This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

From Easy to Hard: Progressive Active Learning Framework for Infrared Small Target Detection with Single Point Supervision

Chuang Yu[1,2,3], Jinmiao Zhao[1,2,3], Yunpeng Liu[2†], Sicheng Zhao[4], Yimian Dai[5], Xiangyu Yue[6,7†]

[1]Key Laboratory of Opto-Electronic Information Processing, Chinese Academy of Sciences
[2]Shenyang Institute of Automation, Chinese Academy of Sciences [3]University of Chinese Academy of Sciences
[4]Tsinghua University [5]Nankai University [6]MMLab, The Chinese University of Hong Kong [7]CPII under InnoHK

Abstract

*Recently, single-frame infrared small target (SIRST) detection with single point supervision has drawn wide-spread attention. However, the latest label evolution with single point supervision (LESPS) framework suffers from instability, excessive label evolution, and difficulty in exerting embedded network performance. Inspired by organisms gradually adapting to their environment and continuously accumulating knowledge, we construct an innovative **P**rogressive **A**ctive **L**earning (**PAL**) framework, which drives the existing SIRST detection networks progressively and actively recognizes and learns harder samples. Specifically, to avoid the early low-performance model leading to the wrong selection of hard samples, we propose a model pre-start concept, which focuses on automatically selecting a portion of easy samples and helping the model have basic task-specific learning capabilities. Meanwhile, we propose a refined dual-update strategy, which can promote reasonable learning of harder samples and continuous refinement of pseudo-labels. In addition, to alleviate the risk of excessive label evolution, a decay factor is reasonably introduced, which helps to achieve a dynamic balance between the expansion and contraction of target annotations. Extensive experiments show that existing SIRST detection networks equipped with our PAL framework have achieved state-of-the-art (SOTA) results on multiple public datasets. Furthermore, our PAL framework can build an efficient and stable bridge between full supervision and single point supervision tasks. Our code is available at* https://github.com/YuChuang1205/PAL

1. Introduction

Single-frame infrared small target (SIRST) detection is one of the key technologies in infrared search and tracking systems [12, 42, 66], which has been widely used in traffic monitoring [31, 74], early warning [33, 63], and maritime rescue [73, 78]. However, this task faces many challenges, such as complex backgrounds, scarcity of intrinsic features and insufficient annotated data [31, 66–68, 74, 79]. Existing research mainly focuses on full supervision, but the cost of pixel-level dense annotation is expensive. Therefore, achieving high-performance SIRST detection with single point supervision is both challenging and significant.

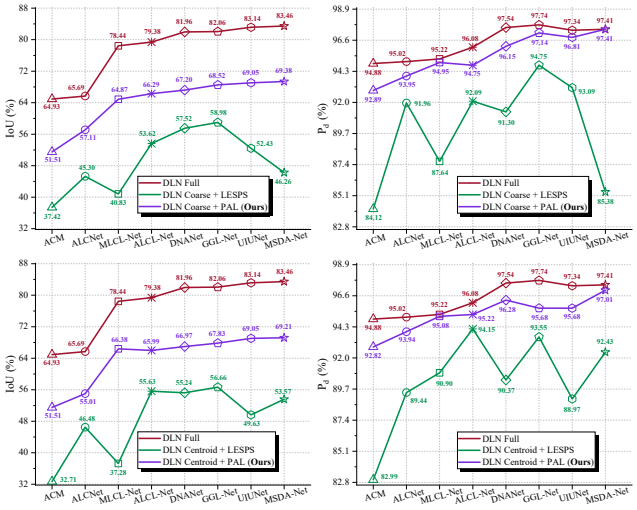

Figure 1. Comparison of different methods on the SIRST3 dataset. *DLN Full*, *DLN Coarse*, and *DLN Centroid* denote DLN-based methods with full supervision, coarse and centroid point supervision. The curve trend of DLNs equipped with the PAL framework is **basically consistent** with that of full supervision, whereas DLNs with the LESPS framework [66] is not. In addition, compared with the LESPS, using our PAL framework can improve by **8.53%-29.1%** and **1.07%-12.03%** in the IoU and P_d.

Early SIRST detection focus on model-driven methods [2–4, 8, 10, 13, 16, 20, 24, 45, 47, 53, 60, 72], which are prone to instability due to static background assumptions and hyperparameter sensitivity. Compared with model-driven methods, deep learning networks (DLNs) are data-driven and can automatically learn the nonlinear relationship between input images and labels [30]. Different from general targets, infrared small targets are small in size and lack of intrinsic features. It is difficult to achieve satisfactory detection results [66] by directly using existing general target detection methods such as the YOLO series [49, 50, 56] and SSD [36]. Therefore, existing meth-

† Corresponding authors.

ods focus on building proprietary SIRST detection networks by leveraging domain knowledge [15, 67, 68, 74, 79, 81]. However, existing SIRST detection methods focus on full supervision, which requires expensive pixel-level dense annotations. To reduce the annotation cost, Ying *et al*. recently propose a LESPS framework [66] that combines the existing excellent SIRST detection network with continuous pseudo-label evolution to achieve SIRST detection with single point supervision. However, from the Sec. A of Supplementary Materials and Fig. 1, the LESPS framework has problems such as instability, excessive label evolution, and difficulty in exerting embedded network performance.

To address these problems, we attempt to build a new, efficient and stable SIRST detection framework. Inspired by organisms gradually adapting to their environment and continuously accumulating knowledge, we came up with an idea: ***The learning of network models should also be from easy to hard?*** *We believe that an excellent learning process should be from easy to hard and take into account the learning ability of the current learner (model) rather than directly treating all tasks (samples) equally.* On this basis, we construct an innovative progressive active learning framework, which drives the existing SIRST detection networks progressively and actively recognizes and learns harder samples. Our PAL can be classified as automatic curriculum learning (CL) [57]. However, unlike previous automatic CL that focuses on full supervision or other weakly supervised tasks [27, 28, 52, 58, 84], this paper is the first attempt to explore its derivation to single point supervision tasks and construct an effective learning framework for SIRST detection with single point supervision. From Fig. 1, our PAL achieves significant performance improvements and builds an efficient and stable bridge between full supervision and single point supervision tasks.

Specifically, **first**, to enable the network model to have basic task-specific learning capabilities in the initial training phase, we propose a model pre-start concept. Different from the existing methods that directly use all samples with point labels for training, it focuses on automatically selecting a portion of easy samples and put them into the training pool for initial training. For this task, an easy-sample pseudo-label generation (EPG) strategy that does not require manual annotation is constructed to achieve the initial selection of easy samples and the generation of their corresponding pseudo-labels. **Second**, considering that the model after model pre-start phase has the ability to recognize some hard samples and the pseudo-labels of samples in the training pool need to be refined, we construct a fine dual-update strategy that includes coarse outer updates (COU) and fine inner updates (FIN). Unlike the existing methods that perform only label evolution, the fine dual-update strategy not only dynamically selects relatively easy samples from hard samples into the training pool based on the capability of the current model, but also performs fine label updates on samples that have entered the training pool. **Finally**, we observe an interesting phenomenon: the existing framework has the problem that the evolution area does not shrink after over-expansion. To solve this problem, a decay factor is rationally introduced to achieve a dynamic balance between the expansion and contraction of target annotations. The contributions can be summarized as follows:

- We construct an innovative PAL framework that can transfer existing SIRST detection networks to single point supervision tasks and drive them to progressively and actively learn harder samples, thereby achieving continuous performance improvements.
- A model pre-start concept is proposed, which focuses on automatically selecting a portion of easy samples and can help models have basic task-specific learning capabilities. For this task, we construct an EPG strategy to achieve the initial selection of easy samples and the generation of corresponding pseudo-labels.
- A fine dual-update strategy is proposed, which can promote reasonable learning of harder samples and continuous refinement of pseudo-labels.
- To alleviate the risk of excessive label evolution, a decay factor is reasonably introduced, which helps achieve a dynamic balance between the expansion and contraction of target annotations.

