MLSA-YOLOV8: an efficient infrared remote sensing small object detection based on label assignment strategy and tiny head

Yi Li, Huajun Wang

***Abstract*—Infrared remote sensing images, captured from significant distances, often exhibit low spatial resolution and contain small abnormal objects. Despite the impressive performance and efficiency of CNN-based infrared image target detection methods, they struggle with detecting such tiny targets. The conventional label assignment strategy is inadequate for this task, and the low spatial resolution results in the loss of small target information, making it difficult for the standard detection head to identify these targets. Consequently, accurate detection of small targets remains a challenge. a new model is put forward for small object detection in remote-sensing infrared images, incorporating an enhanced label assignment strategy. To do this, we first present a matching-based label assignment strategy, aimed at improving the precision of small target label assignments. This strategy facilitates the learning of small target positive samples and enhances localization accuracy. Subsequently, we devise a prediction head structure with high spatial resolution to facilitate the detecting rate of small objects. Ultimately, our model was assessed comprehensively using the SISRT-V2 and the results obtained are promising. This innovative approach paves the way for more precise and effective small object detection in the domain of infrared remote sensing images.**

***Index Terms*—Infrared small object detection, Label Assignment, Prediction Head, Anchor Free**

I. Introduction

INFRARED imaging systems can detect objects at any time of day, making them a significant technological advancement. Their capacity to capture heat sources further enhances their functionality, providing a comprehensive view of the environment. Coupled with their extensive range, these systems offer unparalleled surveillance capabilities. These features make them critical optoelectronic components in surveillance systems[1]. Robust small object detection is important to infrared search and tracking applications. Moreover, it has been an investigatory hot spot [2]. Generic object detectors are categorized into: anchor-based and anchor-free patterns. However, their performance on small-target tasks suffers from unreasonable label assignment of small targets and low resolution after CNN drop sampling [3].

A. Label Assignment Strategy

Location, as one of the main tasks of detection, is expressed as a regression problem in most detection paradigms, where the location branch is designed as an output boundary box offset, and the joint intersection (lou) metric is usually used to evaluate the accuracy[4]. FasterRCNN and RetinaNet select the IoU threshold of pre-defined anchors and ground truth to distinguish positive and negative samples [5][6]. Considering the distance between center points, overlapping parts, along with aspect ratio. yolov4 chose CIoU as the label Assignment Tactic can fulfill greater convergence velocity and precision on the BBox regression [7]. FCOS chooses anchors located in the central region of ground truth as its positive sample[8]. The IoU metric, however, exhibits a significant bias towards larger objects and demonstrates high sensitivity to small infrared targets owing to their diminutive size . As shown in the1, small deviations (6 pixels along diagonal orientation) in small object prediction box caused a significant drop in loU (from 72% to 38%) compared to normal objects (96% and 92%). In other words, Compared with regular objects IoU is vulnerable to light offset between two boundary boxes, and this ambiguous anchors assignment can affect the learning of regression branches[9]. We call this problem Low tolerance for bounding box perturbation [2]. Given the above problems, Xu put forward dot distance, they think very small objects could be regarded as points and the importance of height and breadth is far lower than that of the locations of center points. Hence, DotD centers uniquely on the positional correlation between center points, and the correlation becomes more appropriate for tiny targets of absolute size below 16 pixels[9]. However, the dot method needs to calculate each pixel point. When the target is too large, the computational complexity is high. To tackle the performance difference in infrared tiny object detection, a new label assignment plan is brought forward, which is known as the Multi-scale label assignment strategy. We continue to utilize CIOU as the label allocation index for low-resolution large maps and reference DotD for high-resolution feature maps in order to further enhance the detection rate of minimal targets.

