations are required. To handle this problem, in this paper, we make the first attempt to achieve infrared small target detection with point-level supervision. Interestingly, during the training phase supervised by point labels, we discover that CNNs first learn to segment a cluster of pixels near the targets, and then gradually converge to predict groundtruth point labels. Motivated by this "mapping degeneration" phenomenon, we propose a label evolution framework named label evolution with single point supervision (LESPS) to progressively expand the point label by leveraging the intermediate predictions of CNNs. In this way, the network predictions can finally approximate the updated pseudo labels, and a pixel-level target mask can be obtained to train CNNs in an end-to-end manner. We conduct extensive experiments with insightful visualizations to validate the effectiveness of our method. Experimental results show that CNNs equipped with LESPS can well recover the target masks from corresponding point labels, and can achieve over 70% and 95% of their fully supervised performance in terms of pixel-level intersection over union (IoU) and object-level probability of detection (P_d), respectively. Code is available at https://github.com/XinyiYing/LESPS.

1. Introduction

Infrared small target detection has been a longstanding, fundamental yet challenging task in infrared search and tracking systems, and has various important applications in civil and military fields [49,57], including traffic monitoring [24, 54], maritime rescue [52, 53] and military surveillance [7, 47]. Due to the rapid response and robustness to fast-moving scenes, single-frame infrared small target (SIRST) detection methods have always attracted much more attention, and numerous methods have been proposed. Early methods, including filtering-based [9, 40], local contrast-based [3, 16] and low rank-based [11, 43] methods, require complex handcrafted features with carefully tuned hyper-parameters. Recently, compact deep learning has been introduced in solving the problem of SIRST detection [24, 45, 54]. However, there are only a few attempts, and its potential remains locked, unlike the extensive explorations of deep learning for natural images. This is mainly due to potential reasons, including lack of large-scale, accurately annotated datasets and high stake application scenarios.

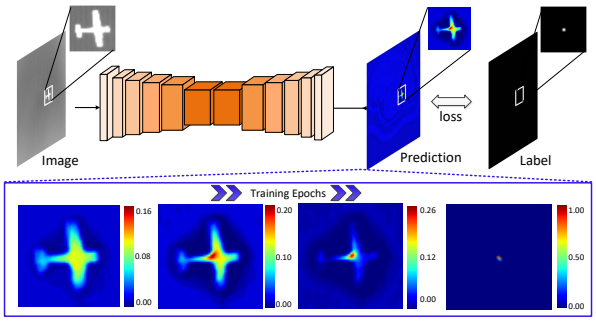

Figure 1. An illustration of mapping degeneration under point supervision. CNNs always tend to segment a cluster of pixels near the targets with low confidence at the early stage, and then gradually learn to predict GT point labels with high confidence.

Infrared small targets are usually of very small size, weak, shapeless and textureless, and are easily submerged in diverse complex background clutters. As a result, directly adopting existing popular generic object detectors like RCNN series [13, 14, 19, 39], YOLO series [25, 37, 38] and SSD [29] to SIRST detection cannot produce satisfactory performance. Realizing this, researchers have been focusing on developing deep networks tailored for

This work was supported by National Key Research and Development Program of China No. 2021YFB3100800.

infrared small targets by adequately utilizing the domain knowledge. However, most existing deep methods for SIRST detection [8, 24, 54] are fully supervised, which usually requires a large dataset with accurate target mask annotations for training. Clearly, this is costly [5, 26].

Therefore, a natural question arises: *Can we develop a new framework for SIRST detection with single point supervision*? In fact, to substantially reduce the annotation cost for object detection tasks, weakly supervised object detection methods with point supervision [4, 5, 26, 56] have been studied in the field of computer vision. Although these weakly supervised methods achieve promising results, they are not designed for the problem of SIRST detection, and the class-agnostic labels (*i.e.,* only foreground and background) of infrared small targets hinder their applications [42, 58]. Therefore, in this work, we intend to conduct the first study of weakly supervised SIRST detection with single-point supervision.

A key motivation of this work comes from an interesting observation during the training of SIRST detection networks. That is, with single point labels serving as supervision, CNNs always tend to segment a cluster of pixels near the targets with low confidence at the early stage, and then gradually learn to predict groundtruth (GT) point labels with high confidence, as shown in Fig. 1. It reveals the fact that region-to-region mapping is the intermediate result of the final region-to-point mapping[1]. We attribute this "mapping degeneration" phenomenon to the special imaging mechanism of infrared system [24, 54], the local contrast prior of infrared small targets [3, 8], and the easy-to-hard learning property of CNNs [44], in which the first two factors result in extended mapping regions beyond the point labels, and the last factor contributes to the degeneration process.

Based on the aforementioned discussion, in this work, we propose a novel framework for the problem of weakly supervised SIRST detection, dubbed label evolution with single point supervision (LESPS). Specifically, LESPS leverages the intermediate network predictions in the training phase to update the current labels, which serve as supervision until the next label update. Through iterative label update and network training, the network predictions can finally approximate the updated pseudo mask labels, and the network can be simultaneously trained to achieve pixel-level SIRST detection in an end-to-end[2] manner.

Our main contributions are summarized as: (1) We present the first study of weakly supervised SIRST detection, and introduce LESPS that can significantly reduce the annotation cost. (2) We discover the mapping degeneration phenomenon, and leverage this phenomenon to automatically regress pixel-level pseudo labels from the given point labels via LESPS. (3) Experimental results show that our framework can be applied to different existing SIRST detection networks, and enable them to achieve over 70% and 95% of its fully supervised performance in terms of pixel-level intersection over union (IoU) and object-level probability of detection (P_d), respectively.