2. Related Work

SIRST detection can be divided into model-driven non-deep learning methods [2–4, 8, 10, 13, 16, 20, 24, 45, 47, 53, 60, 72] and data-driven deep learning methods [14, 15, 31, 62, 67, 68, 74, 79, 81]. Early research focus on model-driven methods, including background suppression-based methods [2, 4, 10, 47, 53], human visual system-based methods [3, 8, 16, 45, 60], and image data structure-based methods [13, 20, 24, 72]. However, these methods require careful hyperparameter tuning. Subsequently, data-driven methods gradually replaced model-driven methods. Its research focuses on incorporating domain knowledge [15, 67, 68, 74, 79, 81] and optimizing network architectures [14, 31, 62]. These methods perform well in SIRST detection with full supervision. However, pixel-wise dense annotations are expensive, which motivates us to exploit the more cost-effective point-level weak supervision [5].

Point-based supervision is of great significance and has been widely studied in object detection [9, 43, 76], object counting [1, 29, 71], and semantic segmentation [5, 11, 17, 19, 39, 46, 59, 64, 75, 77]. We focus mainly on image segmentation in this paper. For weakly-supervised semantic segmentation with point labels, existing methods focus on general object segmentation [5, 11, 75] and panoptic segmentation [19, 39, 46], where targets exhibit rich colors, fine textures and often require multiple annotated points.

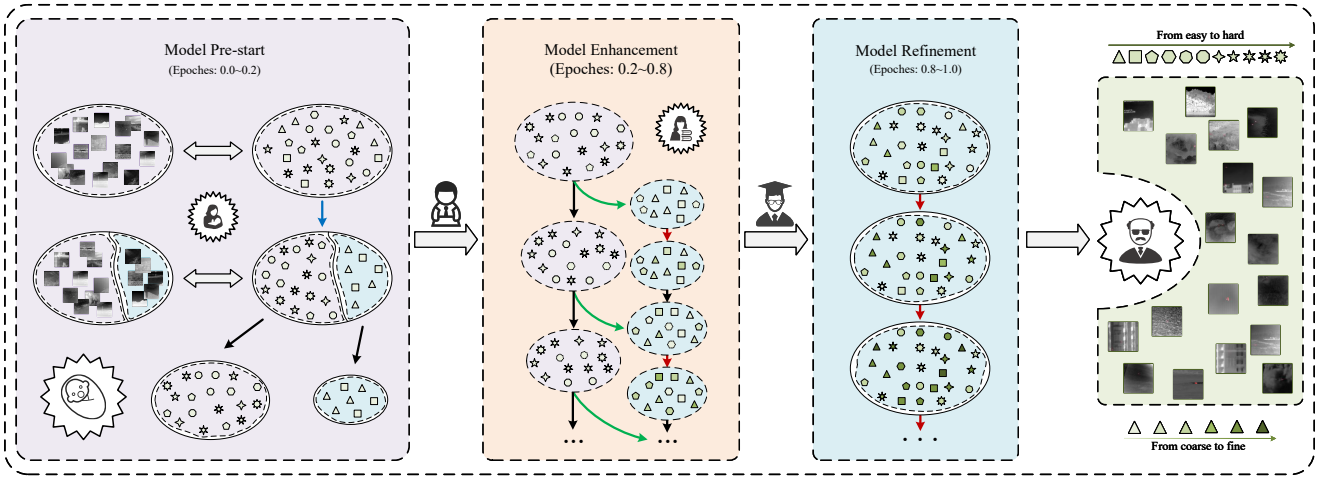

Figure 2. The PAL framework. The purple ellipse area is the preparation pool, which contains samples that cannot be trained temporarily. The blue ellipse area is the training pool, which contains samples that can be trained. Each shape refers to a different level of difficulty, and the color depth refers to the refinement of the corresponding pseudo-label. The *blue* arrow denotes the EPG strategy. The *green* arrows and *red* arrows denote the *coarse outer updates* and *fine inner updates*. 0.0, 0.2, 0.8 and 1.0 denote the division of the total training epochs.

Given the target characteristics and the single-point label of this task, the above method is difficult to apply directly. Currently, there is little research on this task, which can be mainly divided into static pseudo-label generation methods [26, 32, 70] and dynamic pseudo-label evolution methods [65, 66]. The former often rely on additional prior or auxiliary information, while the latter (such as the LESPS framework) is more adaptable and has received wide-spread attention. However, existing dynamic methods suffer from instability, excessive label evolution, and difficulty in exerting embedded network performance. Therefore, we attempt to construct a new, efficient and stable framework.

Curriculum Learning (CL) is a training paradigm inspired by the structured learning progression in human curricula, where models are trained from easier to more difficult data [6]. It consists of two components: a Difficulty Measurer and a Training Scheduler [57], and is categorized into predefined CL [18, 22, 25, 35, 37, 44, 48, 54, 55] and automatic CL [21, 23, 27, 28, 40, 51, 52, 58, 61, 83–85]. Predefined CL follows a manually designed training sequence, while automatic CL dynamically adjusts learning based on model feedback, offering greater flexibility [27]. Unlike existing automatic CL methods that focus on full or other weakly supervision, this work is the first to explore its derivation to single point supervision, specifically for SIRST detection. Meanwhile, conventional automatic CL suffers from early-stage misjudgments in sample difficulty [57]. The negative impact on our task will become more significant. Additionally, the single-point label format necessitates a rethinking of Difficulty Measurer and Training Scheduler design. Therefore, we attempt to address these challenges and construct an innovative framework for infrared small target detection with single point supervision.

3. Methods

3.1. Overview

For fine-grained SIRST detection with single point supervision, we propose an end-to-end PAL framework adaptable to all existing SIRST detection networks. From Fig. 2 and Fig. 3, the training process consists of three phases: model pre-start, model enhancement, and model refinement.

The model pre-start phase aims to give model the initial task-specific learning ability. Like a newborn, an untrained model has potential but requires simple knowledge first. This is why existing methods that directly use all samples with point labels for training have problems such as unstable training and poor accuracy. Additionally, low-performance models may misidentify hard samples, thereby compromising final performance. Inspired by early maternal guidance, we construct an EPG strategy to automatically generate pseudo-labels for all samples, and divide them into easy samples and hard samples according to the quality of pseudo-labels. Only easy samples enter the training pool for training with a fixed epochs (0.2 × total epochs). For details, please see Sec. 3.2.

The model enhancement phase aims to make the model progressively and actively recognize and learn harder samples to ensure the accumulation of new knowledge, and to update the pseudo-labels of samples in the training pool. Specifically, The model evaluates hard samples in the preparation pool, transferring those with reliable pseudo-labels to the training pool through periodic coarse outer updates. Simultaneously, all pseudo-labels in the training pool are iteratively refined via fine inner updates. It is like students adjusting their misperceptions of acquired information as their abilities grow. For details, please see Sec. 3.3.

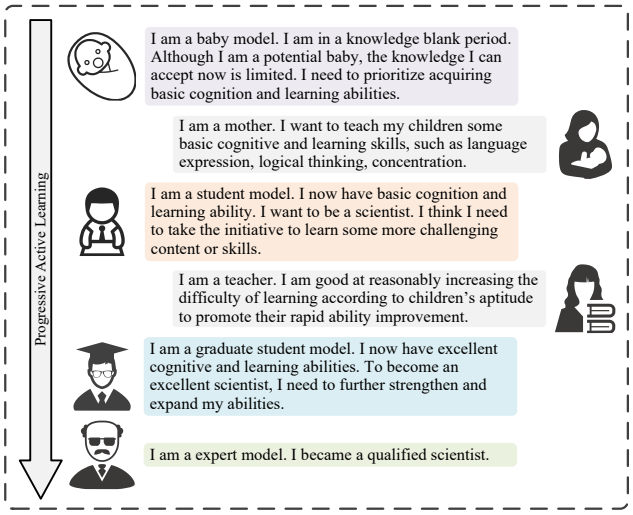

Figure 3. Schematic representation of progressive active learning.