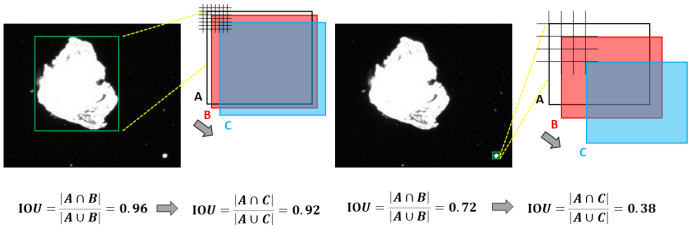

Fig. 1: The sensitivity analysis of IOU on small and normal scale objects. Each grid denotes a pixel. box A denotes the ground truth bbox, box B, C denote the predicted bbox with 1 pixel and 3 pixels right-down deviation respectively.

B. Tiny Head

In the realm of computer vision applications, object detection aims to address the inquiry of "what targets and where are positioned." In the age of deep learning (DL), a predominant paradigm among modern object detectors involves a backbone for feature extraction and a head dedicated to localization and classification tasks. Under the influence of computational cost, former lightweight detectors invariably employ characteristic maps with low detection resolutions (38 X 38 in SSDLite, 19 × 19 in Pelee, 20 × 20 in ThunderNet). Nonetheless, tiny characteristic maps showing low spatial resolution, can not offer spatially matched characteristics for targets in arbitrary locations, particularly for tiny targets[10] Such information loss hardly affects large or medium-sized targets. Unfortunately, this is fatal for small objects, as the detector heads struggle to give accurate predictions over highly structured representations, in which the weak signal from small objects is all but eliminated[2]. Furthermore, the SISRT-V2 dataset is explored, and it is discovered that it encompasses numerous highly tiny cases, such as those with bounding boxes of 3 X 3 5 X 5, which might include no characteristic points through a pervasive characteristic map stride [1]. Thereby, to handle the difficulty of investigating tiny targets in complex scenes, a new detection head is raised3 which consists of a small target detection module. Unlike the conventional detection head that performs classification and regression simultaneously, we decoupled these two tasks into separate branches. Moreover, we added an extra prediction head to the original three-layer detection structure, which is specifically designed for tiny object detection. The performance distinctions in infrared small target detection is largely triggered by two elements. For starters, present label assignment patterns, whether anchor-based or anchor-free, tend to inaccurately mark highly small ground-truth objects as background, which causes the detectors to attach less importance to tiny objects. Secondly, The detection heads designed by conventional object detection methods are difficult to give accurate predictions based on highly structured representations, where small objects fail to detect tiny targets. In conclusion, our conclusions are listed below:

- An effective multi-scale label assignment(MSLA) tactic is introduced. The MSLA can easily displace the point-based label assignment tactics in mainstream detectors and standardized boxes, thereby facilitating their performance on Infrared small targets.
- we designed a multi-scale small target detection head, and it is capable of improving the detecting performance of tiny objects without losing large targets.
- Our experiment results on the SISRT-V2 datasets reached SOTA.

II. METHODOLOGY

The one-stage detection framework ultralytics was selected as our datum line since it is the most convenient and remarkable one-stage detector[11]. The infrared small targets have a small sample size and few categories, so we add tiny CSDarknet as the backbone. we still keep FPN and PanNet architecture as our head, which is comprised of some bottom-up pathways and some top-down pathways [12][13]. To further optimize the whole architecture, we design an effective Mutiscale label assignment(MSLA) strategy. Then, we designed a multi-scale tiny object detection head to enhance the detecting rate of small targets. While we train the model through the SISRT-V2 dataset by means of a data augmentation tactic (MixUp, Mosaic).

A. MSLA strategy in tiny Object

Since the most extensive metric is adopted in target detection, IoU has restrictions for assessing the positional correlation between two boundary boxes [9]. After continuous improvement, CIOU has been adopted as a mainstream label for assigning losses. However, the enhancements may not address the issue that small targets are vulnerable to IoU. Therefore, for the problem of tiny and medium-sized infrared objects, we adopt the combination of CIOU and DotD. DotD is introduced as the evaluation index for high-resolution feature images, and for low-resolution images, CIOU is used as the index. This is called the MSLA strategy.

