激光与光电子学进展 先进成像

基于关键点的半监督红外图像目标检测算法

沈一选，金韬*，但俊
浙江大学信息与电子工程学院，浙江 杭州 310027

摘要 为提高红外图像中目标检测的精度，提出一种基于CenterNet与OMix增强的半监督红外图像目标检测算法(IRCC-OMix)。针对红外图像中锚框先验信息难以确定的问题，利用CenterNet作为主干模型，通过关键点检测红外图像中的目标。由于红外图像标注成本昂贵，引入基于教师学生网络互学习的半监督学习方法，设计基于CenterNet与基于一致性的半监督红外图像目标检测(IRCC)模型。IRCC模型中的随机擦除(cutout)增强可能导致红外图像中的小目标消失，影响模型检测性能，因此采用一种基于目标的图像混合增强方法，提升算法对小目标的检测能力。在公开数据集FLIR上的实验结果表明，IRCC模型的平均精度均值(mAP)达到55.3%，与仅使用有标签数据训练情况相比，mAP提升1.9个百分点，说明该模型能够充分利用无标签数据、提高模型的鲁棒性。基于OMix增强的IRCC模型的mAP为56.8%，与使用cutout增强的IRCC模型相比提高1.5个百分点，取得了良好的检测性能。

关键词 图像处理；目标检测；卷积神经网络；红外图像；半监督学习
中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP221605

Semi-Supervised Infrared Image Target Detection Algorithm Based on Key Points

Shen Yixuan, Jin Tao*, Dan Jun
College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, *Zhejiang, China*

Abstract A semi-supervised target detection algorithm for infrared images based on CenterNet and OMix enhancement (IRCC-OMix) is proposed to improve target detection accuracy. The prior information of the anchor frame in the infrared image is difficult to determine. Therefore, CenterNet is used as the backbone model to detect the target in the infrared image through key points. A semi-supervised learning method based on teacher-student-network mutual learning is introduced owing to the high cost of infrared image annotation, and a semi-supervised infrared image target detection (IRCC) model based on CenterNet and consistency is designed. The random erasure enhancement in the IRCC model may lead to the disappearance of small targets in the infrared image, which affects the detection performance of the model. Therefore, an object-based image mixing enhancement method is adopted to improve the detection ability of the algorithm for small targets. The experimental results on the public dataset, FLIR, show that the average precision mean (mAP) of the IRCC model reaches 55.3%, which is 1.9 percentage points higher than that of the training using only labeled data. This indicates that the model can fully utilize unlabeled data and improve its robustness. The mAP of the OMix-enhanced IRCC model is 56.8%, which is 1.5 percentage points higher than that of the cutout-enhanced IRCC model and achieves good detection performance.

Key words image processing; object detection; convolutional neural network; infrared image; semi-supervised learning

1 引 言

随着红外技术与计算机视觉的快速发展，基于红外图像的目标检测技术在许多领域得到广泛应用，如消防救援、安防监控和自动驾驶等场景[1]。由于红外图像用途广泛，在不同应用场景下会产生大量的红外图像数据，如仅使用红外热成像仪全天候对街道进行安防监测，就会产生海量的红外图像[2]。对于大量质

收稿日期：2022-05-16；**修回日期**：2022-07-15；**录用日期**：2022-08-04；**网络首发日期**：2022-08-14
基金项目：国家自然科学基金(61675180)
通信作者：*jint@zju.edu.cn

量参差不齐的红外图像，仅仅依靠人工对所有图像进行标注需要昂贵的人力成本。半监督学习能够通过大量的无标签数据提升模型性能，因此，研究基于半监督学习的红外图像目标检测技术具有重要的意义。

近年来，在深度学习理论的支持下，各种基于深度学习的目标检测算法层出不穷，其主要分为两大类：基于锚框的目标检测算法和基于关键点的目标检测算法。基于锚框的目标检测算法近年得到广泛研究，其中最具代表性的是基于区域的两阶段算法 Faster R-CNN[3]和基于回归思想的单阶段 YOLO 系列算法[4]。而基于关键点的目标检测算法，如 CenterNet[5]将一个中心点视作单个形状不可知的锚点，通过回归获取目标的中心与大小。基于锚框的检测算法通常需要构建大量的锚框来确定目标在图像中的位置，算法性能的提升严重依赖于锚框形状、大小等先验信息[6]。与可见光图像相比，红外图像中存在更多的仅占有部分像素的小目标，锚框的信息更加难以确定。因此，引入基于关键点的目标检测模型 CenterNet 作为主干模型进行学习，不依赖红外图像的锚框信息，具有更高的准确率和泛化性能。