[1] "region-to-region mapping" represents the mapping learned by CNNs from target regions in images to a cluster of pixels near the targets, while "region-to-point mapping" represents the mapping from target regions in images to the GT point labels.

[2] Different from generic object detection [32, 59], "end-to-end" here represents achieving point-to-mask label regression and direct pixel-level inference in once training.

2. Related Work

SIRST Detection. In the past decades, various methods have been proposed, including early traditional paradigms (*e.g.,* filtering-based methods [9, 40], local contrast-based methods [3, 15–17, 33, 34], low rank-based methods [6, 11, 28, 43, 50, 51]) and recent deep learning paradigms [7, 8, 20, 21, 24, 45, 52–54]. Compared to traditional methods, which require delicately designed models and carefully tuned hyper-parameters, convolutional neural networks (CNNs) can learn the non-linear mapping between input images and GT labels in a data-driven manner, and thus generalize better to real complex scenes. As the pioneering work, Wang *et al.* [45] first employed a generative adversarial network to achieve a better trade-off between miss detection and false alarm. Recently, more works focus on customized solutions of infrared small target. Specifically, Dai *et al.* [7] specialized an asymmetric contextual module, and further incorporated local contrast measure [8] to improve the target contrast. Li *et al.* [24] preserved target information by repetitive feature fusion. Zhang *et al.* [54] aggregated edge information to achieve shape-aware SIRST detection. Zhang *et al.* [52, 53] explored cross-level correlation and transformer-based method [18] to predict accurate target mask. Wu *et al.* [46] customized a UIU-Net framework for multi-level and multi-scale feature aggregation. In conclusion, existing works generally focus on compact architectural designs to pursue superior performance in a fully supervised manner. However, due to the lack of a large number of public datasets [7, 24, 45] with per-pixel annotations, the performance and generalization of CNNs are limited. In addition, per-pixel manual annotations are time-consuming and labor-intensive. Therefore, we focus on achieving good pixel-level SIRST detection with weaker supervision and cheaper annotations.

Weakly Supervised Segmentation with Points. Recently, point-level annotation has raised more attention in dense prediction tasks such as object detection [4, 12, 56], crowd counting [1, 23, 30, 48] and image segmentation [2, 5, 10, 26, 31, 36, 55]. We mainly focus on image segmentation in this paper. Specifically, Bearman *et al.* [2] made the first attempt to introduce an objectiveness potential into a pointly supervised training

loss function to boost segmentation performance. Qian *et al.* [36] leveraged semantic information of several labeled points by a distance metric loss to achieve scene parsing. Zhang *et al.* [55] proposed an inside-outside guidance approach to achieve instance segmentation by five elaborate clicks. Cheng *et al.* [5] designed to provide ten randomly sampled binary point annotations within box annotations for instance segmentation. Li *et al.* [26] encoded each instance with kernel generator for panoptic segmentation to achieve 82% of fully-supervised performance with only twenty randomly annotated points. In contrast to these approaches employing complicated prior constraints to segment large generic objects with rich color and fine textures by several elaborate points, we fully exploit the local contrast prior of infrared small target to progressively evolve pseudo masks by single coarse point without any auxiliaries in an end-to-end manner.

3. The Mapping Degeneration Phenomenon

In this section, we first describe the mapping degeneration phenomenon together with our intuitive explanation. Then we conduct experiments under single-sample and many-sample training schemes to demonstrate the generality of degeneration, and investigate the influence of generalization on degeneration.

As shown in Fig. 1, given an input image and the corresponding GT point label, we employ U-Net [41] as the baseline SIRST detection network for training. It can be observed that, in the early training phase, network predicts a cluster of pixels near the targets with low confidence. As training continues, the network prediction finally approximates GT point label with gradually increased confidence. We name this phenomenon as "mapping degeneration", and attribute the following reasons to this phenomenon. 1) *Special imaging mechanism of infrared systems* [24, 54]: Targets only have intensity information without structure and texture details, resulting in highly similar pixels within the target region. 2) *High local contrast of infrared small targets* [3, 8]: Pixels within the target region are much brighter or darker with high contrast against the local background clutter. 3) *Easy-to-hard learning property of CNNs* [44]: CNNs always tend to learn simple mappings first, and then converge to difficult ones. Compared with region-to-point mapping, region-to-region mapping is easier, and thus tends to be the intermediate result of region-to-point mapping. In conclusion, the unique characteristics of infrared small targets result in extended mapping regions beyond point labels, and CNNs contribute to the mapping degeneration process.

It is worth noting that the mapping degeneration phenomenon is a general phenomenon in various scenes with infrared small targets. Specifically, we use the training datasets (including 1676 images and their corresponding

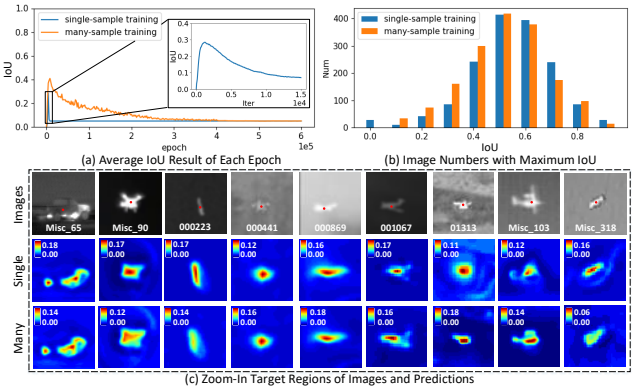

Figure 2. Quantitative and qualitative illustrations of mapping degeneration in CNNs.