1) DIoU: Distance Intersection over Union(DIoU) considers the overlapping velocity, distance between the target and the anchor, and scale [14]. The loss function is listed below:

$$\mathcal{L}_{DIOU} = 1 - IOU + \frac{\rho^2\left(\mathbf{b}, \mathbf{b}^{gt}\right)}{C^2} \tag{1}$$

where b and b^{gt} represent the center point of the prediction box and the real box. ρ means the Euclidean distance between the center of b and b^{gt}, C means diagonal length of the tiniest enclosing box encompassing the two boxes. As it is used in loss function, the regression of the object box is swifter and more steady.

2) CIoU: considers aspect ratio on the grounds of DIOU metrics[14]. The loss function is defined as below

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2\left(\mathbf{b}, \mathbf{b}^{gt}\right)}{c^2} + \alpha v \tag{2}$$

here α represents an active trade-off parameter,

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{3}$$

v determines the uniformity of aspect ratio

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{4}$$

where w^{gt},h^{gt} denote width and height of the bounding box ground truth, w,h denote width and height of boundary box anchor.

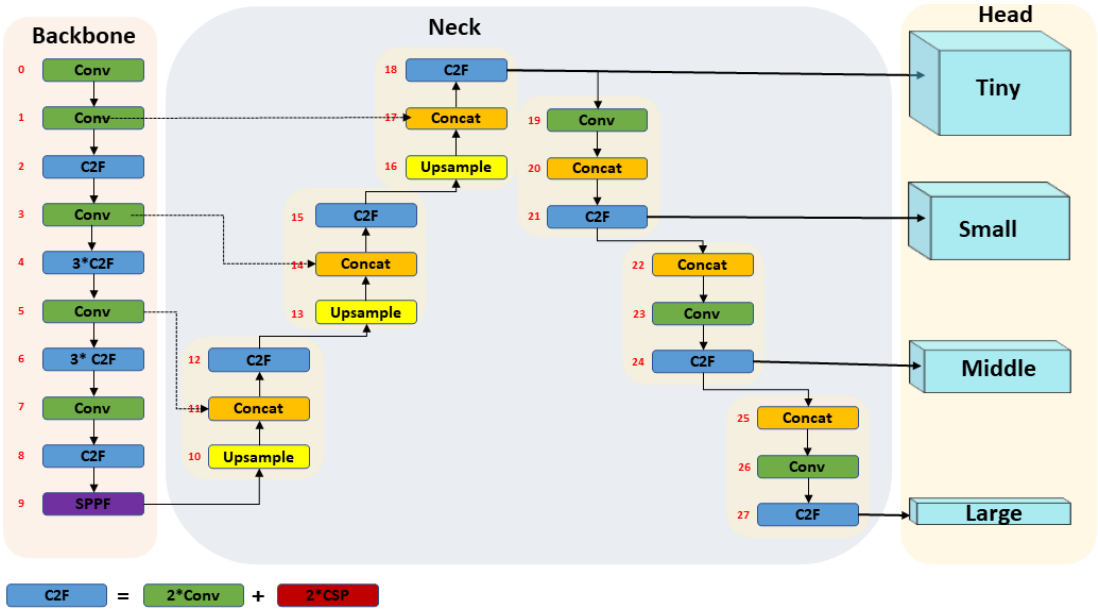

Fig. 2: The architecture of the MSHN. a) CSPDarknet53 backbone with four multi-scale detection heads at the end. b) The Neck uses a structure like PANet.