半监督学习使用大规模的无标签数据来提高模型性能，其主要思想是一致性正则化，伪标签在训练过程中起到关键作用。教师学生网络互学习（TSML）的学习框架作为当前半监督学习的研究热点，不同于半监督学习中传统伪标签的训练方法，教师网络可以作为学生网络的时间集成模型，通过产生伪标签对学生网络进行监督训练[7]。将 TSML 的半监督学习框架引入红外图像目标检测能够充分利用无标签数据，显著提高模型性能。图像数据增强在半监督学习中扮演着重要角色，最近的许多研究[8-10]都已证明数据增强在半监督学习中的有效性。基于 TSML 的半监督目标检测中，弱增强图像作为教师网络的输入用于产生可靠的伪标签，而学生网络使用强增强图像提升模型的鲁棒性能。图像增强的方法主要包括简单的仿射变换（旋转、平移）和复杂的插值变换[随机擦除（cutout）[11]、图像混合（mixup）[12]等]。cutout 操作通过随机擦除一块区域的像素点提升模型的鲁棒性，然而在红外图像中存在大量的小目标，cutout 操作可能导致小目标[13]被意外删除，使模型对小目标的检测性能下降。为解决这个问题，提出一种基于目标的图像混合增强（OMix）方法对红外图像进行增强，能够充分利用红外图像中小目标的像素点，提升模型对小目标的检测能力。

针对红外图像目标检测的难点，本文提出一种基于 CenterNet 与 OMix 增强的半监督红外图像目标检测（IRCC-OMix）算法。CenterNet 作为一种基于关键点的目标检测模型，利用中心点的特征信息进行目标分类与边界回归，在彩色图像目标检测中有较好的实时性与检测精度[5]，将其作为主干模型对红外图像进行学习，可以降低模型对锚框信息的依赖，具有较好的鲁棒性。为了通过无标签数据提升模型性能，引入基于 TSML 的半监督学习框架，设计基于 CenterNet 与基于一致性的半监督学习的红外图像目标检测模型（IRCC）。针对半监督学习中 cutout 增强操作产生的小目标消失问题，受到 mixup 增强[12]的启发，使用 OMix 增强方法，通过 mixup 方式减少像素点丢失、增强模型对小目标的检测性能。所提模型利用无标签图像显著提升红外图像中目标检测的精度，尤其是对较小目标的检测能力，在消防救援、安防监控等场景下有较高的应用价值。

2　基于 CenterNet 的红外图像目标检测模型

2.1　全监督目标检测

CenterNet 是一种简单有效的基于关键点的目标检测框架，其模型主要分为两个部分：特征图提取网络和目标检测网络[5]，如图 1 所示。特征图提取网络是一种基于编码解码（encoder-decoder）的特征提取结构[14]。红外图像首先通过高效的特征提取网络（如 ResNet[15]、Hourglass[16-17]）获取图像的高级语义信息，之后对其进行反卷积上采样得到高分辨率特征图。目标检测网络利用高分辨率特征图生成概率热力图、目标大小以及中心点偏移。选取热力图的峰值作为目标