centroid point label, see details in Section 5.1) to train U-Net under a single-sample training scheme (*i.e.,* training one CNN on each image). For quantitative analyses, we employ the IoU results between positive pixels in predictions (*i.e.,* pixels with confidence higher than half of its maximum value) and GT mask label. Average IoU results of 1676 CNNs at each epoch are shown by the blue curve in Fig. 2(a), while the number of training samples with maximum IoU during training phase falling in a threshold range of $[i, i+0.1], (i = 0, 0.1, \cdots, 0.9)$ is illustrated via blue bars in Fig. 2(b). It can be observed from the zoom-in curve and bars that mapping degeneration is a general phenomenon with point supervision, and U-Net can achieve $IoU > 0.5$ on more than 60% of the training images.

In addition, we conduct experiments to train U-Net under a many-sample training scheme (*i.e.,* training one CNN using all images which contain abundant targets with various sizes and shapes) to investigate the effect of generalization on mapping degeneration. Average IoU results of 1676 images are shown by orange curve in Fig. 2(a). It can be observed that many-sample training scheme needs more time to converge. Moreover, Fig. 2(b) shows that orange bars are slightly lower than blue ones on larger IoU values (*i.e.,* 0.5-1.0). It is demonstrated that generalization decelerates but aggravates mapping degeneration. Figure 2(c) shows some zoom-in target regions of images and their predictions under these two training schemes. It can be observed that CNNs can effectively segment a cluster of target pixels under both training schemes in a size-aware manner.

Therefore, an intuitive assumption arises: Can we leverage the intermediate results of CNNs to regress masks? A simple early stopping strategy seems to be a positive answer but is indeed unpractical since mapping degeneration is influenced by various factors, including target intensity, size, shape, and local background clutter (see details in Section 5.2.1). Consequently, there is no fixed optimal stopping epoch for all situations. These

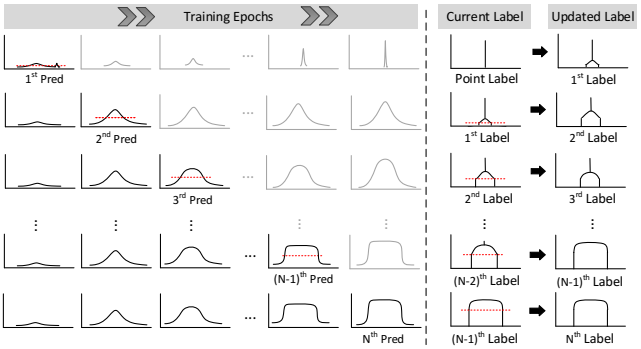

Figure 3. An illustration of label evolution with single point supervision (LESPS). During training, intermediate predictions of CNNs are used to progressively expand point labels to mask labels. Black arrows represent each round of label updates.

observations motivate us to design a special label evolution framework to well leverage the mapping degeneration for pseudo mask regression.

4. The Label Evolution Framework

Motivated by mapping degeneration, we propose a label evolution framework named label evolution with single point supervision (LESPS) to leverage the intermediate network predictions in the training phase to update labels. As training continues, the network predictions approximate the updated pseudo mask labels, and network can simultaneously learn to achieve pixel-level SIRST detection in an end-to-end manner. Here, we employ a toy example of 1D curves for easy understanding. As shown in Fig. 3, sub-figures on the left of the dotted line represent the network predictions. Note that, the black curves denote the intermediate predictions within LESPS, while the gray curves represent virtual results produced by the network without label update. On the right of the dotted line, the first and second columns of sub-figures represent current labels and updated labels, respectively, and black arrows represent each round of label update. The overall framework can be summarized as follows. With point label serving as supervision, in the 1^{st} round label update after initial training, the predictions are used to update the current point label to generate the 1^{st} updated label, which is then used to supervise the network training until the 2^{nd} round label update. Through iterative label updates and network training, CNNs can incorporate the local contrast prior to gradually recover the mask labels. From another viewpoint, label evolution consistently updates the supervision to prevent mapping degeneration, and promotes CNNs to converge to the easy region-to-region mapping.

Taking the n^{th} update as an example, given the current label L_n and the network prediction P_n, we perform label update for each target, which consists of three steps: candidate pixel extraction, false alarm elimination, and weighted summation between candidate pixels and current

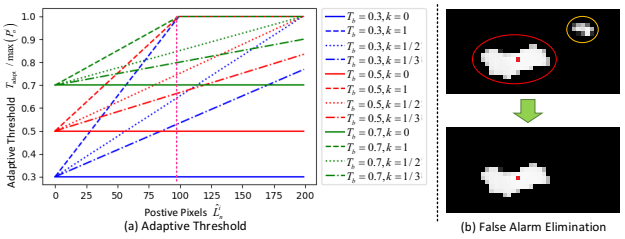

Figure 4. (a) Adaptive threshold T_{adapt} with respect to positive pixels $\hat{L}_n^i$ and hyper-parameters k, T_b. Pink dotted line represents the constant hwr. (b) An illustration of false alarm elimination. Red circle and dot represent positive pixels and centroid point of label. Orange circle represents false alarms.

