applied sciences

MDPI

Article

Borrow from Source Models: Efficient Infrared Object Detection with Limited Examples

Ruimin Chen [1,2,3], Shijian Liu [1,2,*], Jing Mu [1,2,3], Zhuang Miao [1,2,3] and Fanming Li [1,2]

[1] Key Laboratory of Infrared System Detection and Imaging Technology, Chinese Academy of Sciences, Shanghai 200083, China; chenruimin@mail.sitp.ac.cn (R.C.); mujing@mail.sitp.ac.cn (J.M.); akkomz@mail.sitp.ac.cn (Z.M.); lifanming@mail.sitp.ac.cn (F.L.)
[2] Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: Shj_liu@ustc.edu

Abstract: Recent deep models trained on large-scale RGB datasets lead to considerable achievements in visual detection tasks. However, the training examples are often limited for an infrared detection task, which may deteriorate the performance of deep detectors. In this paper, we propose a transfer approach, Source Model Guidance (SMG), where we leverage a high-capacity RGB detection model as the guidance to supervise the training process of an infrared detection network. In SMG, the foreground soft label generated from the RGB model is introduced as source knowledge to provide guidance for cross-domain transfer. Additionally, we design a Background Suppression Module in the infrared network to receive the knowledge and enhance the foreground features. SMG is easily plugged into any modern detection framework, and we show two explicit instantiations of it, SMG-C and SMG-Y, based on CenterNet and YOLOv3, respectively. Extensive experiments on different benchmarks show that both SMG-C and SMG-Y achieve remarkable performance even if the training set is scarce. Compared to advanced detectors on public FLIR, SMG-Y with 77.0% mAP outperforms others in accuracy, and SMG-C achieves real-time detection at a speed of 107 FPS. More importantly, SMG-Y trained on a quarter of the thermal dataset obtains 74.5% mAP, surpassing most state-of-the-art detectors with full FLIR as training data.

Keywords: infrared object detection; limited training examples; knowledge transfer

Citation: Chen, R.; Liu, S.; Mu, J.; Miao, Z.; Li, F. Borrow from Source Models: Efficient Infrared Object Detection with Limited Examples. *Appl. Sci.* **2022**, *12*, 1896. https://doi.org/10.3390/app12041896

Academic Editor: Yue Wu

Received: 10 January 2022
Accepted: 10 February 2022
Published: 11 February 2022

1. Introduction

Recently, thermal infrared cameras have become increasingly popular in security and military surveillance operations [1,2]. Thus, infrared object detection, including both classification and localization of the targets in thermal images, is a critical problem to be invested in. With the advent of Convolution Neural Network (CNN) in many applications [3–7] such as action recognition and target tracking, a number of advanced models [8–10] based on CNN are proposed in object detection. Those detectors lead to considerable achievements in visual RGB detection tasks because they are mainly driven by large training data, which are easily available in the RGB domain. However, the relative lack of large-scale infrared datasets restricts CNN-based methods to obtain the same level of success in the thermal infrared domain [1,11].

One popular solution is finetuning an RGB pre-trained model with limited infrared examples. Many researchers firstly initialize a detection network with parameters trained on public fully-annotated RGB datasets, such as PASCAL-VOC [12] and MS-COCO [13]. Then, the network is finetuned by limited infrared data for specific tasks. To extract infrared object features better, most of the infrared detectors improve existing detection frameworks by introducing some extra enhanced modules such as feature fusion and background suppression. For example, Zhou et al. [14] apply a dual cascade regression mechanism to fuse high-level and low-level features. Miao et al. [15] design an auxiliary foreground

prediction loss to reduce background interference. To some extent, the aforementioned modules are effective for infrared object detection. However, it is hard for simple finetuning with inadequate infrared examples to eliminate the difference between thermal and visual images, which hinders the detection of infrared targets.

An alternative solution is to borrow some features from a rich RGB domain. Compared to the finetuning, this method leverages abundant features from the RGB domain to boost accuracy in infrared detection. Konig et al. [16] and Liu et al. [17] combine visual and thermal information by constructing multi-modal networks. They feed paired RGB and infrared examples into the network to detect the objects in thermal images. However, the paired images from two domains are difficult to be obtained, which hampers the development of the multi-modal networks. To tackle this problem, Devaguptapu et al. [1] employ a trainable image-to-image translation framework to generate pseudo-RGB equivalents from thermal images. Although this pseudo multi-modal detector is feasible in the absence of large-scale available datasets, the complicated architecture is difficult to train and thus rarely reaches advanced performance.

In this work, we address this problem from a novel perspective, knowledge transfer. Our proposed approach, named Source Model Guidance (SMG), is the first transfer learning solution for infrared limited-examples detection, to the best of our knowledge. By leveraging existing RGB detection models as source knowledge, we convert recent state-of-the-art RGB detectors to infrared detectors with inadequate thermal data. The basic idea is that if we already have an RGB model with strong ability to distinguish foreground from background, the model can be used as a source model to supervise another network training for infrared detection. Then, the problems becomes how to transfer the source knowledge between different domains and where to add the source supervision.

We first observe modern RGB detection frameworks including anchor-based (Faster RCNN [8], SSD [9], YOLOv3 [18]) and anchor-free (CenterNet [19], CornorNet [20], ExtremeNet [21], FCOS [22]) methods. All of them consist of two main modules, a Feature Extraction Network (FEN) to calculate feature maps and a Detection Head (DH) to generate results. Many researchers have trained those frameworks with large-scale RGB datasets and exposed network weights as common RGB object detection models. Despite the fact that an RGB model is designed for visual images, it still can detect most infrared targets when given a thermal image. However, the precise categories and bounding boxes are hard to be predicted by it due to the difference between two domains. Therefore, we combine all category predictions as a foreground soft label, which is regarded as the source knowledge to be transferred. Then, we look for where to add the source supervision. Different from ground-truth supervision on the final DH, we propose a Background Suppression Module (BSM) to receive the source knowledge. BSM is inserted after FEN to enhance the feature maps and produce a foreground prediction at the same time. By calculating the transfer loss between the foreground prediction and the soft label, we introduce source supervision into the training process of the infrared detector, as shown in Figure 1.

Theoretically, our transfer approach SMG can be implemented in any visual detection networks effortlessly. In this paper, we choose two popular frameworks, CenterNet [19] and YOLOv3 [18], as instantiations, and the frameworks we proposed are named SMG-C and SMG-Y, respectively. To validate the performance of SMG, we conduct extensive experiments on two infrared benchmarks, FLIR [23] and Infrared Aerial Target (IAT) [15]. Experimental results show that SMG is an effective method to boost detection accuracy especially when there are limited training examples. On FLIR, using only a quarter of training data, SMG-Y obtains higher mAP than the original YOLOv3 finetuned on the entire dataset. Furthermore, compared to other infrared detectors, both SMG-C and SMG-Y achieve state-of-the-art accuracy and inference speed.

The main contributions are described as the following three folds:

- First, we propose a cross-domain transfer approach SMG, which easily converts a visual RGB detection framework to an infrared detector.

- Second, SMG decreases the data dependency for an infrared network. The detectors with SMG maintain remarkable performance even if trained on the small-scale datasets.
- Third, two proposed instantiations of SMG, SMG-C and SMG-Y, outperform other advanced approaches in accuracy and speed, showing that SMG is a preferable strategy for infrared detection.

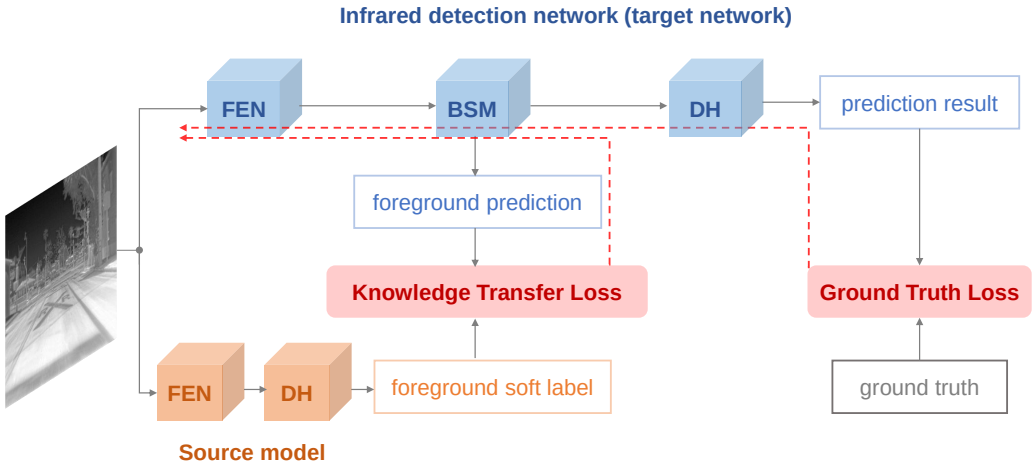

Figure 1. The overall framework of SMG, which mainly consists of two parts: a source model to provide source knowledge and a target network to predict infrared detection results. Red arrows indicate the backpropagation pathways.