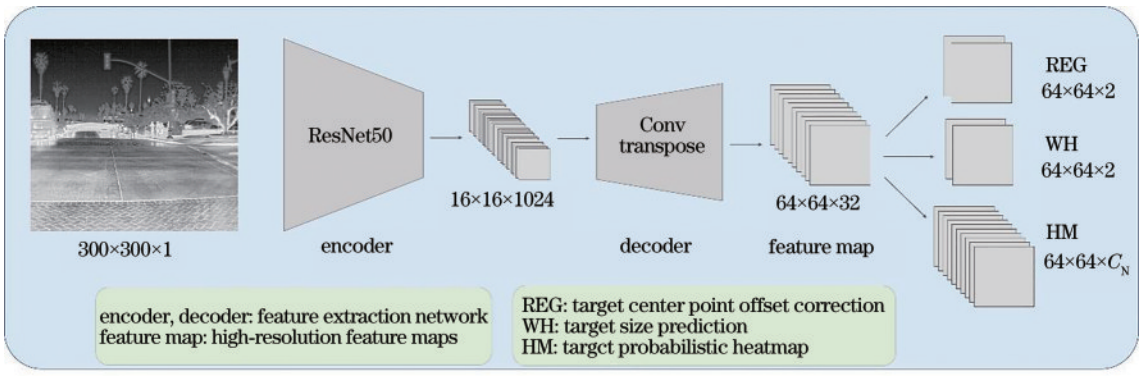

图 1　基于 CenterNet 的红外图像目标检测模型

Fig. 1　Infrared image target detection model based on CenterNet

中心，结合对应点的其他特征确定目标在原图中的大小及位置。训练和推断过程只需简单提取每个目标对象的对应中心点，不需要非极大抑制或其他后处理。

假设输入有标签红外图像的尺寸为 $W\times H$，W、H 分别为输入图像的宽和高。其通过高分辨率特征图生成的热力图为 $\hat{\boldsymbol{Y}}\in[0,1]^{\frac{W}{R}\times\frac{H}{R}\times C_{\mathrm{N}}}$，其中，$R$ 是输出的步长，C_{N} 是关键点种类的个数。假设一个种类 c 在热力图 $\{x,y\}$ 坐标的预测点为 $\hat{Y}_{xyc}$，那么 $\hat{Y}_{xyc}=1$ 意味着该关键点为目标点，否则该点为背景点。令 N 为特征图中关键点的数目，$\boldsymbol{p}$ 为原图中的一个点，则 $\tilde{\boldsymbol{p}}=\left\lfloor\frac{\boldsymbol{p}}{R}\right\rfloor$ 为点 $\boldsymbol{p}$ 在特征图中的位置。对于有标签的数据，Y_{xyc} 代表利用目标中心位置和尺寸通过高斯核在每个种类热力图上生成的标签，则分类损失 L_{T} 可以定义为

$$L_{\mathrm{T}}=\frac{-1}{N}\sum_{xyc}\begin{cases}\left(1-\hat{Y}_{xyc}\right)^{\alpha_c}\log\left(\hat{Y}_{xyc}\right), & Y_{xyc}==1\\ \left(1-Y_{xyc}\right)^{\beta_c}\left(\hat{Y}_{xyc}\right)^{\alpha_c}\log\left(1-\hat{Y}_{xyc}\right), & \text{else}\end{cases},\tag{1}$$

式中：α_c 和 β_c 是调整难易样本分类权重的超参数[18]。

为了解决特征图中由于不同步长 R 导致的位置偏差，通过在高分辨率特征图中的每个像素点引入局部偏差 $\hat{\boldsymbol{O}}\in\mathbf{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$ 进行修正，位置偏差修正损失函数定义为

$$L_{\mathrm{off}}=\frac{1}{N}\sum_{\boldsymbol{p}}\left|\hat{\boldsymbol{O}}_{\tilde{\boldsymbol{p}}}-\left(\frac{\boldsymbol{p}}{R}-\tilde{\boldsymbol{p}}\right)\right|。\tag{2}$$

此外，对于种类 c 的第 k 个目标 c_k，N_c 为种类 c 的关键点个数，假设其中心点为 $\boldsymbol{p}_{c_k}$，则目标框的大小为 $\boldsymbol{s}_{c_k}=\left(w_{c_k},h_{c_k}\right)$。通过回归得到所有关键点的目标框为 $\hat{\boldsymbol{s}}\in\mathbf{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$，则目标框尺寸的回归损失 L_{size} 可以表示为

$$L_{\mathrm{size}}=\frac{1}{N}\sum_{c=1}^{C_{\mathrm{N}}}\sum_{k=1}^{N_c}\left|\hat{\boldsymbol{s}}_{c_k}-\boldsymbol{s}_{c_k}\right|。\tag{3}$$