labels. Specifically, the $d \times d$ local neighborhoods of the i^{th} target in label L_n and prediction P_n are cropped based on the centroid of the positive pixels[3] in label (*i.e.,* $\hat{L}_n$). Then to reduce error accumulation for label update (see Section 5.2.2 for details), we employ an adaptive threshold (the red dotted line in Fig. 3) to extract the local neighborhood candidate pixels (predictions higher than the red dotted line in Fig. 3). The process can be defined as:

$$C_n^i = P_n^i \odot (P_n^i > T_{adapt}), \tag{1}$$

where C_n^i is the candidate pixels, and $\odot$ represents element-wise multiplication. T_{adapt} is the adaptive threshold that correlated to the current prediction P_n^i and the positive pixels in label $\hat{L}_n^i$, and can be calculated according to:

$$T_{adapt} = max(P_n^i)(T_b + k(1 - T_b)\hat{L}_n^i/(hwr)), \tag{2}$$

where h, w are the height and width of input images, and r is set to 0.15% [7, 24]. As shown in Fig. 4 (a), T_b is the minimum threshold, and k controls the threshold growth rate. An increasing number of $\hat{L}_n^i$ leads to the increase of the threshold, which can reduce error accumulation of low contrast targets and strong background clutter.

To eliminate false alarms by local neighborhood noise, we exclude the eight connective regions of candidate pixels that have no intersection with positive pixels of labels, as shown in Fig. 4 (b). This process is defined as:

$$E_n^i = C_n^i \odot F_n^i, \tag{3}$$

where E_n^i is the candidate pixels after false alarm elimination, and F_n^i is the mask against false alarm pixels.

We then perform average weighted summation between candidate pixels E_n^i and current label L_n^i to achieve label update. The process can be formulated as:

$$L_{n+1}^i = L_n^i \odot (1 - N_n^i) + \frac{L_n^i + E_n^i}{2} \odot N_n^i, \tag{4}$$

where L_{n+1}^i is the updated label in the n^{th} round, which serves as new supervision for training in the $n + 1^{th}$ round,

[3]The value of a pixel is higher than 0.5, which represents that the pixel is more likely to be positive than negative [8, 24]

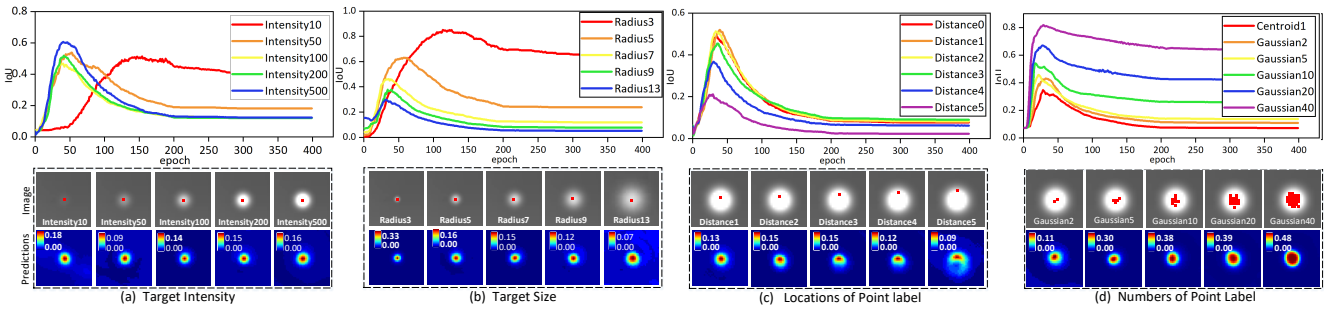

Figure 5. IoU and visualize results of mapping degeneration with respect to different characteristics of targets (*i.e.,*(a) intensity, (b) size,) and point labels (*i.e.,*(c) locations and (d) numbers). We visualize the zoom-in target regions of input images with GT point labels (*i.e.,* red dots in images) and corresponding CNN predictions (in the epoch reaching maximum IoU).

and $N_n^i = (P_n^i > T_{adapt}) \odot F_n^i$. Note that, the first term represents GT labels below red dotted lines, and the second term represents the average weighted summation between predictions and GT labels above red dotted lines.

It is worth noting that we provide three conditions to ensure network convergence: 1) Average weighted summation between predictions and labels promotes CNNs to converge as predictions approximate labels. 2) Pixel-adaptive threshold increases with the increase of positive pixels in updated labels, which slows down or suspends the label update. 3) As label evolution introduces more target information for training, CNNs grow to mature, and learn to distinguish targets from backgrounds.

We start label evolution after initial evolution epoch T_{epoch}, and perform label update every f epoch until the end of training. Note that, our epoch-based threshold T_{epoch} is a coarse threshold to ensure that networks attend to targets instead of background clutter.

5. Experiments

In this section, we first describe the implementation details, and then make comprehensive analyses of the mapping degeneration phenomenon and our label evolution framework. In addition, we apply our method to the state-of-the-art SIRST detection methods with point supervision, and make comparisons with their fully supervised counterparts. Moreover, we make comparisons of our dynamic updated pseudo labels with fixed pseudo labels, and discuss the calculation of loss function.

5.1. Implementation Details

Three public datasets NUAA-SIRST [8], NUDT-SIRST [24], and IRSTD-1K [54] are used in our experiments. We followed [24] to split the training and test sets of NUAA-SIRST and NUDT-SIRST, and followed [54] to split IRSTD-1K. We employed two pixel-level metrics (*i.e.,* intersection over union (IoU) and pixel accuracy (PA)) and two target-level metrics (*i.e.,* probability of detection (P_d) and false-alarm rate (F_a)) for performance evaluation.