The structure of this paper is as follows. In Section 2 , we briefly present some aspects related to our work. Section 3 shows the proposed method SMG in detail. Extensive experiments and ablation studies are conducted in Sections 4 and 5, respectively. We explain why SMG works well and analyze the failure cases of our detectors in Section 6. Finally, the summary is drawn in Section 7.

2. Related Work

In this section, we briefly introduce recent object detection frameworks including both visual and infrared methods. In addition, we describe the knowledge transfer, which is the inspiration of our method.

2.1. Object Detection

Current object detection frameworks can be divided into two groups: anchor-based methods such as Faster RCNN [8], SSD [9], and YOLOv3 [18] and anchor-free methods represented by CenterNet [19], CornorNet [20], ExtremeNet [21], and FCOS [22]. Anchor-based methods firstly define a series of rectangle bounding boxes, called anchors, as proposal candidates. Then, all potential object detections are enumerated exhaustively according to proposed anchors. Finally, additional Non-Maximum Suppression (NMS) [24] is used to remove duplicated locations for the same instance. To avoid the redundant design of anchors and lessen the computation burden, anchor-free methods regard the detection problem as a keypoint estimation without pre-defined anchors. For example, CenterNet [19] predicts the center point of an object and then regresses to other properties such as object size. Although those algorithms achieve remarkable performance, they are mainly driven by extensive public training data and focus on detecting the targets in standard visual RGB images. For infrared detection, the lack of large-scale labeled thermal images hinders the power of detectors based on CNN. Researchers cope with this problem from two aspects: one is finetuning a pre-trained model [14,15], the other is introducing corresponding RGB images as supplements [1,16]. The first strategy hardly makes full use of the information from the RGB domain, and the sophisticated structures in the second method are difficult to be performed. Different from two solutions, our SMG not only

leverages existing RGB models as the guidance for infrared detectors but also is easily plugged in any modern detection framework.

2.2. Knowledge Transfer

Knowledge transfer is a popular strategy to tackle various problems, such as object classification [25–27], model compression [28,29], and detection [30–32]. It first distills knowledge from a trained model (source) and then transfers the knowledge to another network (target). Hinton et al. [25] introduce the concept of soft label as the guidance in knowledge transfer for classification tasks. In comparison with the hard label such as ground truths, the soft label is a softened version of the final output from the source model. Benefiting from the soft label, the target network can learn how the source model classifies different objects. Many methods [28,29] with soft label obtain achievement in classification and retain accuracy in model compression. However, applying transfer techniques to object detection is challenging because detection is a more complex task that combines regression, region proposals, and classification. To tackle this problem, Chen et al. [31] designed a novel teacher bounded regression loss for knowledge transfer and adaptation layers to better learn from the source model. Although this method is easy to be applied in object detection, the method is driven by large-scale training datasets. Some researchers try to perform transfer learning in few-shot detection and construct a target-domain detector with very few training data. Chen et al. [32] alleviate transfer difficulties in low-shot detection by adding a background-depression regularization and designing a deep architecture, a combination of SSD and Faster RCNN, called LSTD. However, LSTD is suitable for RGB object detection without involving the transfer between different domains. Additionally, it just masks feature maps with the ground-truth bounding boxes in the background-depression regularization, which damages the features extracted from the backbone. Different from LSTD, our SMG introduces an independent block BSM to enhance the foreground features of thermal infrared images by taking advantage of the knowledge from the visual RGB domain.

3. Method

In this section, we detail our method Source Model Guidance (SMG). First, we introduce the structure of SMG, including the overall framework and proposed Background Suppression Module (BSM). Then, we describe the training details of SMG, including how to transfer knowledge from the source model to the target network and how to train the whole network. Finally, we show two explicit instantiations of SMG, SMG-C and SMG-Y.

3.1. Overall Framework

As illustrated in Figure 1, we train an infrared object detector by using the knowledge of a source model. The source model is a high-capacity RGB detection model, which has been trained with large-scale RGB datasets. The source model is composed of two modules, a Feature Extraction Network (FEN) for feature map calculation and a Detection Head (DH) to generate the prediction. We choose two popular detection models, CenterNet [19] and YOLOv3 [18], as source models to guide different infrared detectors, named SMG-C and SMG-Y, respectively.

Compared to the source model, the infrared detection network not only consists of FEN and DH but also has an extra part, Background Suppression Module (BSM). The structure of FEN is flexible, and it can be the same or different from the source model. The DH in an infrared detection network is similar to the source model except for the predicted category. For BSM, it is a novel part with two functions, predicting the foreground and enhancing the feature map from FEN.

3.2. BSM

The BSM in the infrared detection network (target network) is a key module to receive the knowledge transferred from the source model. We describe the principle of BSM, as shown

in Figure 2. The idea of BSM is inspired by the concept of attention mechanism [33–37], and thus, its main structure is a transformation mapping from the input $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ to an enhanced feature map $\mathbf{X}' \in \mathbb{R}^{H \times W \times C}$. In addition, an extra prediction, named foreground prediction $\mathbf{P_{FG}} \in \mathbb{R}^{H \times W \times k}$, is obtained in BSM. The $\mathbf{P_{FG}}$ is defined as the combination of ground-truth targets based on anchors, where k is the number of anchors and k is 1 for anchor-free methods.

To be specific, the input $\mathbf{X}$ first passes two convolutional layers to produce an intermediate feature map. Then, it is fed into to two different branches: one for predicting foreground and the other for feature enhancement. The foreground prediction is achieved by a convolution with sigmoid function to generate a score $\mathbf{P_{FG}}$. The intermediate feature map is also employed to re-weight the input feature map over spatial dimension because it reflects the feature of the foreground. After a 1×1 convolution for channel transformation, we use an average pooling to squeeze global information into channel-wise weights. Finally, the enhanced feature map branch $\mathbf{X}'$ is obtained by rescaling input $\mathbf{X}$ with the weights.

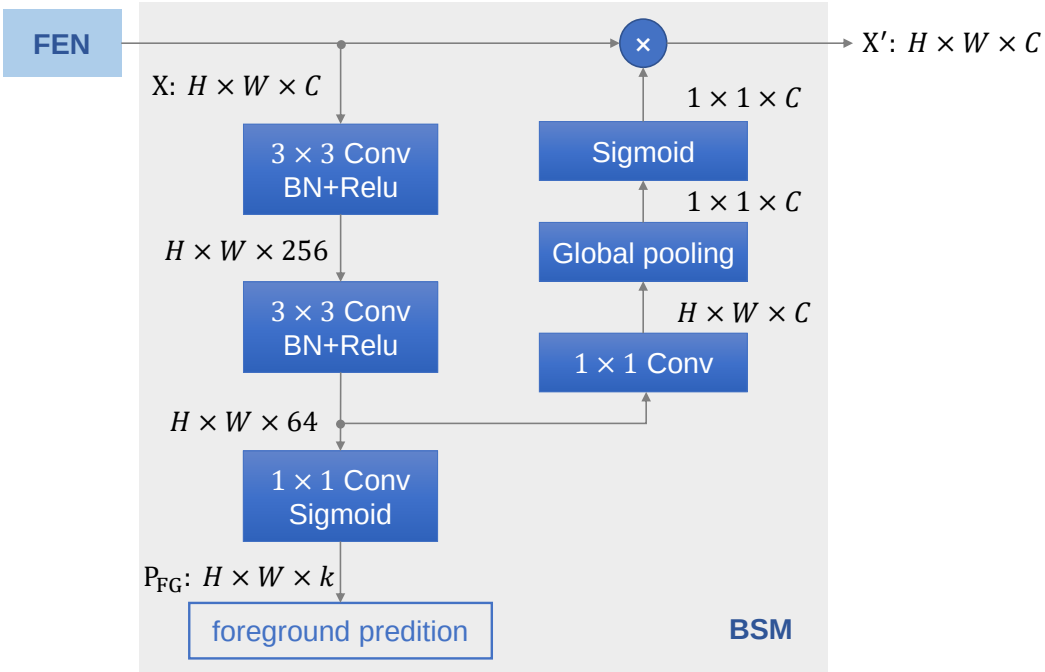

Figure 2. The network structure of BSM.