3) Dot Distance: Dot Distance (DotD) is derived from the observation that absolute and relative sizes of small boundary boxes are much smaller than those of big or medium-sized boundary boxes. The loss function is as below:

$$\mathcal{L}_{DotD} = 1 - e^{-\frac{\rho(\mathbf{b},\mathbf{b}^{gt})}{S}} \tag{5}$$

ρ denotes the Euclidean distance between the center of b and b^{gt} as below, and $\mathcal{S}$ is the average size of tiny objects in a certain dataset.

$$\mathcal{S} = \sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} w_{ij} \times h_{ij}}{\sum_{i=1}^{M} N_i}} \tag{6}$$

Where $w_i j$ and $h_i j$ denote the width and height of j-th bounding box within i-th image. M stands for the quantity of images in some dataset, N_i stands for the quantity of the marked boundary boxes within the i-th image.

B. Prediction head for tiny objects

The backbone as a categorization network, can not fulfill the positioning assignment. Furthermore, the head is devised to take responsibility for exploring the position and type of the subject through the characteristics maps from the backbone[4]. As shown in 3, Integrated with other three predicting heads, our four-head construction can alleviate the passive effect of violent target scale variance. the added predicting head (a) originates in an elevated-resolution, low-degree, characteristic map, which becomes more vulnerable to highly targets.

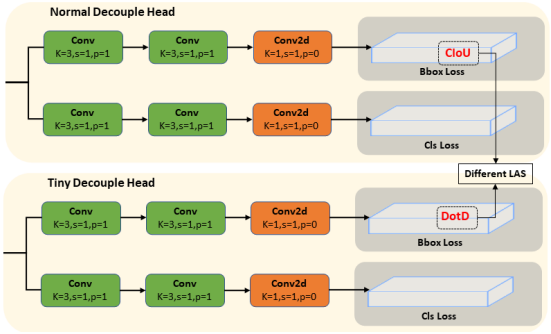

Fig. 3: The architecture of Head, contains two main blocks, a normal decouple head block and a tiny decouple head block. Normal decouple head help the network converge better and prevent the network from over fitting. Tiny decouple head can help the current node pay attention to tiny objects.

III. EXPERIMENTS

A. Dataset

SIRST-V2 is a specific dataset, and it is designed for single-frame infrared tiny object detection, where the graphs are chosen from thousands of infrared sequencing for disparate scenes. This dataset involves 1024 typical graphs largely at a resolution of 1280 × 1024. These graphs are acquired from the truthful videos of diverse scenes and they offer a tough trial bed to infrared tiny object detection in intricate real scenes.

B. Experimental Setting

The total experiments are executed through PyTorch on Ubuntu20 with two RTX 3090. The optimizer is designated as SGD showing a learning velocity of 1e - 2. The target detection is largely assessed by Precision, Recall, and mAP. Target detection outcomes are categorized into TP, FP, FN,

TABLE I: Comparison of performance among the proposed algorithm and the other five algorithms in SIRST-V2.

Method	$\mathbf{mAP_{50-90}}(\%)$	$\mathbf{AP_{50}}(\%)$	Recall(%)	Precies
FasterRCNN	62.28	80.7	59.2	0.68
FCOS	63.54	77.5	63.72	0.69
OSCAR	78.8	88.8	84.43	0.83
YOLOV5	71.2	83.5	79.51	0.78
YOLOV8	82.4	91.4	90.14	0.89
Ours	**83.9**	**93.1**	**94.0**	**93.0**

along with TN. TP denotes the positive specimen of the accurate categorization; FP denotes the positive specimen of the misclassification; FN denotes the passive specimen of the misclassification; and TN denotes the passive specimen of the accurate categorization. Precision is the proportion of the specimens which are accurately categorized as positive specimens to the whole quantity of detection specimens. The recall is the proportion of specimens of this category which are accurately categorized and recognized to the quantity of specimens in the whole object test sets of this category. In addition, the lost detection rate is used to recognized and assess the algorithm performance for the object frame. The formula is listed as below:

$$Re = \frac{TP}{TP + FN} \tag{7}$$

$$Pr = \frac{TP}{TP + FP} \tag{8}$$

We seek to gain the total potential values of Precision and Recall and Precision-Recall curve through calculation. The mean network precision is calculated as the area under the curve. Furthermore, the value varies between 0 and 1

$$AP_{IoU} = \int_0^1 P(R)dR \tag{9}$$

Precision and Recall values and figure out Precision-Recall curve

Precision and Recall values and calculate the Precision-Recall curve

C. Small-target Detection and Analysis

To affirm the feasibility and superiority of our proposed model, we compared it with five models on the SIRST-V2 dataset The specific experimental results are presented in Table I. As can be seen in TablesI our model has an AP50 value of 93.1%, which is significantly better than other detection models.

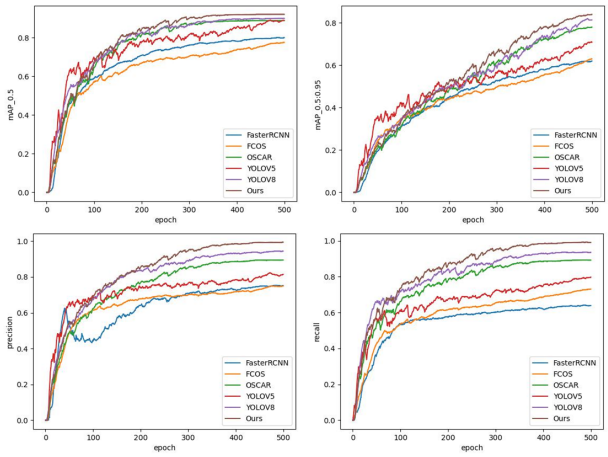

Fig. 4: Comparison of training results among the proposed algorithm and the other five algorithms in SIRST-V2.

IV. Conclusion

we introduce a effective Muti-scale label assignment(MSLA) tactic. The MSLA easily displace the standard box and point-based label task tactics in mainstream detectors. Hence, their performance on Infrared small target are facilitated. Our designed a multi-scale small target detection head, which can enhance the detecting performance of tiny objects without losing large targets. Widespread tests and analyses on SIRST-V2 dataset deeply verify the efficiency of MLSA-YOLOV8.

References

[1] Yimian Dai et al. “One-Stage Cascade Refinement Networks for Infrared Small Target Detection”. In: *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), pp. 1–17.

[2] Mingjing Zhao et al. “Single-frame infrared small-target detection: A survey”. In: *IEEE Geoscience and Remote Sensing Magazine* 10.2 (2022), pp. 87–119.

[3] Chang Xu et al. “RFLA: Gaussian receptive field based label assignment for tiny object detection”. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer. 2022, pp. 526–543.

[4] Xingkui Zhu et al. “TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 2778–2788.

[5] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015).