During training, all images were normalized and randomly cropped into patches of size 256×256 as network inputs. We augmented the training data by random flipping and rotation. Due to the extreme positive-negative sample imbalance (less than 10 vs. more than 256×256) in SIRST detection with point supervision, we employed focal loss[4] [27] to stabilize the training process. All the networks were optimized by the Adam method [22]. Batch size was set to 16, and learning rate was initially set to 5×10^{-4} and reduced by ten times at the 200^{th} and 300^{th} epochs. We stopped training after 400 epochs. All models were implemented in PyTorch [35] on a PC with an Nvidia GeForce 3090 GPU.

5.2. Model Analyses

5.2.1 Analyses of Mapping Degeneration

We use synthetic images (simulated targets and real backgrounds [24]) to investigate the mapping degeneration phenomenon with respect to different characteristics of targets (*i.e.,* intensity and size[5]) and point labels (*i.e.,* locations and numbers). We employ U-Net [41] as the baseline detection network, and use centroid point as GT label if not specified. We calculate IoU between positive pixels in predictions and GT mask labels of each epoch for quantitative analyses. In addition, we visualize the zoom-in target regions of simulated images with GT point labels (*i.e.,* red dots) and corresponding CNN predictions (in the epoch reaching maximum IoU). To reduce training randomness, we show the average IoU results and visualization results over 100 runs.

Target Intensity. We simulate Gaussian-based extended targets with different peak values (*i.e.,* 10, 50, 100, 200, 500) to investigate the influence of target intensity on mapping degeneration. Quantitative results in Fig. 5(a) show that intensity higher than 100 leads to a positive correlation between intensity and maximum IoU, while

[4]Focal loss is calculated between current evolved and GT labels to supervise the network training until the next round label update.

[5]Shape, and local background clutter are investigated in supplemental material.

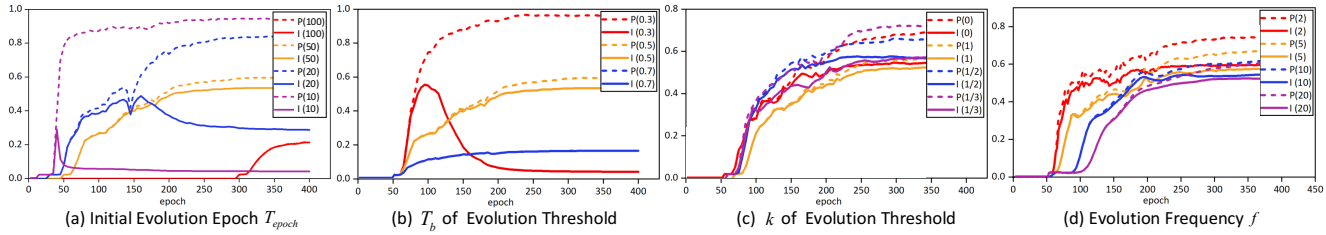

Figure 6. PA (P) and IoU (I) results of LESPS with respect to (a) initial evolution epoch T_{epoch}, (b) T_b and (c) k of evolution threshold, and (d) evolution frequency f.

Table 1. Average IoU ($\times 10^2$), P_d ($\times 10^2$) and F_a($\times 10^6$) values on NUAA-SIRST [8] NUDT-SIRST [24] and IRSTD-1K [54] achieved by DNA-Net with (w/) and without (w/o) LESPS under centroid, coarse point supervision together with full supervision.

Method	Centroid			Coarse			Full		
	IoU	P_d	F_a	IoU	P_d	F_a	IoU	P_d	F_a
w/o LESPS	5.12	89.19	0.68	2.96	49.89	0.30	75.67	96.18	22.94
w/ LESPS	57.34	91.87	20.24	56.18	91.49	18.32			

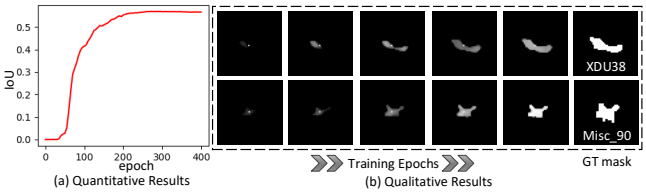

Figure 7. Quantitative and qualitative results of evolved target masks.

lower intensity leads to a negative one. In addition, curve "intensity10" reaches maximum IoU at around epoch 150 while others are less than 50, which demonstrates that over-small intensity decelerates degeneration. Visualization results show that our method can well highlight target regions under various intensities.

Target Size. We simulate Gaussian-based extended targets with different radii (*i.e.,* 3, 5, 7, 9, 13) to investigate the influence of target size on mapping degeneration. Quantitative results in Fig. 5(b) show that larger target size leads to lower maximum IoU and less time to reach maximum IoU. That is because, size discrepancy between targets and GT point labels increases as target size increases, which aggravates and accelerates mapping degeneration. Visualization results show that CNNs can predict a cluster of pixels in a size-aware manner, and the peak values of predictions decrease as target size increases.

Locations of Point Label. We simulate a Gaussian-based extended target (with intensity 500 & radius 13), and place point labels at different distances away from the target centroid to investigate the influence of point label locations on mapping degeneration. Results in Fig. 5(c) show that small perturbations of label locations (less than 3 pixels) have a minor influence on the maximum IoU results. However, severe location perturbations (larger than 3 pixels) can lead to a significant maximum IoU drop, and the drop is more obvious when the point label is close to the edge. Note that, the same targets with different label locations reach maximum IoU at the same time, which demonstrates that the speed of mapping degeneration is irrelevant to the position of labels.