In addition, inspired by teacher-guided learning, we introduce additional guidance to dynamically adjust the difficulty threshold based on the model's capabilities, accelerating learning and ensuring that the to-be-learned content is learned in a timely manner.

The model refinement phase aims to allow the model to fully learn hard samples and further refine pseudo-labels. Specifically, the hard samples are periodically and gradually added to the training pool in the model enhancement phase. For hard samples that enter later, it is difficult to be fully learned in a few periods. At the same time, as the model continues to learn hard samples, its stronger detection ability can help further refine the pseudo-labels. Therefore, we perform the fine inner updates periodically. This phase resembles graduates refining their expertise through repeated review and practice to achieve proficiency.

3.2. Model pre-start

Compared to directly training with point labels, selecting easy samples that can generate high-quality pseudo-labels for initial training can help the network establish basic SIRST detection capabilities. Therefore, different from existing methods [66, 82], we propose a model pre-start concept. Considering the detection scene, imaging characteristics and target properties, infrared small targets are characterized by small size, high brightness and a lack of strong semantic contextual relationships with the surrounding environment [15, 31, 68, 81]. Some target areas can be detected by model-driven methods. However, these methods often lack robustness and are prone to false detections [81]. Therefore, we propose an EPG strategy to achieve the selection of easy samples, as shown in Fig. 4.

To reduce background interference, we process local image patches centered on the given point label instead of the entire image. To accurately detect the target area, we first use the Gaussian filtering to appropriately suppress the noise. Second, the Canny operator [7] is used to extract the target's contour. Finally, a morphological closing operation mitigates contour fragmentation and fills the enclosed area to obtain the segmentation result.

To select easy samples and refine pseudo-labels, we make full use of the point labels and the segmentation results. **First**, the given point label is used to assess connected areas in the segmentation result. Areas containing the label point and below a set threshold are identified as true targets, while others are removed as false detections. **Second**, we calculate the target-level recall rate, which is the proportion of real targets that are correctly detected. If it is greater than and equal to the set threshold (0.8), it is regarded as an easy sample. Otherwise, it is a hard sample. **Third**, each segmented image patch of the easy sample is added to a pure black background with the original image size to generate a pseudo-label. **Finally**, point labels are added to compensate for missed targets, thereby refining the pseudo-labels.

The proposed EPG strategy selects easy samples with high-quality pseudo-labels for initial training, enabling the model to develop basic task-specific capabilities. This also prevents hard samples from introducing excessive noise or misleading information during the early training phase.

3.3. Fine dual-update strategy

The quality of pseudo-labels directly impacts the final model's performance. Unlike existing frameworks that directly iteratively evolve all point labels [66, 82], we propose a fine dual-update strategy, consisting of a *Coarse outer updates* and a *Fine inner updates*.

1) Coarse outer updates (COU): After the model pre-start phase, the model has basic SIRST detection capabilities. Based on it, we assess the remaining hard samples in the preparation pool to select "easy samples" whose generated pseudo-labels meet the selection rules into the training pool. The COU can be divided into three steps: determining recognizable samples, eliminating false targets, and supplementing missed targets.

First, we need to assess hard samples in the preparation pool that the current model can recognize:

$$S = \begin{cases} I \in Easy\,Sample & \text{if } R_m \le T_m \,\&\, R_f \le T_f \\ I \in Hard\,Sample & \text{otherwise} \end{cases} \quad (1)$$

where R_m and R_f denote the target-level missed detection rate (the ratio of missed real targets) and false detection rate (the ratio of incorrectly detected targets to real targets). T_m and T_f denote the corresponding thresholds. When there is an intersection between the true point label and the predicted target area, the target is detected correctly.

Second, the selected easy samples undergo false target elimination. Specifically, predicted target areas that do not intersect with any true point label are directly removed. The

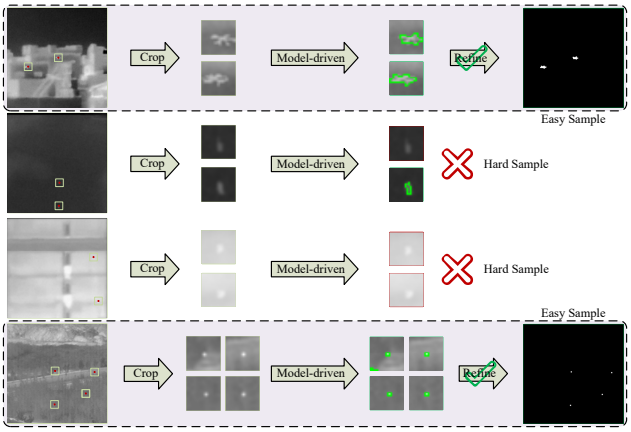

Figure 4. The EPG strategy. The *green* boxes denote correct detections. The *red* boxes denote error detections.

formula expression is as follows:

$$\hat{P}_b = P_b \backslash \{A_f \in P_b | Instersection(A_f, L_{true}) = \phi\} \quad (2)$$

where P_b and $\hat{P}_b$ denote the detection results before and after false target elimination. A_f denotes the predicted target area. $\backslash$ denotes the difference of sets.

Finally, the ground-truth point labels are incorporated to mitigate missed targets and ensure effective learning.

Considering that the model will have stronger cognitive and learning abilities after learning hard samples, we periodically perform COU during the model enhancement phase. This drives the model to progressively and actively recognize and learn more harder samples without catastrophically forgetting easy samples.

2) Fine inner updates (FIU): Whether it is the easy samples that enter the training pool during the model pre-start phase or the hard samples that continuously enter the training pool during the model enhancement phase, the pseudo-labels they initially generate need to be refined. To solve this problem, we construct a FIU. It is inspired by [66] and effectively solves the potential problem of excessive label evolution as shown in the Supplementary Materials. The FIU is divided into three steps: candidate area extraction, false area elimination and pseudo-label update.

For candidate area extraction, we first obtain the binary label for the n_{th} pseudo-label iteration. Second, the centroid of each target connected area in the pseudo-label is extracted and the corresponding $d \times d$ local area in the pseudo-label and prediction result is cropped. Finally, to highlight the target area and reduce accumulated errors, we use adaptive thresholding to extract the local candidate areas from the prediction results:

$$C_n^i = (P_n^i > T_{adapt}) \quad (3)$$

$$T_{adapt} = max(P_n^i)(T_b + k(1 - T_b)L_n^i/(hwr)) \quad (4)$$

where C_n^i denotes candidate areas. P_n^i and L_n^i denote the prediction result and pseudo-label of the image patch. h and w denote the height and width of the image. T_b and k are both set to 0.5 [66]. r is set to 0.15% [14, 31].

For false area elimination, similar to Eq.(2), we use the centroid of the target area in the pseudo-label to check for intersections with the candidate areas. Candidate areas without intersections are eliminated.

For pseudo-label update, we first restore the positions of all refined candidate areas and use the maximum method to combine them into complete candidate areas. Then, the pseudo-label is updated:

$$N_n = \max(\{N_n, \hat{C}_n^i\}) \quad (5)$$

$$L_{n+1} = \lambda L_n \odot (1 - N_n) + \frac{L_n + P_n}{2} \odot N_n \quad (6)$$

where λ is the decay factor. N_n is initialized to a zero matrix with the same size as the label. $\hat{C}_n^i$ denotes the refined candidate areas of the image patch. P_n and L_n denote the prediction result and pseudo-label.