3.3. Transfer-Knowledge Regularization

Although the foreground enhancement in BSM can alleviate the disturbance of background, the foreground prediction $\mathbf{P_{FG}}$ from BSM should be supervised in the limited-examples scenario. For this reason, we propose a novel transfer-knowledge regularization by leveraging the source model as a guidance.

In this paper, the foreground prediction $\mathbf{P_{FG}}$ with values within 0 and 1 is supervised by the foreground soft label $\mathbf{S_{FG}}$ generated from the source model. Different from the hard label in ground-truth supervision, we adopt the soft label in knowledge transfer because it contains hidden information about how the source model makes inferences when given samples. In every position of $\mathbf{S_{FG}}$, the value of the soft label is in $[0, 1]$ based on anchor, while the hard label is either 0 or 1.

For different source models, we choose different methods to obtain the foreground soft label $\mathbf{S_{FG}}$. We sum the label prediction (heatmap) for all positions in SMG-C and use the anchor confidence directly in SMG-Y, as shown in Figures 3 and 4. The soft label $\mathbf{S_{FG}}$ is the foreground score based on anchor and has the same size with foreground prediction $\mathbf{P_{FG}}$ from the target network. We take $\mathbf{S_{FG}}$ as source-domain knowledge to

regularize the training of target network. Mean Squared Error (MSE) is applied as a transfer-knowledge regularization:

$$\mathcal{L}_{TK} = MSE(\mathbf{S_{FG}}, \mathbf{P_{FG}}). \quad (1)$$

In this case, the trained RGB detection model can be integrated into the training procedure of the infrared detector, which achieves cross-domain transfer in SMG.

3.4. Training Algorithm

The whole loss $\mathcal{L}$ of SMG consists of two parts: one is the standard detection loss with ground truth supervision $\mathcal{L}_{GT}$, and the other is the transfer-knowledge loss $\mathcal{L}_{TK}$ mentioned in the above subsection:

$$\mathcal{L} = \mathcal{L}_{GT} + \lambda \mathcal{L}_{TK}. \quad (2)$$

The weight λ represents hyper-parameters to control the balance between different losses. We fix it to be 1 in SMG-C. In SMG-Y, λ is 0.3 because we introduce 3 BSMs to generate the transfer-knowledge loss in SMG-Y, as explained in the following subsection.

During the training, we first initialize the source model with public parameters trained on COCO, which is a large-scale RGB detection dataset. For the target network, the FEN is initialized with ImageNet pretrained parameters, and other modules are randomly initialized. Then, training loss is calculated according to Equation (2). Finally, we update the weights of target network in the back propagation. It is notable that the source model is not updated, and thus, we just employ the target network as an infrared detector in the inference.

3.5. Instantiations

SMG can be implemented in standard visual RGB detection networks and convert those networks to infrared detectors. To illustrate this point, we apply SMG in both anchor-free and anchor-based detection frameworks, which is described next.

We first consider CenterNet [19], an anchor-free model, as an instantiation, and the framework we proposed is named SMG-C. As shown in Figure 3, CenterNet predicts center points of targets directly by producing a heatmap $\hat{Y} \in [0,1]^{H \times W \times class}$, where $class$ is the number of categories (for RGB models trained on COCO, $class = 80$). Therefore, the sum of the heatmap represents foreground prediction, and we use it as $\mathbf{S_{FG}}$ to transfer knowledge. For the infrared detection network of SMG-C, only a BSM is inserted in between FEN and DH in comparison with CenterNet.

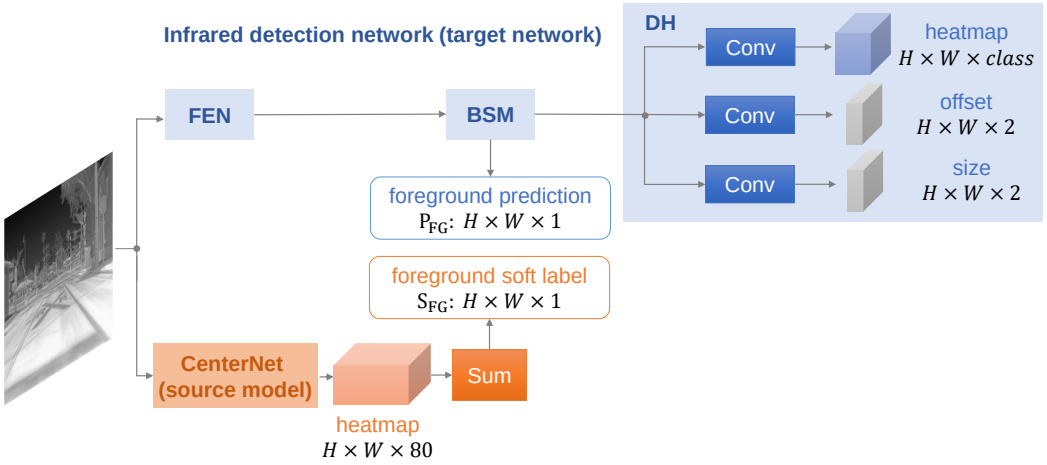

Figure 3. The framework of SMG-C.

SMG is also applied in YOLOv3 [18], an anchor-based model, and Figure 4 shows the framework of SMG-Y. YOLOv3 predicts bounding boxes at 3 different scales by extracting

features from 3 scales. As a result, we add 3 BSMs in the infrared detection network. Furthermore, YOLOv3 sets k anchors with different sizes, and thus, the prediction in every scale is a k-d tensor encoding location, confidence, and class. The confidence reflects whether there is an object in the anchor, and we adapt it as the foreground soft label $\mathbf{S_{FG}}$ directly. In this work, we set $k = 3$ according to the original paper [18].

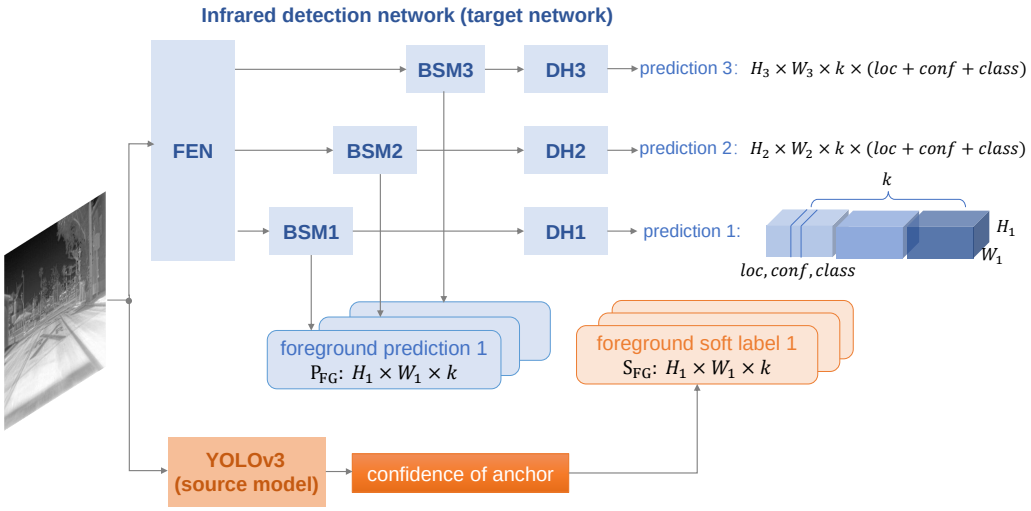

Figure 4. The framework of SMG-Y.

4. Experiments

In this section, we first introduce experimental details and the training datasets we use in this paper. Then, we conduct extensive experiments to evaluate the detection performance of two frameworks, SMG-C and SMG-Y. Finally, our method is compared with some popular detectors on the public FLIR benchmark.

4.1. Dataset and Experimental Setup

We adopt the public FLIR dataset [23] and self-build IAT dataset [15] for our experimental studies.

FLIR [23] collects 9214 infrared images with annotations, where the labeled objects contain a person, car, and bicycle. It is acquired via a thermal camera mounted on a vehicle, and all images are taken on the streets and highways, as illustrated in Figure 5. To evaluate the capability of our method with limited data, we perform experiments with full, half, and one-quarter of training examples in FLIR. The statistics of the training datasets are shown in Table 1. Although the numbers of training images are different in the three datasets, their test sets are the same as those provided in the FLIR benchmark.

Table 1. Numbers of instances on FLIR datasets.