Number of Points in Label. We simulate a Gaussian-based extended target (with intensity 500 & radius 13) and set different numbers of points in labels to investigate its influence on mapping degeneration. Quantitative results in Fig. 5(d) show that as the number of points increases, CNNs can learn more target information to achieve higher maximum IoU results. In addition, the speed of mapping degeneration is irrelevant to the point number. Visualization results show that peak values of predictions increase as the number of points increases, which demonstrates that stronger supervision alleviates mapping degeneration. The conclusion well supports our label evolution framework.

5.2.2 Analyses of the Label Evolution Framework

In this subsection, we conduct experiments to investigate the effectiveness and the optimal parameter settings of our label evolution framework (*i.e.,* LESPS). We employ PA and IoU between the positive pixels in updated labels and the GT mask labels to quantitatively evaluate the accuracy and expansion degree of the current label.

Effectiveness. We compare the average results of NUAA-SIRST [8], NUDT-SIRST [24], and IRSTD-1K [54] datasets achieved by DNA-Net with (*i.e.,* w/) and without (*i.e.,* w/o) LESPS under centroid and coarse point supervision, respectively. Note that, the results of DNA-Net w/o LESPS are calculated under extremely low threshold[6] (*i.e.,* 0.15) while those of DNA-Net w/ LESPS are calculated under the standard threshold (*i.e.,* 0.5 [24, 54]). As shown in Table 1, as compared to full supervision, the results of DNA-Net w/o LESPS are extremely low, which demonstrates that SIRST detection on single point supervision is a challenging task. In contrast, DNA-Net w/ LESPS can achieve significant performance improvement under both point supervisions in terms of all the metrics, which approximate the performance of their fully supervised counterparts. Note that, P_d results of DNA-

[6]With point supervision, the results of DNA-Net w/o LESPS calculated under threshold 0.5 are all zeros.

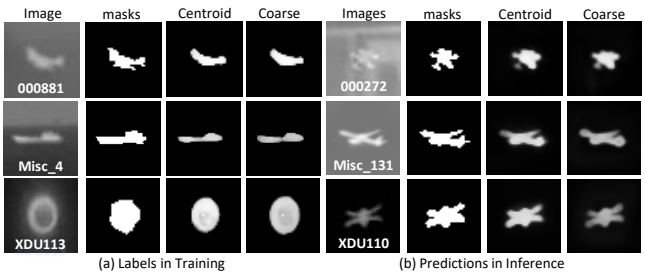

Figure 8. Visualizations of regressed labels during training and network predictions during inference with centroid and coarse point supervision.

Net w/o LESPS under coarse point supervision are over half lower than those under the centroid ones, while the results of DNA-Net w/ LESPS under these two kinds of point supervision are comparable. It demonstrates that LESPS can generalize well to manual annotation errors.

In addition, we make evaluations of evolved target masks during training. Quantitative results in Fig. 7(a) show average IoU values between positive pixels of evolved target masks and GT labels in 20-time training, which demonstrates that the quality of pseudo target masks consecutively increases during training. Qualitative results in Fig. 7(b) demonstrate that networks can effectively learn to expand point labels to mask labels. Furthermore, we visualize the labels regressed by our LESPS during training together with some network predictions during inference in Figs. 8 (a) and (b). As shown in Fig. 8(a), compared with GT mask labels, the evolved labels are more closely aligned with the objects in images (*e.g.,* GT masks of Misc_4, XDU113 exceed the target regions due to visual edge ambiguity), which demonstrates that LESPS can alleviate the manually annotated errors. Figure 8(b) shows that DNA-Net w/ LESPS can effectively achieve accurate pixel-level SIRST detection in an end-to-end manner. Please refer to the supplemental materials for more visual results.

Initial Evolution Epoch. We investigate the optimal value of epoch-based threshold T_{epoch}. Figure 6(a) shows that small initial evolution epoch results in a significant difference between PA and IoU (*i.e.,* 0.94 vs. 0.04 with T_{epoch}=10). That is because, early label evolution introduces many error pixels, resulting in severe error accumulation and network convergence failure. Increasing initial evolution epoch can reduce error accumulation and promote network convergence (0.60 vs. 0.54 with T_{epoch}=50). However, over-large initial evolution epoch (*i.e.,* a high degree of mapping degeneration) results in inferior performance (0.21 vs. 0.21 with T_{epoch}=100). Therefore, T_{epoch} is set to 50 in our method.

Evolution Threshold. We investigate the optimal values of k and T_b in the evolution threshold. T_b is the minimum threshold, and controls evolution speed and error accumulation degree. k determines the maximum

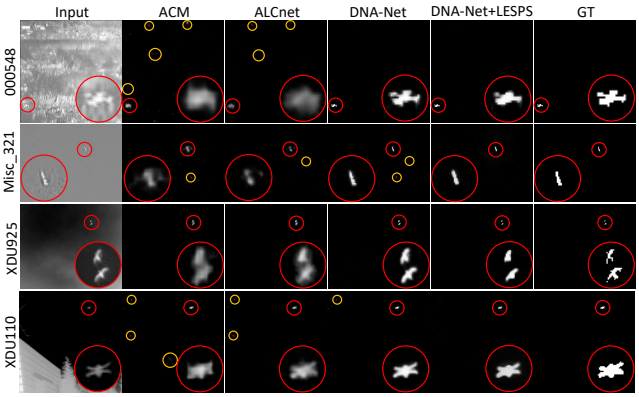

Figure 9. Visual detection results of different methods achieved on NUAA-SIRST [8], NUDT-SIRST [24] and IRSTD-1K [54] datasets. Correctly detected targets and false alarms are highlighted by red and orange circles, respectively.

threshold, and controls the growth rate of the threshold. As shown in Fig. 6(b) and 6(c), both over-large and over-small values of T_b and k result in inferior performance. Therefore, we choose k=1/2 and T_b=0.5 in our method.