[6] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

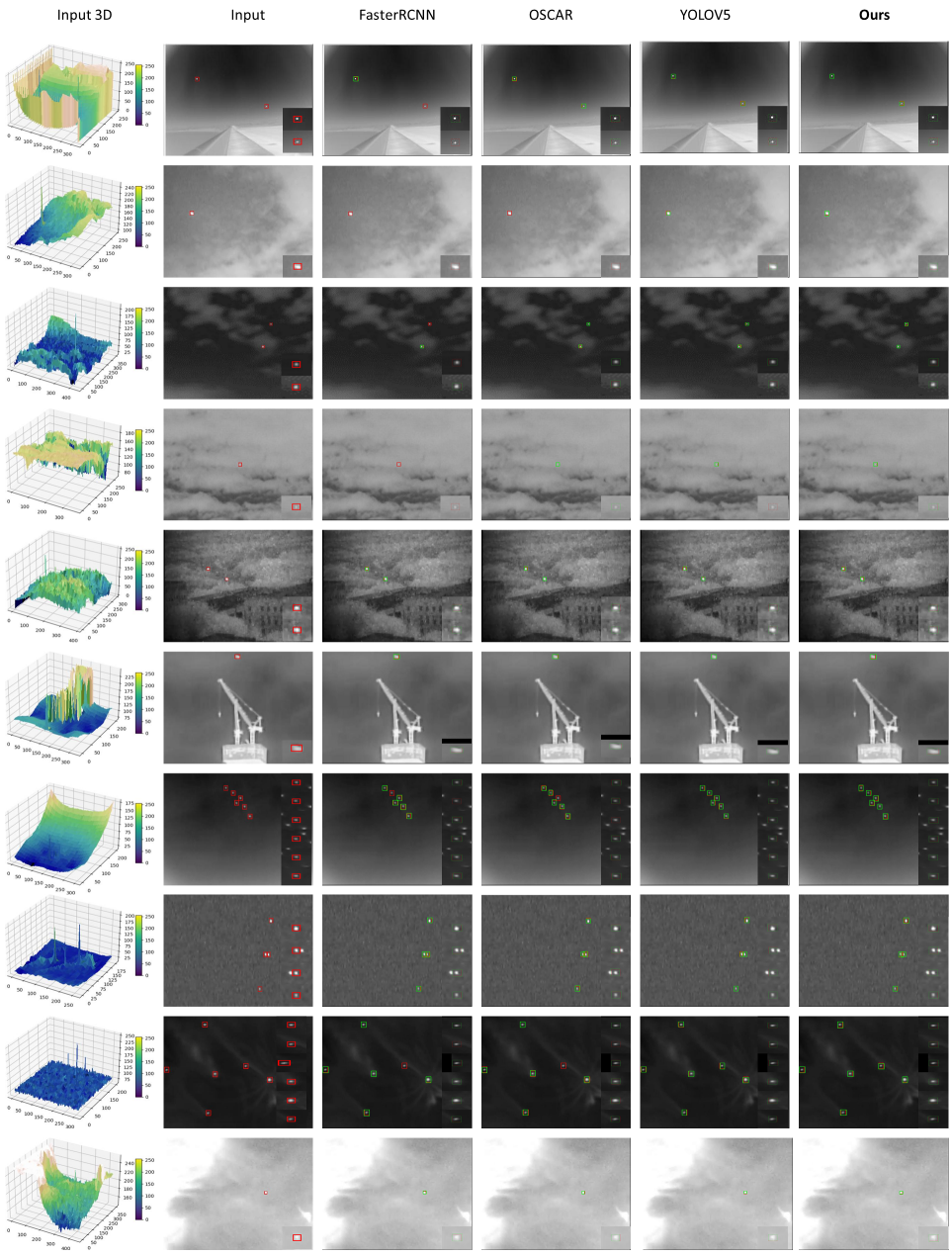

Fig. 5: Comparison of visualization results on SIRST-V2 dataset.

[7] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection". In: *arXiv preprint arXiv:2004.10934* (2020).

[8] Zhi Tian et al. "Fcos: Fully convolutional one-stage object detection". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 9627–9636.

[9] Chang Xu et al. "Dot distance for tiny object detection in aerial images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1192–1201.

[10] Shaoyu Chen et al. *TinyDet: Accurate Small Object Detection in Lightweight Generic Detectors*. 2023. arXiv: 2304.03428 [cs.CV].

[11] Glenn Jocher et al. *ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation*. Version v7.0. Nov. 2022. DOI: 10.5281/zenodo.7347926. URL: https://doi.org/10.5281/zenodo.7347926.

[12] Tsung-Yi Lin et al. *Feature Pyramid Networks for Object Detection*. 2017. arXiv: 1612.03144 [cs.CV].

[13] Junfeng Yang et al. "PanNet: A Deep Network Architecture for Pan-Sharpening". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 1753–1761. DOI: 10.1109/ICCV.2017.193.

[14] Zhaohui Zheng et al. *Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression*. 2019. arXiv: 1911.08287 [cs.CV].