Dataset	Person	Car	Bicycle
FLIR	22,372	41,260	3986
FLIR-1/2	10,997	20,700	1979
FLIR-1/4	5574	10,286	928

The IAT [15] consists of 2750 infrared images with aerial targets, including five categories: airline, bird, fighter, helicopter, and trainer. All images are captured by ground-to-air infrared cameras, and some samples on IAT are shown in Figure 6. Different from the images with target occlusions in FLIR, IAT contains small targets in complex aerial backgrounds, and the main challenge of it is background interference. We split IAT with the ratio of 7:3 as the training set and test set, respectively. Similar to FLIR, we use all and half of the training images to implement experiments, as presented in Table 2.

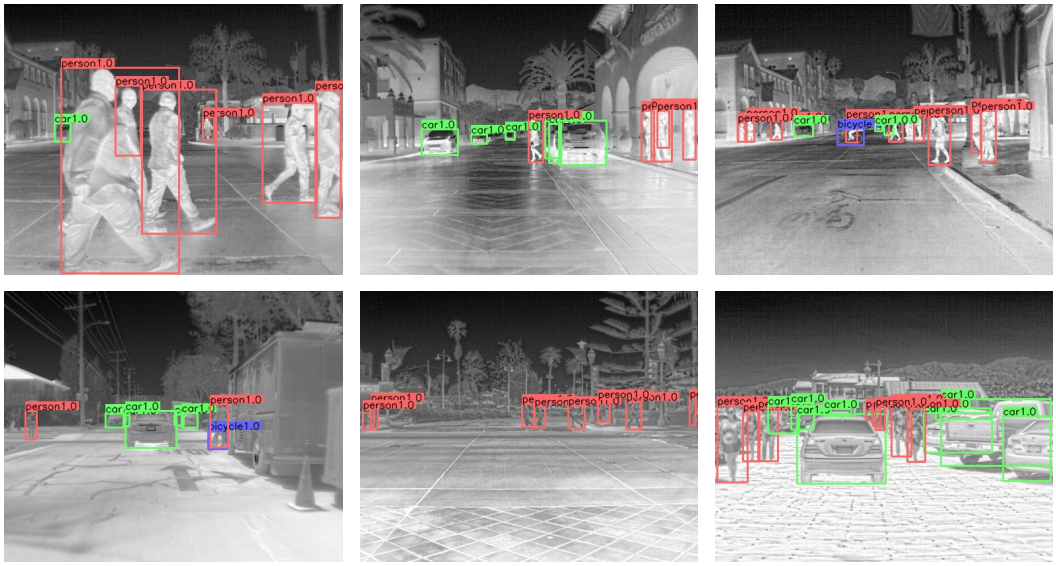

Figure 5. Samples on FLIR dataset.

Table 2. Numbers of instances on IAT datasets.

Dataset	Airline	Bird	Fighter	Helicopter	Trainer
IAT	121	535	667	310	469
IAT-1/2	64	277	321	152	242

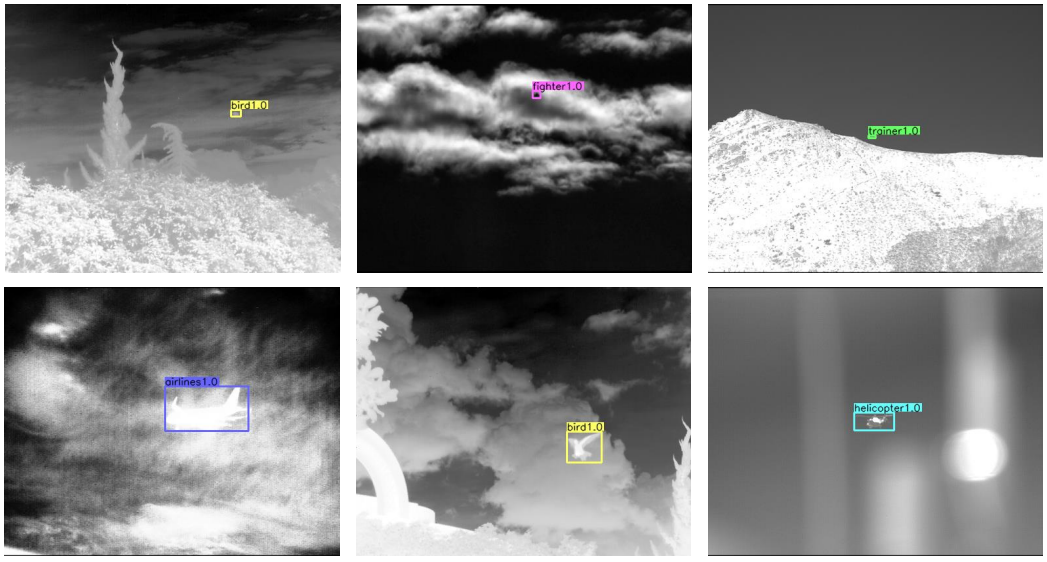

Figure 6. Samples on IAT dataset.

All experiments are implemented on a PC with an i7-8700K CPU and a signal GTX1080Ti GPU. For SMG-C, we adopt CenterNet with ResNet-18 [19] as the source model, because it is light-weight and enough to provide the guidance. The FEN of the target network in SMG-C is the fully convolutional upsampling version of Deep Layer Aggregation (DLA-34) [38]. For SMG-Y, YOLOv3 with DarkNet-53 [18] is used as the source model and the backbone of the target network is DarkNet-53. The source models of two frameworks are RGB detection models trained on COCO [13].

The input resolution is set to 512 × 512 in SMG-C and 416 × 416 in SMG-Y. During the training process of two frameworks, we follow their original papers [18,19] separately for training setting and hyper-parameters, unless specified otherwise. In the inference, we

evaluate the performance with the mean Average Precision (mAP) at IoU of 0.5, which is a common metric for object detection tasks.

4.2. SMG-C Results

We use SMG-C as the detection framework and implement experiments on both FLIR and IAT benchmarks. The baseline method in this subsection is the original CenterNet [19] without SMG.

Table 3 shows the comparison of AP for each class and mAP of SMG-C against the baseline detection network when trained with different numbers of training examples on the FLIR benchmark. One can see that our SMG-C outperforms the baseline detector across all classes when trained with the same dataset. For example, SMG-C on FLIR obtains 75.6% mAP, which is 4.5% higher than the baseline. This can be attributed to the fact that the source model offers sufficient guidance for the infrared detector in SMG.

More importantly, SMG-C achieves outstanding performance when the training data are insufficient. Taking the bicycle as example, we find that its AP maintains 51.5%, although the training examples are reduced to 1/4 of the original. In contrast, the highest bicycle's AP is 51.2% for the baseline method. Furthermore, the mAP of SMG-C trained on FLIR-1/2 obtains 73.3% mAP, surpassing the original CenterNet trained on the entire FLIR (71.1%).

Table 3. Detection results of SMG-C on the FLIR benchmark.

Dataset	Method	mAP (%)	AP (%)		
			Person	Car	Bicycle
FLIR	Baseline	71.1	76.6	85.4	51.2
	SMG-C	75.6	79.0	85.8	62.0
FLIR-1/2	Baseline	68.1	75.1	83.5	45.8
	SMG-C	73.3	78.7	86.0	55.3
FLIR-1/4	Baseline	65.8	71.5	81.8	44.1
	SMG-C	70.9	76.7	84.5	51.5

We also report the results on the IAT benchmark in Table 4. All mAPs of SMG-C exceed 95%, while the highest accuracy of CenterNet is only 93%. When we reduce training datasets to half of the original, the accuracy of the baseline drops to 90.6%, while SMG-C maintains 95.2% in mAP. Furthermore, SMG-C trained on IAT-1/2 surpasses the baseline method trained with the entire training dataset. This demonstrates that SMG-C yields an effective infrared detection method even when there are a lack of available training data.

Some results on IAT-1/2 are visualized in Figure 7. When the target is small, some interference from the background may adversely affect the detection especially in the absence of enough training examples. As shown in Figure 7, the baseline CenterNet hardly overcomes this problem so as to generate many wrong detection results. However, SMG-C guided by the high-performance RGB model suppresses the interference from the background and predicts more precisely than the baseline.

Table 4. Detection results of SMG-C on the IAT benchmark.