Evolution Frequency. We investigate the optimal value of evolution frequency f. Figure 6(d) shows that evolution frequency is positively correlated to PA and IoU. However, high evolution frequency needs more time for label updates. To balance performance and efficiency, we choose f=5 in our method. Interestingly, higher frequency (*i.e.,* f=2) does not cause serve error accumulation, which also demonstrates the effectiveness of the convergence conditions of our LESPS. Please refer to the supplemental materials for more discussions of the convergence issue.

5.3. Comparison to State-of-the-art Methods

Comparison to SISRT detection methods. We apply our LESPS to several state-of-the-art CNN-based methods, including ACM [7], ALCNet [8] and DNA-Net [24]. For fair comparisons, we retrained all models on NUAA-SIRST [8], NUDT-SIRST [24], and IRSTD-1K [54] datasets with the same settings. In addition, we add the results of six fully supervised CNN-based methods (MDvsFA [45], ACM [7], ALCNet [8], DNA-Net [24], ISNet [54], UIU-Net [46]) and six traditional methods (Top-Hat [40], RLCM [15], TLLCM [16], MSPCM [17], IPI [11], PSTNN [51]) as the baseline results.

Quantitative results in Table 2 show that CNN-based methods equipped with LESPS can outperform all the traditional methods. In addition, they can also achieve 71-75% IoU results and comparable P_d and F_a results of their fully supervised counterparts. Qualitative results in Fig. 9 show 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. Please refer to supplemental materials for more quantitative

Table 2. IoU ($\times 10^2$), P_d ($\times 10^2$) and F_a($\times 10^6$) values of different methods achieved on NUAA-SIRST [8], NUDT-SIRST [24] and IRSTD-1K [54] 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 [8]			NUDT-SIRST [24]			IRSTD-1K [54]			Average		
		IoU	P_d	F_a	IoU	P_d	F_a	IoU	P_d	F_a	IoU	P_d	F_a
Top-Hat [40]	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
RLCM [15]	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
TLLCM [16]	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
MSPCM [34]	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 [11]	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
PSTNN [51]	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
MDvsFA [45]	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 [54]	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 [46]	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 [7]	CNN Full	64.92	90.87	12.76	57.42	91.75	39.73	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 [8]	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 [24]	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

Table 3. Average IoU ($\times 10^2$), P_d ($\times 10^2$), F_a($\times 10^6$) values on NUAA-SIRST [8], NUDT-SIRST [24] and IRSTD-1K [54] datasets of DNA-Net trained with pseudo labels generated by input intensity threshold, LCM-based methods [15, 16, 34] and LESPS under centroid and coarse point supervision.

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 [15]	21.43	89.10	**2.67**	22.53	90.56	3.69
TLLCM [16]	21.95	90.96	7.72	26.05	**94.15**	4.27
MSPCM [34]	28.89	**92.62**	3.84	29.79	93.95	**2.28**
LESPS(ours)	**57.34**	91.87	20.24	**56.18**	91.49	18.32

Table 4. Average IoU($\times 10^2$), P_d($\times 10^2$), F_a($\times 10^3$) values on NUAA-SIRST [8], NUDT-SIRST [24] and IRSTD-1K [54] datasets of different methods when computing negative loss on i handcrafted (hand$_i$), j randomly sampled (rand$_j$) and all background points.

Annotations	ACM			ALCNet			DNA-Net		
	IoU	P_d	F_a	IoU	P_d	F_a	IoU	P_d	F_a
hand$_1$	0.54	95.79	47.06	0.16	95.19	262.06	1.43	97.80	26.91
hand$_2$	0.12	**97.24**	295.17	0.15	96.62	248.05	3.41	**98.18**	8.48
hand$_5$	0.11	96.36	316.37	0.08	**97.29**	363.25	3.68	98.13	7.29
rand$_1$	8.06	93.45	3.56	8.57	92.97	3.21	18.74	94.69	0.58
rand$_2$	10.78	92.72	2.22	10.78	91.16	2.71	22.85	94.81	0.42
rand$_5$	**13.39**	92.66	1.35	**11.87**	93.26	0.89	**24.80**	95.00	0.34
All (Ours)	3.95	87.15	**0.02**	4.08	88.93	**0.02**	5.12	89.18	**0.01**

and qualitative results.

Comparison to other pseudo labels. We compare our dynamic updated pseudo labels with fixed pseudo labels generated by input intensity threshold and local contrast-based methods [15, 16, 34]. Specifically, given a GT point label, we only preserve the eight connected regions of detection results that have union pixels with the GT point label. Then, we employ the pseudo labels to retrain DNA-Net [24] from scratch. As shown in Table 3, compared with fixed pseudo labels, dynamic updated pseudo labels by LESPS can achieve the highest IoU values with comparable P_d and reasonable F_a increase.

5.4. Discussion of Loss Function

In this subsection, we investigate the loss function of computing negative loss on different background points. Average results of different baseline methods under centroid point supervision are shown in Table 4. Extremely limited handcrafted background points[7] leads to many false alarms. Random sample[8] introduces more background points and well alleviates class imbalance, resulting in great performance improvements. However, the above simple versions introduce huge false alarms (34-1.8K times of all points), which are not practical for real applications, but inspire further thorough investigation in the future.