综上所述，基于 CenterNet 全监督的红外图像目标检测的损失函数 L_{C} 如下：

$$L_{\mathrm{C}}=L_{\mathrm{T}}+\lambda_{\mathrm{size}}L_{\mathrm{size}}+\lambda_{\mathrm{off}}L_{\mathrm{off}},\tag{4}$$

式中：λ_{size} 和 λ_{off} 分别为位置偏差修正损失函数和目标框大小回归损失函数的平衡系数。

2.2　半监督目标检测

半监督目标检测主要思想是一致性正则化，基于 TSML 的半监督模型能够对学生网络产生有效的监督信号[7]。该模型框架主要分为燃烧期和网络互学习两个阶段，如图 2 所示。在燃烧期阶段，利用现有的有标签红外图像获得一个初始化的目标检测模型，之后对初始模型参数进行复制得到教师网络和学生网络。在网络互学习阶段，教师网络使用生成的伪标签对学生网络进行监督训练，同时学生网络根据从伪标签学习到的知识利用组合指数移动平均值（EMA）更新教师网络的参数。根据半监督一致性[19]，教师模型使用弱增强图像作为输入可以获得更加可靠的伪标签，而学生模型采用强增强图像作为输入通过反向传播更新网络权重。从集成模型角度出发，教师网络可以视为多个不同时间学生子模型的时间集成模型，教师网络的准确率总是优于学生网络，使用 TSML 机制可以获得更好的半监督学习效果。因此，设计了 IRCC 模型，其能够充分利用无标签红外图像提升目标检测的准确率。

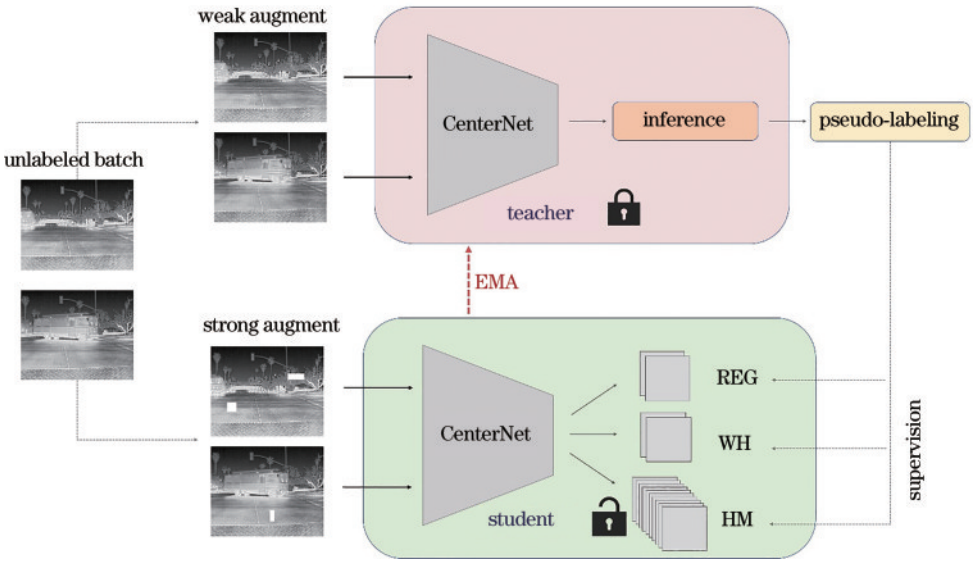

图 2　IRCC 无标签图像学习过程

Fig. 2　Unlabeled infrared image learning process for IRCC

在燃烧期阶段，利用有标签数据获得可靠的初始化模型。该阶段损失函数 L_{Sup} 与全监督 CenternNet 的损失函数相同，即 $L_{\mathrm{Sup}}=L_{\mathrm{C}}$。为了充分利用无标签数据，在 TSML 阶段通过迭代训练提升模型性能。在该