With the improvement of model capabilities and various constraints, the evolution of the target annotation area achieves a dynamic balance between expansion and contraction and generates more refined pseudo-labels.

3.4. Loss Function

For the SIRST detection task, the lack of intrinsic features makes it difficult to accurately locate the target area [15, 68, 81]. Therefore, we introduce an edge-enhanced difficulty-mining (EEDM) loss [80] to guide network optimization. This helps alleviate the imbalance problem of positive and negative pixels, and directs the model's focus to edges and difficult pixels during training. Specifically, EEDM loss emphasizes pixel-level fine-grained detection, which consists of two parts: edge pixel enhancement and difficult pixel mining. On the one hand, the edge pixel enhancement increases the focus on edge details by weighting the loss of edge pixels, thereby enhancing the model's learning of important features. On the other hand, the difficult pixel mining helps the model better focus on difficult-to-detect target areas by filtering pixels with larger loss, thereby promoting the model's learning in complex areas and preventing small targets from being submerged by the background. More details can be found in the Supplementary Materials.

4. Experiments

4.1. Datasets

Four public SIRST detection datasets are used in the experiments, namely SIRST3 [66], NUAA-SIRST [14], NUDT-SIRST [31], and IRSTD-1K [74]. They contain 2755, 427, 1327 and 1001 samples, respectively. We follow the partitioning in [31, 66] for NUAA-SIRST and NUDT-SIRST and follow the partitioning in [66, 74] for IRSTD-1K. The stability under limited samples can be evaluated on these

Table 1. IoU (%), $nIoU$ (%), P_d (%) and F_a (10^{-6}) values of different methods achieved on the SIRST3 dataset. *NUAA-SIRST-Test*, *NUDT-SIRST-Test* and *IRSTD-1K-Test* denote the decompositions of *SIRST3-Test* to verify the robustness of the model.

Scheme	Description	SIRST3-Test				NUAA-SIRST-Test				NUDT-SIRST-Test				IRSTD-1K-Test			
		IoU	$nIoU$	P_d	F_a	IoU	$nIoU$	P_d	F_a	IoU	$nIoU$	P_d	F_a	IoU	$nIoU$	P_d	F_a
FKRW [47]	Filtering	14.46	23.67	60.66	22.50	22.84	29.61	77.19	51.28	14.70	24.11	58.10	106.90	9.25	15.89	54.21	22.32
MPCM [60]	Local Contrast	18.54	26.80	71.96	123.76	42.92	46.48	88.59	91.08	12.55	20.46	64.34	550.90	21.36	26.78	81.48	191.34
NIPPS [13]	Low Rank	18.82	28.59	70.56	15.17	26.82	37.58	80.23	39.42	18.66	26.16	68.89	42.10	14.93	27.07	67.34	38.85
ACM [14]	DLN Full	64.93	64.89	94.88	20.97	67.61	67.57	92.02	10.36	65.77	65.65	95.87	15.95	61.92	59.55	94.28	28.07
	DLN Coarse + LESPS	37.42	35.20	84.12	50.13	43.66	41.51	84.41	37.94	37.22	33.01	83.91	41.09	34.09	35.73	84.51	**60.98**
	DLN Coarse + PAL (Ours)	**51.51**	**54.07**	**92.89**	**39.18**	**58.66**	**60.09**	**87.07**	**24.49**	**54.47**	**54.42**	**94.50**	**15.21**	**42.75**	**46.50**	**92.93**	63.05
ALCNet [15]	DLN Full	65.69	66.68	95.02	34.60	70.33	69.49	94.68	13.72	66.64	68.32	95.56	25.65	61.60	58.28	93.60	47.77
	DLN Coarse + LESPS	45.30	41.81	91.96	48.23	50.15	47.72	90.87	30.80	42.00	39.02	92.38	41.25	**50.26**	44.71	91.58	58.81
	DLN Coarse + PAL (Ours)	**57.11**	**60.22**	**93.95**	**37.20**	**65.54**	**65.79**	**93.16**	**22.09**	**58.60**	**60.68**	**94.50**	**34.88**	49.74	**52.77**	**92.93**	**43.33**
MLCL-Net [67]	DLN Full	78.44	82.01	95.22	17.99	71.28	74.38	91.25	34.57	89.36	90.11	97.14	13.44	63.25	63.38	92.59	17.16
	DLN Coarse + LESPS	40.83	38.34	87.64	45.21	47.73	44.91	91.25	34.44	38.81	35.14	86.46	37.71	41.27	41.90	88.21	54.37
	DLN Coarse + PAL (Ours)	**64.87**	**69.40**	**94.95**	**24.43**	**66.63**	**70.04**	**93.54**	**24.08**	**71.20**	**72.65**	**97.14**	**21.28**	**51.11**	**57.96**	**89.23**	**25.03**
ALCL-Net [68]	DLN Full	79.38	81.09	96.08	18.64	72.79	75.36	93.54	20.92	87.68	88.21	97.57	13.03	68.36	63.66	93.60	22.64
	DLN Coarse + LESPS	53.62	52.15	92.09	48.61	62.70	60.92	91.63	**20.44**	51.95	49.52	92.38	60.58	52.28	51.49	**91.58**	46.52
	DLN Coarse + PAL (Ours)	**66.29**	**68.18**	**94.75**	**18.79**	**66.56**	**69.09**	**91.63**	28.88	**68.95**	**70.58**	**96.83**	**16.20**	**60.75**	**59.29**	90.91	**18.14**
DNANet [31]	DLN Full	81.96	85.90	97.54	9.11	78.06	80.50	97.72	15.71	92.70	93.37	99.05	5.77	64.80	66.94	92.59	10.04
	DLN Coarse + LESPS	57.52	55.09	91.30	19.04	61.80	61.32	92.40	**13.03**	57.65	53.41	93.44	18.66	54.56	54.02	83.50	21.01
	DLN Coarse + PAL (Ours)	**67.20**	**70.20**	**96.15**	**10.86**	**70.35**	**71.99**	**95.82**	24.28	**70.72**	**72.61**	**98.62**	**7.47**	**57.33**	**60.33**	**88.55**	**9.94**
GGL-Net [79]	DLN Full	82.06	85.34	97.74	12.76	78.40	79.99	96.96	20.72	92.07	92.34	99.26	4.87	66.68	67.88	93.60	17.08
	DLN Coarse + LESPS	58.98	57.16	94.75	20.47	66.26	64.80	95.06	20.72	57.81	54.69	95.45	13.56	57.32	57.18	92.26	26.11
	DLN Coarse + PAL (Ours)	**68.52**	**71.69**	**97.14**	**16.69**	**71.21**	**74.02**	**95.44**	**17.42**	**72.02**	**73.92**	**98.52**	**6.50**	**59.90**	**61.86**	**94.61**	**24.90**
UIUNet [62]	DLN Full	83.14	85.29	97.34	15.44	76.64	78.43	95.82	12.69	92.94	93.29	98.41	4.11	70.04	66.19	95.29	25.55
	DLN Coarse + LESPS	52.43	52.86	93.09	16.92	58.55	60.66	90.11	20.37	52.03	50.20	94.50	12.18	49.71	53.35	91.25	19.87
	DLN Coarse + PAL (Ours)	**69.05**	**71.53**	**96.81**	**15.45**	**66.05**	**69.21**	**93.54**	**10.98**	**74.21**	**75.61**	**98.52**	**9.15**	**60.63**	**60.50**	**94.28**	**21.90**
MSDA-Net [81]	DLN Full	83.46	85.97	97.41	17.15	74.81	78.61	95.06	30.94	93.62	94.03	99.26	9.67	70.98	67.16	93.60	19.51
	DLN Coarse + LESPS	46.26	45.73	85.38	36.16	54.91	54.69	92.02	30.25	45.18	42.48	83.92	35.32	43.16	46.92	84.18	38.49
	DLN Coarse + PAL (Ours)	**69.38**	**71.55**	**97.41**	**16.34**	**71.67**	**73.38**	**96.96**	**11.59**	**72.62**	**73.86**	**98.41**	**6.18**	**61.62**	**61.94**	**94.61**	**26.04**

three datasets. SIRST3 is a multi-sample, multi-scene, and multi-target dataset that combines NUAA-SIRST, NUDT-SIRST, and IRSTD-1K. It can more effectively evaluate the comprehensive detection capabilities. Consistent with [66], we process the labels of the training set into two types: *Coarse* point labels and *Centroid* point labels.