Dataset	Method	mAP (%)
IAT	Baseline	93.0
	SMG-C	96.8
IAT-1/2	Baseline	90.6
	SMG-C	95.2

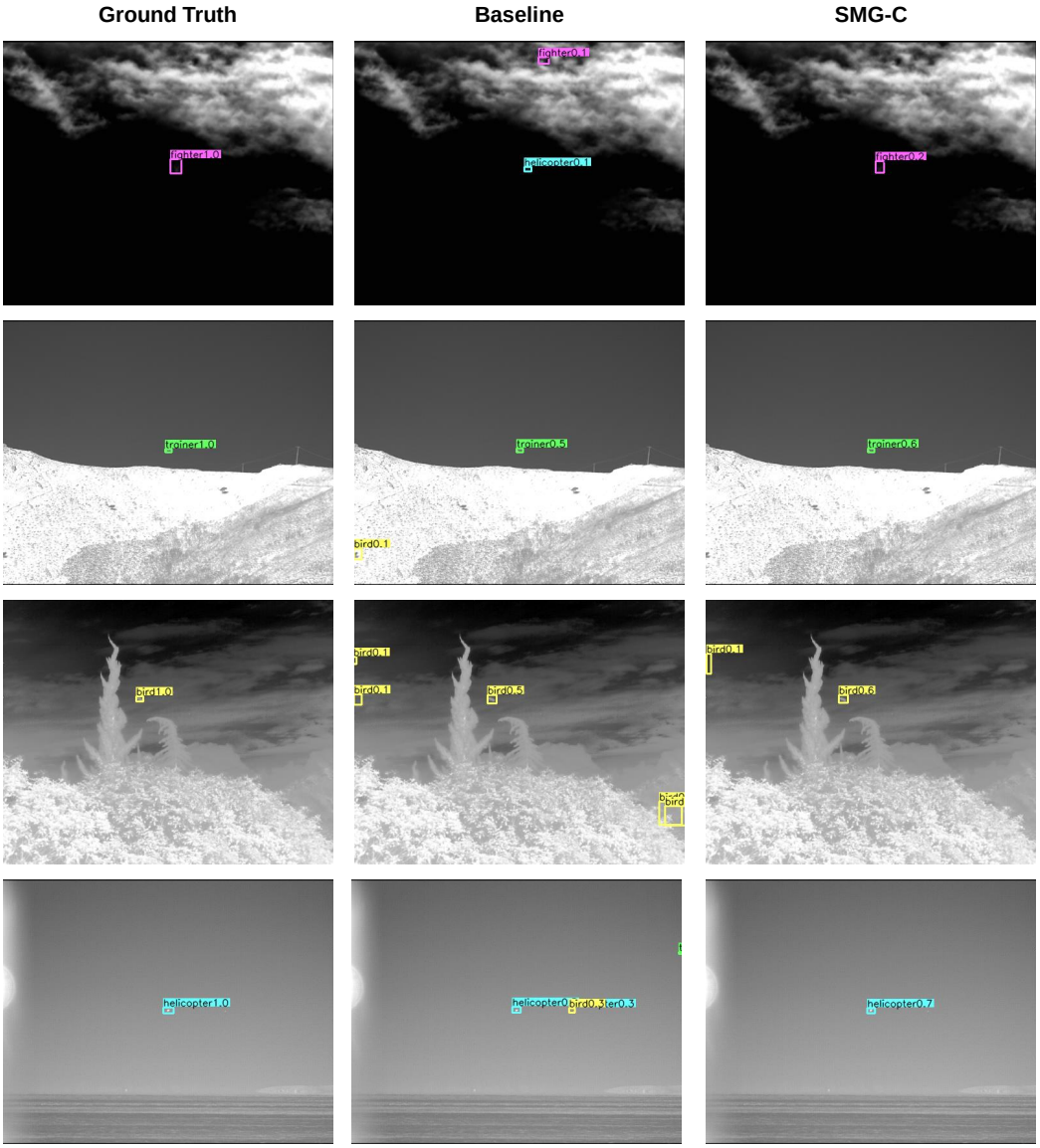

Figure 7. Visualization results on IAT-1/2.

4.3. SMG-Y Results

Similar to SMG-C, we conduct experiments on both FLIR and IAT datasets to evaluate the performance of SMG-Y. SMG-Y is compared with the baseline detector, YOLOv3 [18].

Table 5 presents the results of SMG-Y on the FLIR benchmark. The mAP of SMG-Y exceeds the baseline method nearly 10% on the same dataset, and the gap of them increases with the decrease of training examples. On FLIR-1/4, SMG-Y achieves 62.5% AP in bicycle detection in comparison with 29.1% for the baseline. We also observe that the accuracy of SMG-Y on FLIR-1/4 (74.5% mAP) outperforms the baseline method trained with full FLIR (69.4% mAP), which demonstrates SMG-Y maintains remarkable accuracy with limited training data. When the dataset is reduced to 1/4 of the original, the mAP of SMG-Y decreases by 2.5% (from 77.0% to 74.5%). However, the mAP of the baseline method drops by 13.2% (from 69.4% to 56.2%). The low reduction of SMG-Y indicates that it can take full advantage of the knowledge from the source model and decrease the data dependency of the network.

Table 5. Detection results of SMG-Y on the FLIR benchmark.

Dataset	Method	mAP (%)	AP (%)		
			Person	Car	Bicycle
FLIR	Baseline	69.4	74.5	84.4	49.2
	SMG-Y	77.0	78.5	86.6	65.8
FLIR-1/2	Baseline	64.9	68.5	82.1	44.0
	SMG-Y	75.4	76.9	86.7	62.7
FLIR-1/4	Baseline	56.2	61.2	78.3	29.1
	SMG-Y	74.5	76.6	84.4	62.5

We visualize some results of SMG-Y and its baseline YOLOv3 when both of them are trained on FLIR-1/4, as shown in Figure 8. We find that the baseline method hardly predicts the position of the bicycle because it is always obscured by people. Furthermore, due to insufficient training data, YOLOv3 is difficult to recognize objects with special gestures, such as the sitting woman in the last row of Figure 8 (note that most people in the training dataset are walking or riding). However, SMG-Y overcomes those problems and detects precisely under the circumstances of severe occlusion and appearance change even if the training examples are limited.

Experiments are also conducted on the IAT benchmark, and the results are shown in Table 6. We witness a sharp fall in the baseline accuracy as the number of training instances decreases. In contrast, SMG-Y trained on IAT-1/2 keeps competitive accuracy with 96.2% mAP, which is slightly lower than that trained on the full IAT dataset.

Table 6. Detection results of SMG-Y on the IAT benchmark.

Dataset	Method	mAP (%)
IAT	Baseline	92.5
	SMG-Y	97.8
IAT-1/2	Baseline	88.3
	SMG-Y	96.2

4.4. Comparison of SMG-C and SMG-Y

We compare two instantiations and their baseline methods in Figure 9. It is notable that SMG-Y outperforms SMG-C but YOLOv3 is inferior to CenterNet. In other words, the gap between SMG-Y and its baseline is larger in comparison with SMG-C. To be specific, SMG-Y achieves 77.0% mAP, which is 7.6% higher than its baseline when trained on a full FLIR. In contrast, SMG-C obtains 75.6% mAP, exceeding its baseline by 4.5%. We attribute this phenomenon to the fact that three different BSMs are added in SMG-Y to receive knowledge from different scales, and only one BSM is inserted in SMG-C.

Additionally, the data dependency for a detector can be reflected in the performance degradation when we reduce the training examples, which is also the slope of the curves in Figure 9. The decline of CenterNet is less than that of YOLOv3 due to the different principles between two frameworks: one is anchor-free and the other is anchor-based. We observe that the curves of both SMG-Y and SMG-C are smoother than their baselines. For example, a slight reduction in mAP can be witnessed in SMG-Y while its baseline accuracy drops dramatically, which indicates that SMG is an efficient strategy to decrease the data dependency for an infrared detection network.