阶段，首先将初始化的模型复制产生教师模型（模型参数为 θ_t）和学生模型（模型参数为 θ_s）。与半监督分类任务相同，先利用置信度 δ 作为教师网络的阈值产生当前图片检测目标的伪标签[19]。假设通过教师模型生成的伪标签锚框 $\boldsymbol{p}_u \in \mathbf{R}^2$，根据其在低分辨率特征中的位置 $\tilde{\boldsymbol{p}}_u = \left\lfloor \frac{p_u}{R} \right\rfloor$，利用高斯核生成对应关键点的伪标签，定义为 Y^u_{xyc}：

$$Y^u_{xyc} = \exp\left[\frac{\left(x-\tilde{p}^u_x\right)^2+\left(y-\tilde{p}^u_y\right)^2}{2\sigma_p^2}\right]。\tag{5}$$

则 IRCC 模型中无监督损失函数 L_{Unsup} 可表示为

$$L_{\text{Unsup}} = L^u_T + \lambda^u_{\text{size}} L^u_{\text{size}} + \lambda^u_{\text{off}} L^u_{\text{off}}。\tag{6}$$

根据式（6）中的无监督损失函数，IRCC 的损失函数 L_{IRCC} 可以定义为

$$L_{\text{IRCC}} = L_{\text{Sup}} + \lambda_u L_{\text{Unsup}},\tag{7}$$

式中：λ_u 为无监督损失函数的平衡系数。

学生模型中的参数 θ_s 可以利用式（7）通过反向传播进行更新：

$$\theta_s \leftarrow \theta_s + \gamma_s \frac{\partial\left(L_{\text{IRCC}}\right)}{\partial\theta_s},\tag{8}$$

式中：γ_s 代表学生模型的学习率。

最后，为得到稳定可靠的伪标签，采用 EMA 方法更新教师模型的参数 θ_t[13]。教师模型是连续学生模型参数的权重平均值，也可被视为训练迭代过程中学生模型时间的集成模型。在每次迭代中，教师网络的参数通过 EMA 从学生网络中缓慢更新。迭代过程可以表示为

$$\theta_t^{i+1} \leftarrow \alpha_t \theta_t^i + (1-\alpha_t)\theta_s,\tag{9}$$

式中：θ_t^i 和 α_t 分别为第 i 次迭代教师网络的参数和 EMA 的系数。

3　基于 OMix 增强的半监督目标检测

3.1　算法简述

TSML 框架中，教师模型使用弱增强图像作为输入获得可靠的伪标签，学生网络采用强增强图像作为输入将学习到的知识通过 EMA 反馈给教师模型。由于红外图像中存在大量的小目标，这些小目标会因为强图像增强过程中的 cutout 操作失去所有特征点，会导致教师学生模型一致性退化、降低模型性能[13]。mixup[12]是一种通过混合多个样本分布进行增强的方式，两个训练样本的线性组合应该随其混合的权重线性变化，以此获得更加平滑的决策边界，提升模型的鲁棒性。基于图像混合的思想，采用 OMix 增强方法替代 cutout 对教师模型中输入图像进行增强，以此增强目标检测模型对红外图像中小目标识别的能力。

OMix 增强采用两张无标签图像 $\{I_a, I_b\}$ 作为一组进行训练，如图 3 所示。与基于 CenterNet 的一致性半监督目标检测模型相同，先利用弱增强图像 $\{w(I_a), w(I_b)\}$ 通过教师网络获得当前两张无标签图像的伪标签 $\{p^a_u, p^b_u\}$。根据伪标签中目标锚框的位置，可以生成两张图片对应前景和背景的 OMix 掩码 $\{m(I_a), m(I_b)\}$：

$$m(I) = \begin{cases} 1, & p_u > \delta \\ 0, & \text{else} \end{cases},\tag{10}$$