4.2. Implementation Details

Experimental Settings: We use the AdamW [38] optimizer with initial learning rate of $1e^{-3}$. The batch size is set to 16. The total epochs is 400, which is divided into 3 parts: *0-0.2*, *0-0.8* and *0.8-1*. The cropped image patch in EPG is 20×20 pixels. We interestingly found that DLN-based models exhibit minimal false detections but significant missed detections for this task. Since our strategy eliminates false positive areas, we set T_f to a larger value to reduce its constraint. To ensure all samples enter the training pool during model enhancement phase, the initial threshold of T_m is set to 0.2 and increases linearly to 1 as the epoch increases. The epoch period of COU and FIU is set to 5. During training, all images are normalized and randomly cropped into 256×256 pixel patches as network inputs.

Evaluation Metrics: We used two pixel-level evaluation metrics of intersection over union (IoU) and normalized IoU (nIoU) [14] and two target-level evaluation metrics of detection probability (P_d) and false alarm rate (F_a) [31]. In addition, since the magnitude of F_a is 10^{-6}, its value is small and prone to fluctuations. Therefore, we also evaluate the model using a threshold-based approach, and the model is judged invalid when $F_a > 1e^{-4}$ [69, 82]. Invalid results are highlighted in red.

4.3. Comparison with SOTA Methods

We embed multiple excellent SIRST detection networks [14, 15, 31, 62, 67, 68, 79, 81] into our framework. For fair comparison, we retrain all models with the same settings. Meanwhile, we also conduct detailed comparative experiments with various traditional methods [13, 47, 60] and the latest LESPS framework [66]. More experiments such as MCLC [32] can be found in the Supplementary Materials.

Evaluation on the SIRST3 dataset. From Tab. 1, when the networks (UIU-Net, MSDA-Net) are embedded into the LESPS framework for single point supervision tasks, they fail to exert their performance advantages on full supervision. This is because LESPS directly uses all point labels for training and starts label evolution at fixed loss or epoch values, which can not accurately control the optimal timing for label evolution. In contrast, our PAL framework addresses this issue effectively. The performance trend of DLNs equipped with our PAL closely matches that of full supervision. At the same time, compared to LESPS, our PAL improves the IoU by 9.68%-24.04%, the nIoU by 14.53%-31.06%, and the P_d by 1.99%-12.03% on the comprehensive SIRST3-Test. The improvement is obvious. Compared with the fully supervised task on the SIRST3-Test, our PAL framework can reach 79.33%-86.94% on IoU, 81.72%-90.30% on nIoU, and has comparable performance on P_d. Additionally, from the results on the three decomposed test subsets, our PAL has stable performance. Furthermore, from Fig. 5, DLNs equipped with our PAL have target-level detection effects similar to those of full supervision, and they are significantly better than the LESPS.

Table 2. IoU (%), $nIoU$ (%), P_d (%) and F_a (10^{-6}) values of different methods achieved on the separate NUAA-SIRST, NUDT-SIRST, and IRSTD-1k datasets. *(213:214)*, *(663:664)* and *(800:201)* denote the division of training samples and test samples.

Scheme	Description	NUAA-SIRST (213:214)				NUDT-SIRST (663:664)				IRSTD-1K (800:201)			
		IoU	$nIoU$	P_d	F_a	IoU	$nIoU$	P_d	F_a	IoU	$nIoU$	P_d	F_a
ACM [14]	DLN Full	65.67	63.74	90.11	24.01	65.33	65.12	95.87	12.92	60.45	53.70	92.26	46.06
ALCNet [15]	DLN Full	66.41	65.18	91.63	35.26	69.74	70.67	97.46	11.15	62.47	55.25	88.89	36.79
MLCL-Net [67]	DLN Full	74.68	76.50	95.82	28.74	94.03	93.97	98.73	7.721	64.86	63.35	91.25	23.23
	DLN Coarse + LESPS	41.31	41.13	90.87	43.22	53.55	48.70	90.05	21.92	44.64	42.04	88.22	36.42
	DLN Coarse + PAL (Ours)	**68.76**	**70.75**	**93.92**	**39.10**	**74.00**	**73.71**	**95.45**	**9.40**	**58.90**	**58.79**	**90.91**	**24.63**
ALCL-Net [68]	DLN Full	72.84	72.59	95.06	17.01	92.80	93.01	99.05	2.21	65.56	65.03	91.58	9.68
	DLN Coarse + LESPS	37.99	41.91	90.87	35.54	45.89	43.66	90.48	34.45	49.46	45.00	82.49	26.72
	DLN Coarse + PAL (Ours)	**67.49**	**67.67**	**93.54**	**24.01**	**70.62**	**72.56**	**98.31**	**7.93**	**57.98**	**53.36**	**88.22**	**20.61**
DNANet [31]	DLN Full	76.40	78.32	96.20	20.72	95.17	95.19	98.94	2.00	69.06	65.22	91.58	11.56
	DLN Coarse + LESPS	9.29	7.36	19.77	415.72	48.10	48.84	81.80	85.19	50.99	47.61	85.52	25.49
	DLN Coarse + PAL (Ours)	**66.57**	**69.73**	**90.49**	**27.44**	**73.29**	**74.70**	**98.10**	**22.08**	**58.40**	**60.29**	**91.25**	**24.52**
GGL-Net [79]	DLN Full	75.47	75.93	97.34	20.65	94.86	94.93	99.47	1.03	69.09	65.52	92.59	13.68
	DLN Coarse + LESPS	55.17	54.51	89.35	26.27	57.39	56.67	91.75	23.92	50.90	48.12	86.20	**27.73**
	DLN Coarse + PAL (Ours)	**63.46**	**64.31**	**90.11**	**25.73**	**73.10**	**74.83**	**98.20**	**13.12**	**60.72**	**57.05**	**90.24**	40.20
UIUNet [62]	DLN Full	78.02	76.85	96.58	14.68	95.07	95.10	98.73	0.21	70.94	64.32	91.25	10.57
	DLN Coarse + LESPS	7.11	8.02	72.62	550.80	35.89	33.12	78.41	89.05	21.49	18.29	73.74	227.06
	DLN Coarse + PAL (Ours)	**62.25**	**61.68**	**91.25**	**34.99**	**74.89**	**76.11**	**98.62**	**7.10**	**61.70**	**59.99**	**92.57**	**21.60**
MSDA-Net [81]	DLN Full	76.73	77.78	96.20	21.75	95.27	95.18	99.15	1.72	70.98	65.70	93.94	33.95
	DLN Coarse + LESPS	43.39	49.33	88.59	27.37	41.60	38.25	79.79	34.42	43.29	38.16	83.16	54.98
	DLN Coarse + PAL (Ours)	**66.78**	**65.92**	**92.78**	**22.77**	**74.48**	**75.37**	**97.46**	**16.80**	**62.72**	**56.71**	**89.23**	**28.64**