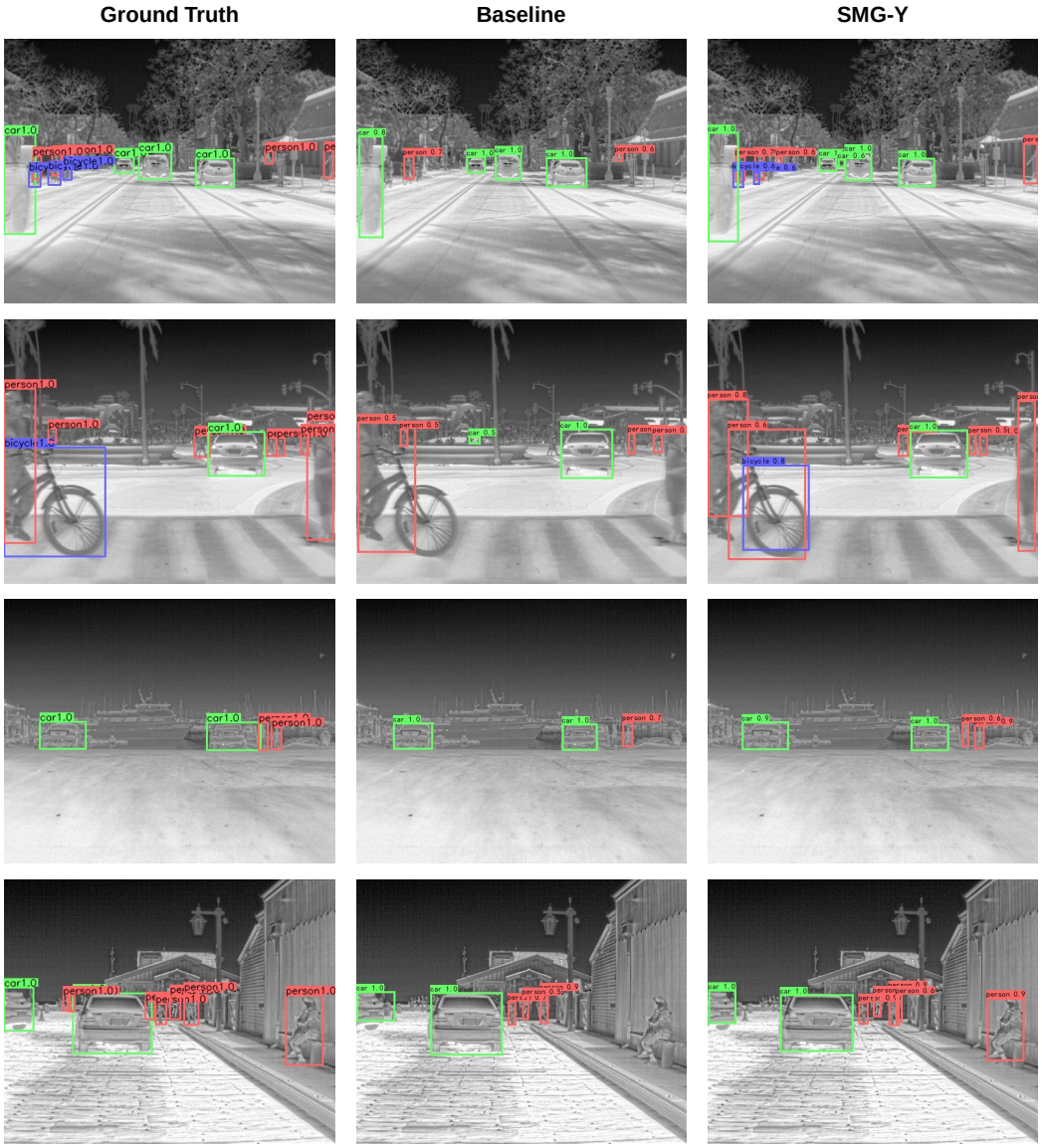

Figure 8. Visualization results on FLIR-1/4.

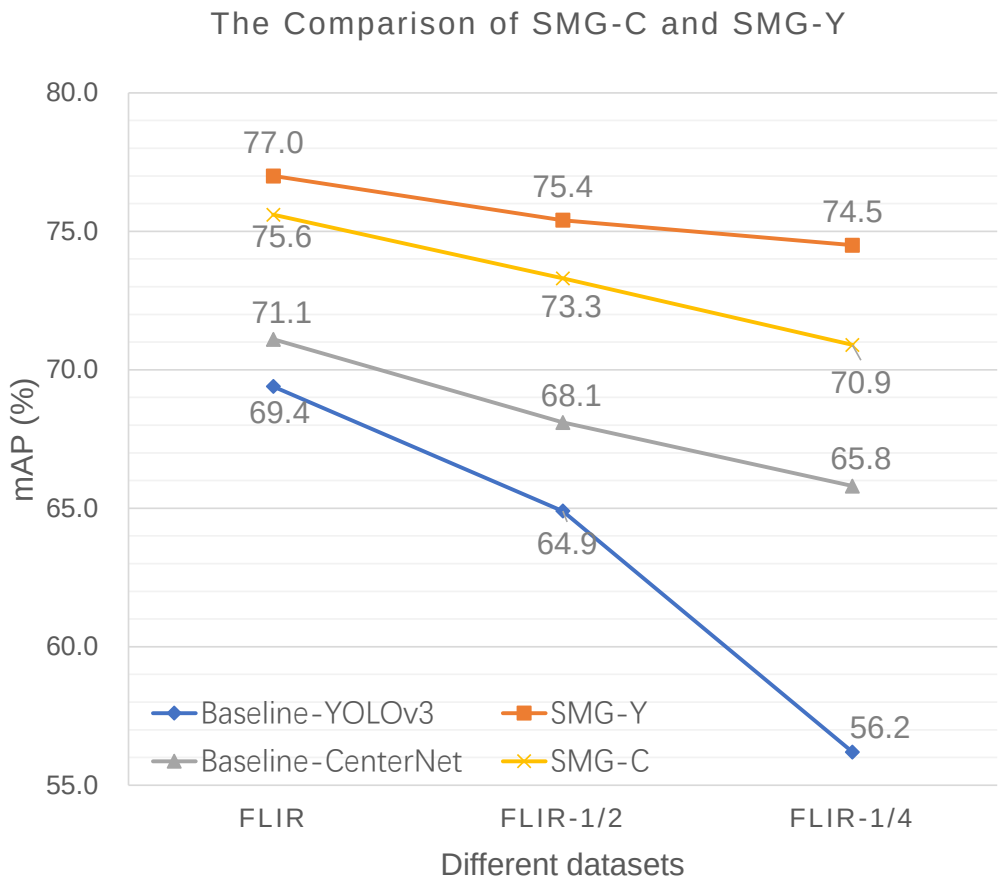

Figure 9. The comparison of SMG-C and SMG-Y in terms of mAP.

4.5. Comparison with State-of-the-Arts

Our frameworks (SMG-C and SMG-Y) are compared with some recent state-of-the-art detectors on the FLIR benchmark. The compared trackers are divided into two categories, visual and infrared detectors. The visual detectors such as SSD [9], YOLOv3 [18], Faster-RCNN [8], CenterNet [19], and RefineNet [39] are designed for RGB object detection and finetuned on the training set of FLIR. The infrared detectors, including MMTOD-CG [1], MMTOD-UNIT [1], Effi-YOLOv3 [40] and Pesudo-two-stage [14], are applied for thermal images directly.

We present the qualitative results in Table 7. It is remarkable that two proposed detection frameworks achieve outstanding performance. Specifically speaking, SMG-Y obtains the highest mAP with 77.0% and the AP of person, car, and bicycle are 78.5%, 86.6%, and 65.8%, respectively. It outperforms advanced detectors in mAP, and the speed of it maintains 40 frames per second (FPS), keeping the balance of accuracy and speed. Despite the slightly lower mAP (75.6%) in comparison with SMG-Y, SMG-C runs at the speed of 107 FPS, which is five times faster than other infrared detectors. Compared to the high-speed detector CenterNet [19], SMG-C gains 4.5% improvement in mAP, which shows that SMG-C is an efficient real-time detector.

More importantly, SMG-Y with 1/4 training data also achieves 74.5% mAP, surpassing all visual detectors and most infrared detectors trained on full FLIR. The bicycle accuracy in SMG-Y-1/4 is 62.5% AP, which is on par with that of Pseudo-two-stage [14]. Note that the training dataset of SMG-C-1/4 only contains 928 bicycle instances, while Pseudo-two-stage [14] is trained with 3986 examples for bicycle detection.

Table 7. Detection results of different detectors on the FLIR benchmark.

Category	Model	mAP (%)	AP (%)			FPS
			Person	Car	Bicycle	
Visual detectors	SSD [9]	62.1	63.1	75.8	47.5	24
	YOLOv3 [18]	69.4	74.5	84.4	49.2	42
	Faster-RCNN [8]	70.9	71.3	75.8	61.8	8
	CenterNet [19]	71.1	76.6	85.4	51.2	107
	RefineDet [39]	74.3	**79.4**	85.6	58.0	22
Infrared detectors	MMTOD-CG [1]	61.4	63.3	70.6	50.3	-
	MMTOD-UNIT [1]	61.5	64.5	70.7	49.4	-
	Effi-YOLOv3 [40]	70.8	74.5	84.7	53.2	22
	Pseudo-two-stage [14]	75.6	78.7	85.5	62.5	21
	SMG-C	75.6	79.0	85.8	62.0	**107**
	SMG-Y	**77.0**	78.5	**86.6**	**65.8**	40
	SMG-Y-1/4	74.5	76.6	84.4	62.5	40

SMG-Y-1/4 is trained on FLIR-1/4, and the other detectors are trained on FLIR.

5. Ablation Studies

In this section, we conduct ablation studies with SMG-C to understand the effect of image resolution, guidance, and backbone. All networks are evaluated on the FLIR benchmark, and the source model is CenterNet with ResNet-18 [19].