[7]Points are sampled near targets, and are fixed during training.

[8]Points are randomly sampled, and change in each training iteration.

6. Conclusion

In this paper, we proposed the first work to achieve weakly-supervised SIRST detection using single-point supervision. In our method, we discovered the mapping degeneration phenomenon and proposed a label evolution framework named label evolution with single point supervision (LESPS) to automatically achieve point-to-mask regression. Through LESPS, networks can be trained to achieve SIRST detection in an end-to-end manner. Extensive experiments and insightful visualizations have fully demonstrated the effectiveness and superiority of our method. In addition, our method can be applied to different networks to achieve over 70% and 95% of its fully supervised performance on pixel-level IoU and object-level P_d, respectively. We hope our interesting findings and promising results can inspire researchers to rethink the feasibility of achieving state-of-the-art performance in SIRST detection with much weaker supervision.

References

[1] Peri Akiva, Kristin Dana, Peter Oudemans, and Michael Mars. Finding berries: Segmentation and counting of cranberries using point supervision and shape priors. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 50–51, 2020.

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision (ECCV)*, pages 549–565. Springer, 2016.

[3] CL 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 (TGRS)*, 52(1):574–581, 2013.

[4] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, and Jian Sun. Points as queries: Weakly semi-supervised object detection by points. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8823–8832, 2021.

[5] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2617–2626, 2022.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Pointly-supervised panoptic segmentation. *ArXiv Preprint ArXiv:2210.13950*, 2022.

[11] 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.

[12] Yongtao Ge, Qiang Zhou, Xinlong Wang, Chunhua Shen, Zhibin Wang, and Hao Li. Point-teaching: Weakly semi-supervised object detection with point annotations. *ArXiv Preprint ArXiv:2206.00274*, 2022.

[13] Ross Girshick. Fast r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[15] 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.

[16] 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.

[17] 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.

[18] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:15908–15919, 2021.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[20] Qingyu Hou, Zhipeng Wang, Fanjiao Tan, Ye Zhao, Haoliang Zheng, and Wei Zhang. Ristdnet: Robust infrared small target detection network. *IEEE Geoscience and Remote Sensing Letters (GRSL)*, 19:1–5, 2021.

[21] Moran Ju, Jiangning Luo, Guangqi Liu, and Haibo Luo. Istdet: An efficient end-to-end neural network for infrared small target detection. *Infrared Physics & Technology*, 114:103659, 2021.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[23] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *European Conference on Computer Vision (ECCV)*, pages 547–562, 2018.

[24] 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.

[25] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *ArXiv Preprint ArXiv:2209.02976*, 2022.

[26] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Yukang Chen, Lu Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation with point-based supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[28] Ting Liu, Jungang Yang, Boyang Li, Chao Xiao, Yang Sun, Yingqian Wang, and Wei An. Non-convex tensor low-rank approximation for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 2021.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, pages 21–37. Springer, 2016.

[30] Yongtuo Liu, Dan Xu, Sucheng Ren, Hanjie Wu, Hongmin Cai, and Shengfeng He. Fine-grained domain adaptive crowd counting via point-derived segmentation. *ArXiv Preprint ArXiv:2108.02980*, 2021.

[31] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 616–625, 2018.

[32] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021.

[33] 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.

[34] 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.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[36] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. Weakly supervised scene parsing with point-based distance metric learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 8843–8850, 2019.

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[38] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7263–7271, 2017.

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015.

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

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[43] 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.

[44] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9446–9454, 2018.

[45] 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.

[46] 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.

[47] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Li Liu, Zaiping Lin, and Shilin Zhou. Local motion and contrast priors driven deep network for infrared small target super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTAR)*, 15:5480–5495, 2022.

[48] Mohsen Zand, Haleh Damirchi, Andrew Farley, Mahdiyar Molahasani, Michael Greenspan, and Ali Etemad. Multiscale crowd counting and localization by multitask point supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1820–1824. IEEE, 2022.

[49] Ke Zhang, Shuyan Ni, Dashuang Yan, and Aidi Zhang. Review of dim small target detection algorithms in single-frame infrared images. In *IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, volume 4, pages 2115–2120. IEEE, 2021.

[50] 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.

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

[52] Mingjin Zhang, Haichen Bai, Jing Zhang, Rui Zhang, Chaoyue Wang, Jie Guo, and Xinbo Gao. Rkformer: Runge-kutta transformer with random-connection attention for

infrared small target detection. In *30th ACM International Conference on Multimedia (ACM MM)*, pages 1730–1738, 2022.

[53] Mingjin Zhang, Ke Yue, Jing Zhang, Yunsong Li, and Xinbo Gao. Exploring feature compensation and cross-level correlation for infrared small target detection. In *30th ACM International Conference on Multimedia (ACM MM)*, pages 1857–1865, 2022.

[54] 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.

[55] Shiyin Zhang, Jun Hao Liew, Yunchao Wei, Shikui Wei, and Yao Zhao. Interactive object segmentation with inside-outside guidance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12234–12244, 2020.

[56] Shilong Zhang, Zhuoran Yu, Liyang Liu, Xinjiang Wang, Aojun Zhou, and Kai Chen. Group r-cnn for weakly semi-supervised object detection with points. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9417–9426, 2022.

[57] Mingjing Zhao, Wei Li, Lu Li, Jin Hu, Pengge Ma, and Ran Tao. Single-frame infrared small-target detection: A survey. *IEEE Geoscience and Remote Sensing Magazine (GRSM)*, 2022.

[58] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[59] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2020.