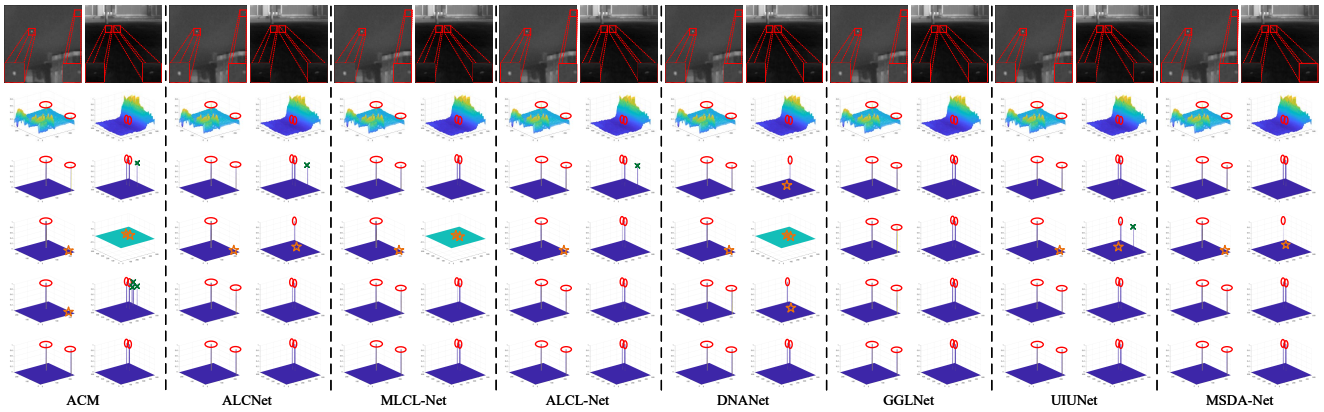

Figure 5. 3D visualization on the SIRST3 dataset. The circles denote correct detections, the crosses denote false detections, and the stars denote missed detections. From top to bottom: *2D image*, *3D image*, *DLN Full*, *DLN Coarse + LESPS*, *DLN Coarse + PAL*, *True label*.

Evaluation on three individual datasets. To further explore the stability of our PAL framework under limited samples, we perform separate training and testing on the three individual datasets. From Tab. 2, The PAL consistently outperforms the LESPS with notably higher stability. These show that our PAL framework can solve the single point supervised SIRST detection task with few training samples, whereas the LESPS framework suffers from a high risk of failure. Specifically, compared with the LESPS, using the PAL framework improves the IoU by 7.41%-57.28%, the nIoU by 8.36%-62.37%, and the P_d by 0.76%-70.72%.

4.4. Ablation Experiment

To fully validate the proposed PAL framework, we perform detailed ablation experiments. More experiments can be found in the Supplementary Materials.

1) Break-Down Ablation. To verify the effectiveness of model pre-start and fine dual-update strategy, Tab. 3 presents detailed ablation experiments. From Scheme 1 and Scheme 2, training directly with point labels will fail, but the network after applying the model pre-start will show significant performance improvement. This verifies the effectiveness of our model pre-start concept and EPG strategy. From Schemes 2-4, whether using COU or FIU alone can significantly improve the performance. In addition, from Schemes 2-5, when the COU and the FIU are used in combination, a significant performance improvement of "$1+1>2$" is achieved. This fully verifies the effectiveness of the proposed fine dual-update strategy.

2) Decay Factor. As presented in Tab. 4, compared with not using the decay factor ($\lambda = 1$), using decay factors can improve the segmentation accuracy. The IoU will increase by 0.04%-1.48% and the nIoU will increase by 1.10%-2.92%. At the same time, using a decay factor within a reasonable range will also lead to an increase in P_d. We finally set λ to 0.97.

Table 3. Break-down ablation experiments on the SIRST3 dataset. *MP* is the model pre-start. *Coarse* denotes coarse point supervision. *Centroid* denotes centroid point supervision.

Scheme	Variants			MSDA-Net Coarse + PAL				MSDA-Net Centroid + PAL			
	MP	COU	FIU	IoU	IoU	P_d	F_a	IoU	$nIoU$	P_d	F_a
Scheme1	-	-	-	1.23	2.60	40.47	**0.26**	2.56	4.94	83.52	**0.61**
Scheme2	✓	✗	✗	56.24	64.20	89.63	31.48	57.25	64.44	91.83	23.83
Scheme3	✓	✓	✗	61.64	67.14	94.02	31.15	60.19	67.17	95.75	18.74
Scheme4	✓	✗	✓	61.66	66.78	93.42	39.47	59.81	65.38	92.89	25.06
Scheme5	✓	✓	✓	**69.38**	**71.55**	**97.41**	16.34	**69.21**	**72.40**	**97.01**	15.70

Table 4. Decay factor investigation on the SIRST3 dataset.

λ	MSDA-Net Coarse + PAL				λ	MSDA-Net Coarse + PAL			
	IoU	$nIoU$	P_d	F_a		IoU	$nIoU$	P_d	F_a
1.00	67.90	69.67	96.68	18.96	0.95	68.96	71.40	97.41	**15.76**
0.99	68.02	70.94	96.74	19.13	0.90	68.92	72.00	96.48	16.35
0.98	68.92	71.20	96.88	19.02	0.85	68.73	**72.59**	96.41	18.66
0.97	**69.38**	71.55	**97.41**	16.34	0.80	67.94	71.30	96.15	20.36
0.96	69.13	70.77	97.28	16.52	0.75	68.11	71.60	95.55	29.45

Table 5. Comparison of different losses on the SIRST3 dataset.

Loss Function	MSDA-Net Coarse + PAL				MSDA-Net Centroid + PAL			
	IoU	$nIoU$	P_d	F_a	IoU	$nIoU$	P_d	F_a
BCE loss	69.26	71.48	97.08	**15.98**	69.19	71.67	96.74	**14.25**
Dice loss	64.38	66.64	94.68	39.83	65.18	67.71	95.61	35.61
Focal loss	68.89	**72.21**	97.01	17.02	69.14	71.85	96.28	16.18
EEDM loss	**69.38**	71.55	**97.41**	16.34	**69.21**	**72.40**	**97.01**	15.70

Table 6. Comparison of the PAL framework with different settings on the SIRST3 dataset. *MP* denotes the proposed model pre-start. *MP-1* and *MP-2* denote the model pre-start with all samples initially enter the training pool by setting the target recall threshold to 0 and directly carry the true point labels for remaining hard samples into the training pool. *S* denotes scheme.

S	Variants					MSDA-Net Coarse + PAL				MSDA-Net Centroid + PAL			
	MP	MP-1	MP-2	COU	FIU	IoU	$nIoU$	P_d	F_a	IoU	$nIoU$	P_d	F_a
S1	✓	✗	✗	✗	✗	56.24	64.20	89.63	31.48	57.25	64.44	91.83	23.83
S2	✗	✓	✗	✗	✗	32.86	45.42	73.62	13.62	32.45	44.18	73.49	13.19
S3	✗	✗	✓	✗	✗	25.72	38.26	61.53	14.14	28.77	41.86	71.56	11.67
S4	✓	✗	✗	✗	✓	61.66	66.78	93.42	39.47	59.81	65.38	92.89	25.06
S5	✗	✓	✗	✗	✓	40.84	51.23	74.82	18.88	41.87	52.96	78.80	14.15
S6	✗	✗	✓	✗	✓	36.67	48.88	71.10	**12.10**	37.13	48.75	75.08	**11.37**
S7	✓	✗	✗	✓	✓	**69.38**	**71.55**	**97.41**	16.34	**69.21**	**72.40**	**97.01**	15.70

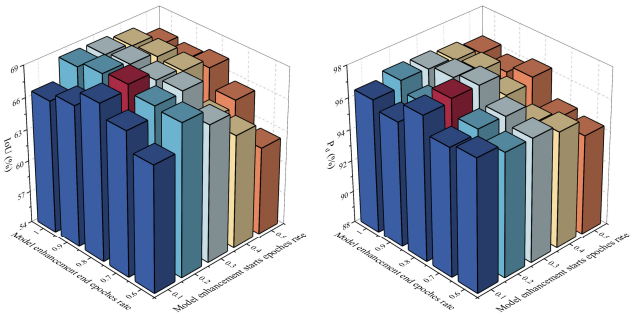

Figure 6. Epoch division investigation on the SIRST3 dataset. *Red* denotes the optimal value.