5.1. Effect of Image Resolution

We employ ResNet-18 as the FEN in the target network, and the compared baseline is the original CenterNet without SMG. Table 8 presents the mAP of two methods when the image resolution is changed from 384×384 to 512×512 . It is obvious that the higher resolution contributes to better accuracy. However, at different resolutions, SMG-C exceeds the baseline more than 5% in mAP. It indicates that the image resolution just affects the performance of the baseline network and has less influence on SMG.

Table 8. Detection results on the FLIR benchmark at different image resolutions.

Input Size	Method	mAP (%)
384×384	Baseline	53.9
	SMG-C	59.0
512×512	baseline	62.7
	SMG-C	68.8

5.2. Guidance with Hard or Soft Label

In SMG, we use the foreground soft label generated from the source model as the guidance. However, the hard label from the ground truth also can be utilized as the guidance. The hard label is the ground-truth foreground score, which is the combination of all ground-truth targets mapped to the heatmap. In every position of heatmap, the value of the hard label is either 0 or 1, which is different from the soft label in $[0, 1]$.

We fix the image resolution at 512×512 and compare the baseline (no guidance) with three different guidance methods, including hard, soft, and both of them in Table 9. The methods with guidance surpass the baseline more than 5% in mAP, which shows that the guidance is an important factor in performance improvement. Furthermore, the soft guidance obtains higher accuracy than other guidance methods. We attribute it to the fact that the soft label contains hidden information about how the source model distinguishes foreground from background, which is exactly what the target network needs to learn. Therefore, we choose the soft guidance in SMG other than hard guidance.

Table 9. Detection results of different guidance methods on the FLIR benchmark.

Guidance Method	mAP (%)
No guidance (baseline)	62.7
Hard	67.7
Hard and soft	68.1
Soft	68.8

5.3. Effect of Backbone

In this subsection, two different backbones, ResNet-18 [19] and DLA-34 [38], are used as FENs in the target networks. Table 10 shows the comparison of the their mAP with corresponding baselines at the image resolution of 512 × 512. The structure of DLA-34 is more complicated than ResNet-18, and thus, higher detection accuracy can be achieved. In spite of different backbones, we observe a significant increase in mAP (over 5%) when SMG is added to the framework. That indicates SMG is an effective strategy no matter which backbone we employ.

Table 10. Detection results of SMG-C with different FENs on the FLIR benchmark.

Backbone	Method	mAP (%)
ResNet-18	Baseline	62.7
	SMG-C	68.8
DLA-34	Baseline	71.1
	SMG-C	75.6

6. Discussions

In this section, we give some insights about why our proposed SMG works well when there are limited training examples. Then, we analyze the failure cases of our methods.

6.1. Why SMG Works Well

In SMG, we suppress the background disturbances by borrowing the knowledge from the source model so as to reduce the data dependency of the target network (infrared detection network). Taking SMG-C as an example, we visualize the soft label generated from the source model and the heatmap of the target network. Figure 10 shows that the source model can filter out the main background, such as roads, houses, and so on. However, it hardly detects specific targets in heavy occlusion, such as people in the crowd, cyclists, and bicycles. In other words, the soft label from the source model can be viewed as effective knowledge to provide supervision, but it cannot be leveraged directly. We solve this problem by inserting a BSM in the target network to receive the knowledge transferred from the source model and enhance the foreground at the same time. The last column in Figure 10 illustrates that the target network with BSM locates center points of targets more precisely than the source model. As a result, the target network can pay more attention to target objects, which is important for training with limited examples.

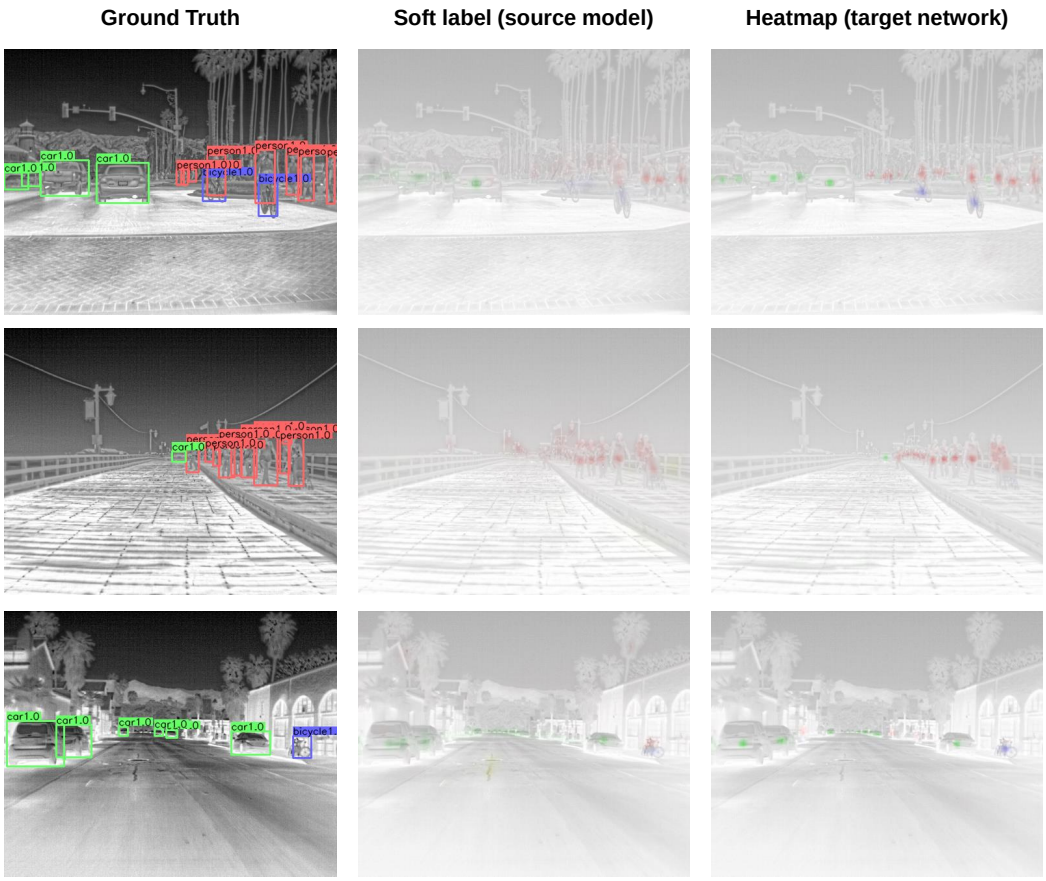

Figure 10. The visualization of soft label and heatmap.

6.2. Missed Detections

Although SMG promotes accuracy in infrared object detection, the limited-examples detection is still a challenging task. By visualizing the results of SMG-Y trained on FLIR-1/4 and full FLIR in Figure 11, we study the missed detections in absence of training examples. We also represent logarithmic average miss rates of SMG-Y and SMG-Y-1/4 in Table 11. The miss rates of SMG-Y-1/4 are slightly higher than those of SMG-Y. When two objects are close to each other, such as two pedestrians walking together, SMG-Y-1/4 may detect them as a single target, while SMG-Y with sufficient training data easily distinguishes them, as shown in Figure 11. Furthermore, we find that both SMG-Y and SMG-Y-1/4 miss the small objects located far from the camera or obscured by others, such as person and bicycle. We attribute this drawback to the fact that their source model YOLOv3 has poor detection performance for small targets. In the future, we will focus on these challenges and try to cope with them.

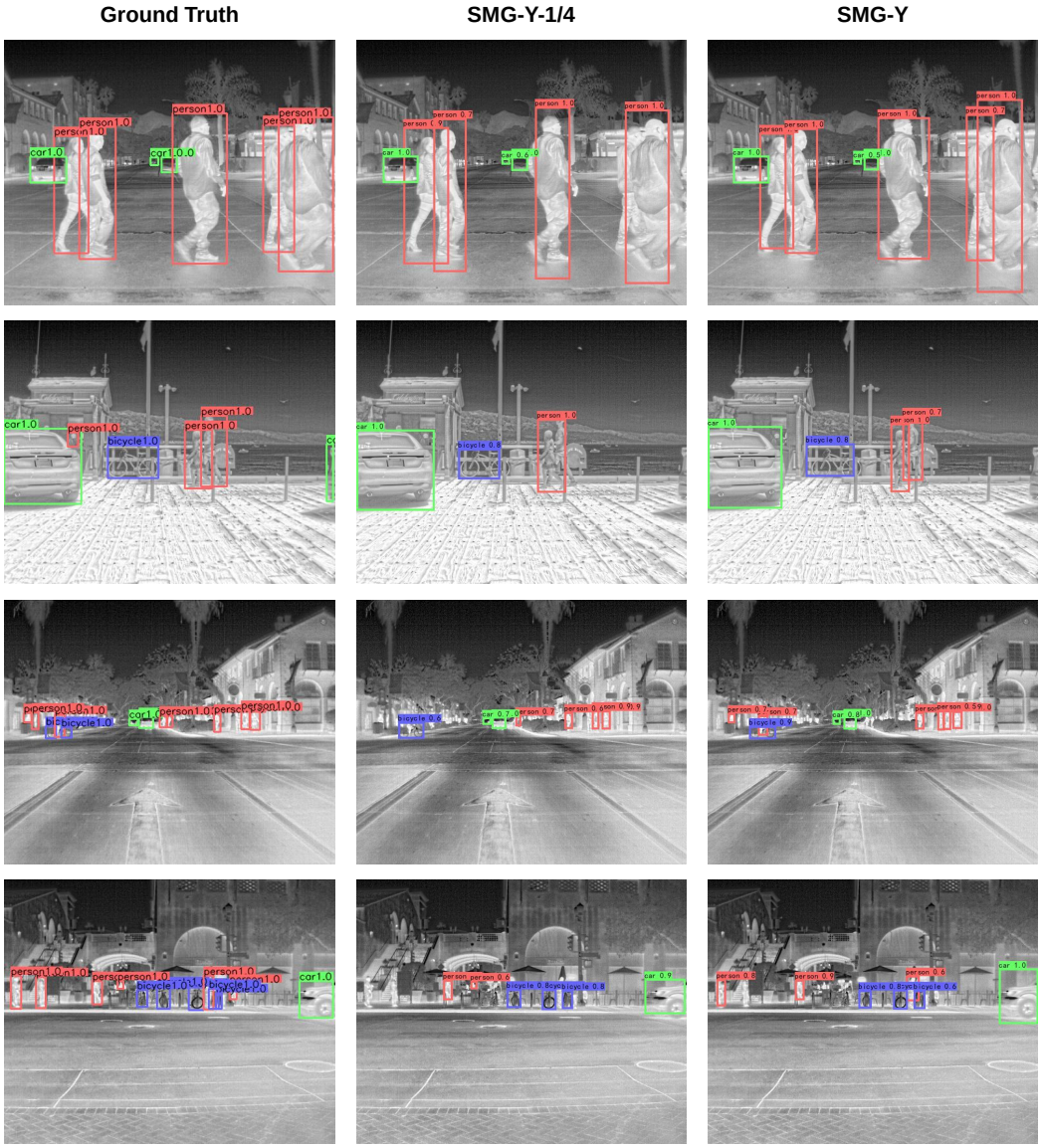

Figure 11. Some examples of missed detections. Note that SMG-Y-1/4 represents SMG-Y trained on FLIR-1/4.

Table 11. Miss rates of SMG-Y and SMG-Y-1/4 on the FLIR benchmark.

Method	Person	Car	Bicycle
SMG-Y	0.53	0.41	0.52
SMG-Y-1/4	0.55	0.43	0.55

7. Conclusions

In summary, we present a novel cross-domain transfer approach SMG to address the problem of infrared detection on small-scale datasets. SMG can convert a visual detection framework into an infrared detector by borrowing the knowledge from the source model, which is a trained RGB detection model. We apply SMG in both anchor-free and anchor-based detection frameworks, named as SMG-C and SMG-Y, respectively. Experiments on FLIR and IAT illustrate that our infrared detectors achieve outstanding performance in lack of available training data. Compared to state-of-the-art detectors, SMG-Y with only 1/4 training data outperforms most of them, demonstrating that SMG is a preferable method for limited-examples infrared detection.

Author Contributions: All of the authors contributed to this study. Conceptualization, R.C. and S.L.; methodology, R.C.; software, R.C.; data curation, J.M. and Z.M.; writing—original draft preparation, R.C.; writing—review and editing, R.C., J.M. and Z.M.; funding acquisition, S.L. and F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shanghai Key Laboratory of Criminal Scene Evidence funded Foundation (Grant No. 2017xcwzk08) and the Innovation Fund of Shanghai In-stitute of Technical Physics (Grant No. CX-321).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Devaguptapu, C.; Akolekar, N.; Sharma, M.M.; Balasubramanian, V.N. Borrow From Anywhere: Pseudo Multi-Modal Object Detection in Thermal Imagery. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 1029–1038. [CrossRef]
2. Zhang, L.; Peng, Z. Infrared Small Target Detection Based on Partial Sum of the Tensor Nuclear Norm. *Remote Sens.* **2019**, *11*, 382. [CrossRef]
3. Rashid, M.; Khan, M.A.; Alhaisoni, M.; Wang, S.H.; Naqvi, S.R.; Rehman, A.; Saba, T. A Sustainable Deep Learning Framework for Object Recognition Using Multi-Layers Deep Features Fusion and Selection. *Sustainability* **2020**, *12*, 5037. [CrossRef]
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
5. Masood, H.; Zafar, A.; Ali, M.U.; Hussain, T.; Khan, M.A.; Tariq, U.; Damaševičius, R. Tracking of a Fixed-Shape Moving Object Based on the Gradient Descent Method. *Sensors* **2022**, *22*, 1098. [CrossRef]
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
7. Hussain, N.; Khan, M.; Kadry, S.; Tariq, U.; Mostafa, R.; Choi, J.; Nam, Y. Intelligent Deep Learning and Improved Whale Optimization Algorithm based Framework for Object Recognition. *Hum.-Centric Comput. Inf. Sci.* **2021**, *11*, 1–18.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
9. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
10. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
11. Chen, R.; Liu, S.; Miao, Z.; Li, F. Infrared aircraft few-shot classification method based on meta learning. *Infrared Millim. Waves* **2021**, *40*, 554–560. [CrossRef]
12. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
13. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland 2014; pp. 740–755.
14. Zhou, T.; Yu, Z.; Cao, Y.; Bai, H.; Su, Y. Study on an infrared multi-target detection method based on the pseudo-two-stage model. *Infrared Phys. Technol.* **2021**, *118*, 103883. [CrossRef]
15. Miao, Z.; Zhang, Y.; Li, W.H. Real-time infrared target detection based on center points. *Infrared Millim. Waves* **2021**, *40*. [CrossRef]
16. Konig, D.; Adam, M.; Jarvers, C.; Layher, G.; Neumann, H.; Teutsch, M. Fully Convolutional Region Proposal Networks for Multispectral Person Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017.
17. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. *arXiv* **2016**, arXiv:1611.02644.
18. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
19. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
20. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 765–781.
21. Zhou, X.; Zhuo, J.; Krahenbuhl, P. Bottom-Up Object Detection by Grouping Extreme and Center Points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
22. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
23. Teledyne FLIR. Flir Thermal Dataset for Algorithm Training [DB/OL]. FLIR. 1 September 2018. Available online: https://www.flir.com/oem/adas/adas-dataset-form/ (accessed on 7 January 2022).

24. Hosang, J.; Benenson, R.; Schiele, B. Learning Non-maximum Suppression. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6469–6477. [CrossRef]
25. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
26. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep Mutual Learning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4320–4328. [CrossRef]
27. Huang, Z.; Pan, Z.; Lei, B. Transfer Learning with Deep Convolutional Neural Network for SAR Target Classification with Limited Labeled Data. *Remote Sens.* **2017**, *9*, 907. [CrossRef]
28. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
29. Yim, J.; Joo, D.; Bae, J.; Kim, J. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7130–7138. [CrossRef]
30. Li, Q.; Jin, S.; Yan, J. Mimicking Very Efficient Network for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7341–7349. [CrossRef]
31. Chen, G.; Choi, W.; Yu, X.; Han, T.; Chandraker, M. Learning efficient object detection models with knowledge distillation. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 742–751.
32. Chen, H.; Wang, Y.; Wang, G.; Qiao, Y. LSTD: A Low-Shot Transfer Detector for Object Detection. Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
34. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539. [CrossRef]
35. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–19. [CrossRef]
36. Zhang, Q.L.; Yang, Y.B. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239. [CrossRef]
37. Wei, D.; Du, Y.; Du, L.; Li, L. Target Detection Network for SAR Images Based on Semi-Supervised Learning and Attention Mechanism. *Remote Sens.* **2021**, *13*, 2686. [CrossRef]
38. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep Layer Aggregation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2403–2412. [CrossRef]
39. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-Shot Refinement Neural Network for Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212. [CrossRef]
40. Qin, P.; Tang, C.; Liu, Y.; Zhang, J.; Xu, Z. Infrared target detection method based on improved YOLOv3. *Comput. Eng.* **2021**, 1–12. [CrossRef]