3) Epoches Division. As presented in Fig. 6, when the model pre-start interval is greater than 0.1 and the model enhancement interval is from 0.1 to 0.4, the performance shows an upward trend. This shows that subdividing all the hard samples multiple times will have a more positive performance improvement. It is like a student model with basic abilities that refines the content to be learned in more detail and then learns from easy to hard. At the same time, when the model refinement phase is eliminated or the model pre-start interval is only 0.1, the performance of the network will degrade. This proves the superiority of the 3-phase design in our PAL framework. In addition, when the model pre-start interval is greater than 0.1 and the model enhancement interval is greater than 0.3, the performance is stable. It demonstrates the robustness of our PAL framework.

4) Loss Function. We compare EEDM loss with the binary cross-entropy loss, Dice loss [41] and focal loss [34]. From Tab. 5, compared with other losses, EEDM loss has the best results. Using EEDM loss can improve the IoU by 0.02%-5.00% and improve the P_d by 0.33%-2.73%.

4.5. Discussion of "From easy to hard"

To further discuss the proposed PAL that focuses on learning from easy samples to hard samples, we conduct detailed experiments. From S1-S6 in Tab. 6, regardless of whether FIU are added or not, the final model generated by all samples entering the training pool in the initial stage is far worse than that of using only easy samples. This further verifies the effectiveness of our model pre-start concept of learning only easy samples in the initial phase. It's like babies should be given pronunciation exercises rather than forced to learn calculus. In addition, from S4 and S7 in Tab. 6, compared to using only easy samples, gradually selecting hard samples into the training pool based on the current capabilities will again achieve significant performance improvements. This further verifies the effectiveness of the idea of learning from easy samples to hard samples. More visualization results can be found in the Supplementary Materials.

5. Conclusion

This paper proposes an innovative PAL framework for the fine-grained SIRST detection with single point supervision, which drives the existing SIRST detection networks progressively and actively recognizes and learns harder samples to achieve significant performance improvements. First, we propose a model pre-start concept, which can help models have basic task-specific learning capabilities in the initial training phase. Second, a refined dual-update strategy is proposed, which can promote reasonable learning of harder samples and continuous refinement of pseudo-labels. Finally, to alleviate the risk of region expansion without shrinkage, a decay factor is reasonably introduced to achieve a dynamic balance between the expansion and shrinkage of target annotations. Extensive comparison results demonstrate the superiority of our PAL framework. We hope that our study can draw attention to the research on weakly supervised SIRST detection.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant No. 62306261), the Shun Hing Institute of Advanced Engineering (SHIAE) Fund (Grant No. 8115074), the LiaoNing Revitalization Talents Program (Grant No. XLYC2201001), and the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

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 *CVPRW*, pages 50–51, 2020. 2

[2] GONZALOR Arce and MICHAELP McLoughlin. Theoretical analysis of the max/median filter. *IEEE transactions on acoustics, speech, and signal processing*, 35(1):60–69, 1987. 1, 2

[3] Xiangzhi Bai and Yanguang Bi. Derivative entropy-based contrast measure for infrared small-target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4): 2452–2466, 2018. 2

[4] Xiangzhi Bai and Fugen Zhou. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognition*, 43(6):2145–2156, 2010. 1, 2

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

[6] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009. 3

[7] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 4

[8] 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*, 52(1):574–581, 2013. 1, 2

[9] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, and Jian Sun. Points as queries: Weakly semi-supervised object detection by points. In *CVPR*, pages 8823–8832, 2021. 2

[10] Zhong Chen, Song Luo, Ting Xie, Jianguo Liu, Guoyou Wang, and Gao Lei. A novel infrared small target detection method based on bemd and local inverse entropy. *Infrared Physics & Technology*, 66:114–124, 2014. 1, 2

[11] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *CVPR*, pages 2617–2626, 2022. 2

[12] Xiaolong Cui, Xingxiu Li, Panlong Wu, Shan He, and Ruohan Zhao. Weakly semi-supervised infrared small target detection guided by point labels. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 1

[13] 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. 1, 2, 6

[14] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 950–959, 2021. 2, 5, 6, 7

[15] 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*, 59(11):9813–9824, 2021. 2, 4, 5, 6, 7

[16] He Deng, Xianping Sun, Maili Liu, Chaohui Ye, and Xin Zhou. Small infrared target detection based on weighted local difference measure. *IEEE Transactions on Geoscience and Remote Sensing*, 54(7):4204–4214, 2016. 1, 2

[17] Anan Du, Tianfei Zhou, Shuchao Pang, Qiang Wu, and Jian Zhang. Pcl: Point contrast and labeling for weakly supervised point cloud semantic segmentation. *IEEE Transactions on Multimedia*, 2024. 2

[18] Rasheed El-Bouri, David Eyre, Peter Watkinson, Tingting Zhu, and David Clifton. Student-teacher curriculum learning via reinforcement learning: Predicting hospital inpatient admission location. In *ICML*, pages 2848–2857. PMLR, 2020. 3

[19] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Pointly-supervised panoptic segmentation. In *ECCV*, pages 319–336. Springer, 2022. 2

[20] 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. 1, 2

[21] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *ICML*, pages 1311–1320. Pmlr, 2017. 3

[22] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *ECCV*, pages 135–150, 2018. 3

[23] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *ICML*, pages 2535–2544. PMLR, 2019. 3

[24] YuJie He, Min Li, JinLi Zhang, and Qi An. Small infrared target detection based on low-rank and sparse representation. *Infrared Physics & Technology*, 68:98–109, 2015. 1, 2

[25] Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A González Ballester, and Gemma Piella. Medical-based deep curriculum learning for improved fracture classification. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 694–702. Springer, 2019. 3

[26] Renke Kou, Chunping Wang, Qiang Fu, Zhanwu Li, Ying Luo, Boyang Li, Wei Li, and Zhenming Peng. Mcgc: A

multiscale chain growth clustering algorithm for generating infrared small target mask under single-point supervision. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12, 2024. 3

[27] M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *NeurIPS*, 23, 2010. 2, 3

[28] M Pawan Kumar, Haithem Turki, Dan Preston, and Daphne Koller. Learning specific-class segmentation from diverse data. In *ICCV*, pages 1800–1807. IEEE, 2011. 2, 3

[29] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *ECCV*, pages 547–562, 2018. 2

[30] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[31] 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*, 32:1745–1758, 2022. 1, 2, 4, 5, 6, 7

[32] Boyang Li, Yingqian Wang, Longguang Wang, Fei Zhang, Ting Liu, Zaiping Lin, Wei An, and Yulan Guo. Monte carlo linear clustering with single-point supervision is enough for infrared small target detection. In *ICCV*, pages 1009–1019, 2023. 3, 6

[33] Yansheng Li and Yongjun Zhang. Robust infrared small target detection using local steering kernel reconstruction. *Pattern Recognition*, 77:113–125, 2018. 1

[34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 8

[35] Cao Liu, Shizhu He, Kang Liu, Jun Zhao, et al. Curriculum learning for natural answer generation. In *IJCAI*, pages 4223–4229, 2018. 3

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 1

[37] Xuebo Liu, Houtim Lai, Derek F Wong, and Lidia S Chao. Norm-based curriculum learning for neural machine translation. *arXiv preprint arXiv:2006.02014*, 2020. 3

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[39] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *CVPR*, pages 616–625, 2018. 2

[40] Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher–student curriculum learning. *IEEE transactions on neural networks and learning systems*, 31 (9):3732–3740, 2019. 3

[41] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 8

[42] Rixiang Ni, Jing Wu, Zhaobing Qiu, Liqiong Chen, Changhai Luo, Feng Huang, Qiujiang Liu, Binxing Wang, Yunxiang Li, and Youli Li. Point-to-point regression: Accurate infrared small target detection with single-point annotation. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 1

[43] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Training object class detectors with click supervision. In *CVPR*, pages 6374–6383, 2017. 2

[44] Gustavo Penha and Claudia Hauff. Curriculum learning strategies for ir: An empirical study on conversation response ranking. In *Proceedings of the Advances in Information Retrieval: 42nd European Conference on IR Research (ECIR)*, pages 699–713. Springer, 2020. 3

[45] Baojun Qi, Tao Wu, and Hangen He. Robust detection of small infrared objects in maritime scenarios using local minimum patterns and spatio-temporal context. *Optical Engineering*, 51(2):027205–027205, 2012. 1, 2

[46] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. Weakly supervised scene parsing with point-based distance metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8843–8850, 2019. 2

[47] 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. 1, 2, 6

[48] Shivesh Ranjan and John HL Hansen. Curriculum learning based approaches for noise robust speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):197–210, 2017. 3

[49] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017. 1

[50] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 1

[51] Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller. Shifting weights: Adapting object detectors from image to video. *NeurIPS*, 25, 2012. 3

[52] Ye Tang, Yu-Bin Yang, and Yang Gao. Self-paced dictionary learning for image classification. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 833–836, 2012. 2, 3

[53] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *ICCV*, pages 839–846. IEEE, 1998. 1, 2

[54] Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. Learning the curriculum with bayesian optimization for task-specific word representation learning. *arXiv preprint arXiv:1605.03852*, 2016. 3

[55] Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Popescu, Dim P Papadopoulos, and Vittorio Ferrari. How hard can it be? estimating the difficulty of visual search in an image. In *CVPR*, pages 2157–2166, 2016. 3

[56] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *NeurIPS*, 37:107984–108011, 2025. 1

[57] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021. 2, 3

[58] Yulin Wang, Yang Yue, Rui Lu, Yizeng Han, Shiji Song, and Gao Huang. Efficienttrain++: Generalized curriculum learning for efficient visual backbone training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 3

[59] Jiacheng Wei, Guosheng Lin, Kim-Hui Yap, Tzu-Yi Hung, and Lihua Xie. Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds. In *CVPR*, pages 4384–4393, 2020. 2

[60] Yantao Wei, Xinge You, and Hong Li. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognition*, 58:216–226, 2016. 1, 2, 6

[61] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *ICML*, pages 5238–5246. PMLR, 2018. 3

[62] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. Uiu-net: U-net in u-net for infrared small object detection. *IEEE Transactions on Image Processing*, 32:364–376, 2022. 2, 6, 7

[63] Hai Xu, Sheng Zhong, Tianxu Zhang, and Xu Zou. Multiscale multilevel residual feature fusion for real-time infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. 1

[64] Yonghao Xu and Pedram Ghamisi. Consistency-regularized region-growing network for semantic segmentation of urban scenes with point-level annotations. *IEEE Transactions on Image Processing*, 31:5038–5051, 2022. 2

[65] Dongning Yang, Haopeng Zhang, Ying Li, and Zhiguo Jiang. Label evolution based on local contrast measure for single-point supervised infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[66] Xinyi Ying, Li Liu, Yingqian Wang, Ruojing Li, Nuo Chen, Zaiping Lin, Weidong Sheng, and Shilin Zhou. Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision. In *CVPR*, pages 15528–15538, 2023. 1, 2, 3, 4, 5, 6

[67] Chuang Yu, Yunpeng Liu, Shuhang Wu, Zhuhua Hu, Xin Xia, Deyan Lan, and Xin Liu. Infrared small target detection based on multiscale local contrast learning networks. *Infrared Physics & Technology*, 123:104107, 2022. 2, 6, 7

[68] Chuang Yu, Yunpeng Liu, Shuhang Wu, Xin Xia, Zhuhua Hu, Deyan Lan, and Xin Liu. Pay attention to local contrast learning networks for infrared small target detection. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 1, 2, 4, 5, 6, 7

[69] Chuang Yu, Yunpeng Liu, Jinmiao Zhao, and Zelin Shi. Lr-net: a lightweight and robust network for infrared small target detection. *arXiv preprint arXiv:2408.02780*, 2024. 6

[70] Shuai Yuan, Hanlin Qin, Renke Kou, Xiang Yan, Zechaun Li, Chenxu Peng, Dongliang Wu, and Huixin Zhou. Beyond full labels: Energy-double-guided single-point prompt for infrared small target label generation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025. 3

[71] Mohsen Zand, Haleh Damirchi, Andrew Farley, Mahdiyar Molahasani, Michael Greenspan, and Ali Etemad. Multiscale crowd counting and localization by multitask point supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1820–1824. IEEE, 2022. 2

[72] Landan Zhang, Lingbing Peng, Tianfang Zhang, Siying Cao, and Zhenming Peng. Infrared small target detection via non-convex rank approximation minimization joint l 2, 1 norm. *Remote Sensing*, 10(11):1821, 2018. 1, 2

[73] Mingjin Zhang, Ke Yue, Jing Zhang, Yunsong Li, and Xinbo Gao. Exploring feature compensation and cross-level correlation for infrared small target detection. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 1857–1865, 2022. 1

[74] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. Isnet: Shape matters for infrared small target detection. In *CVPR*, pages 877–886, 2022. 1, 2, 5

[75] Shiyin Zhang, Jun Hao Liew, Yunchao Wei, Shikui Wei, and Yao Zhao. Interactive object segmentation with inside-outside guidance. In *CVPR*, pages 12234–12244, 2020. 2

[76] 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 *CVPR*, pages 9417–9426, 2022. 2

[77] Yachao Zhang, Yanyun Qu, Yuan Xie, Zonghao Li, Shanshan Zheng, and Cuihua Li. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In *ICCV*, pages 15520–15528, 2021. 2

[78] Enzhong Zhao, Lili Dong, Chenxi Li, and Yingjie Ji. Infrared maritime target detection based on temporal weight and total variation regularization under strong wave interferences. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[79] Jinmiao Zhao, Chuang Yu, Zelin Shi, Yunpeng Liu, and Yingdi Zhang. Gradient-guided learning network for infrared small target detection. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 1, 2, 6, 7

[80] Jinmiao Zhao, Zelin Shi, Chuang Yu, and Yunpeng Liu. Infrared small target detection based on adjustable sensitivity strategy and multi-scale fusion. *arXiv preprint arXiv:2407.20090*, 2024. 5

[81] Jinmiao Zhao, Zelin Shi, Chuang Yu, and Yunpeng Liu. Multi-scale direction-aware network for infrared small target detection. *arXiv preprint arXiv:2406.02037*, 2024. 2, 4, 5, 6, 7

[82] Jinmiao Zhao, Zelin Shi, Chuang Yu, and Yunpeng Liu. Refined infrared small target detection scheme with single-point supervision. *arXiv preprint arXiv:2408.02773*, 2024. 4, 6

[83] Qian Zhao, Deyu Meng, Lu Jiang, Qi Xie, Zongben Xu, and Alexander Hauptmann. Self-paced learning for matrix factorization. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 2015. 3

[84] Peng Zhou, Bicheng Sun, Xinwang Liu, Liang Du, and Xuejun Li. Active clustering ensemble with self-paced learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[85] Sanping Zhou, Jinjun Wang, Deyu Meng, Xiaomeng Xin, Yubing Li, Yihong Gong, and Nanning Zheng. Deep self-paced learning for person re-identification. *Pattern Recognition*, 76:739–751, 2018. 3