式中：δ 为生成伪标签的阈值。

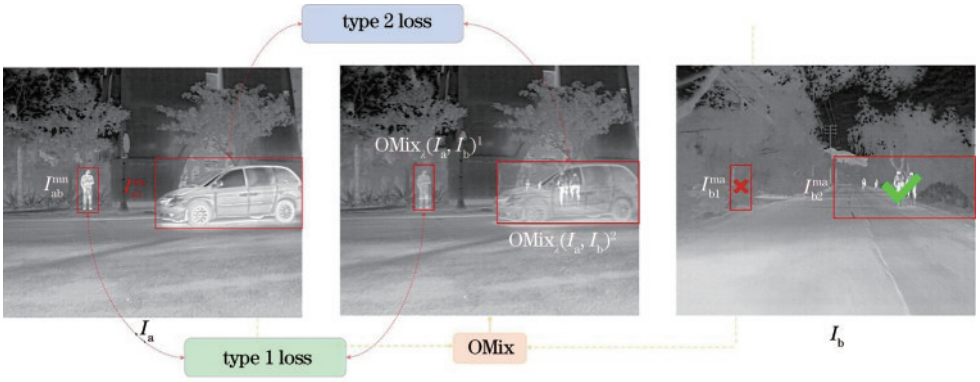

图 3　基于 OMix 增强的半监督目标检测模型

Fig. 3　Semi-supervised infrared image object detection model based on OMix

OMix 掩码 $m(I)=1$ 的区域代表该区域为有效的目标候选框区域，增强操作需要将有效的目标候选框区域与同一组另一张图片中的对应区域进行叠加混合。如对图片 I_a 进行 OMix 增强，则需要将第 k 个有效的前景候选框区域 $I^{ma}_{ak} = I_a m(I_a)$ 与同组中另一张图片 I_b 中对应的 $I^{ma}_{bk} = I_b m(I_a)$ 进行混合，图片 I_a 的混合过程 $\text{OMix}_\lambda(I_a, I_b)$ 和图片 I_b 的混合过程 $\text{OMix}_\lambda(I_b, I_a)$ 分别可描述为

$$\begin{cases} \text{OMix}_\lambda(I_a, I_b) = \sum_{k=1}^{N_a} \lambda I^{ma}_{ak} + (1-\lambda) I^{ma}_{bk} \\ \text{OMix}_\lambda(I_b, I_a) = \sum_{k=1}^{N_b} \lambda I^{mb}_{bk} + (1-\lambda) I^{mb}_{ak} \end{cases},\tag{11}$$

式中：N_a和N_b分别代表I_a和I_b中所有有效前景候选区域的个数；λ为混合系数。与mixup增强相同，OMix的混合系数λ从分布Beta(α,α)中采样而来。

根据教师网络产生的伪标签，OMix增强过程中将会产生不同混合类型[20]，不同OMix混合类型采用不同的损失函数，OMix增强的类型根据式(12)进行分类：

$$\begin{cases}\text{type 1: M1}(I_a)=m\left(I_{ak}^{ma}\right)\otimes\sim m\left(I_{bk}^{ma}\right)\\ \text{type 2: M2}(I_a)=m\left(I_{ak}^{ma}\right)\otimes m\left(I_{bk}^{ma}\right)\end{cases},\tag{12}$$

式中：$\otimes$为哈达玛积；$\sim$为取反运算符。以对图3中图片I_a的操作$\text{OMix}_\lambda(I_a,I_b)$为例。第1类掩码$\text{M1}(I_a)$中，$m\left(I_{a1}^{ma}\right)\otimes\sim m\left(I_{b1}^{ma}\right)$代表图片$I_a$中第1个目标区域$I_{a1}^{ma}$中存在检测的目标(行人)，而图片$I_b$中的对应区域$I_{b1}^{ma}$为背景。第2类掩码$\text{M2}(I_a)$中，$m\left(I_{a2}^{ma}\right)\otimes m\left(I_{b2}^{ma}\right)$意味着图片$I_a$中区域$I_{a2}^{ma}$为前景(汽车)，对应图片$I_b$中区域$I_{b2}^{ma}$也大概率也存在检测的目标。针对OMix增强过程中产生的不同混合类型，设计两种不同的损失函数对其进行处理，以增强模型的检测性能。

3.2 损失函数分类

在OMix增强操作的第1类损失函数中，在高置信度的情况下，一张图片的指定区域为前景，而另一张图片的对应区域为背景。这种情况可以视为FixMatch损失函数[19]的推广形式，没有经过OMix操作的第k个区域分类的结果作为模型学习的目标$f_{\text{heat(s)}}\left(I_a^k\right)$(以$I_a$为前景为例)，而使用OMix增强的输出s

$$\begin{cases}l_1^a=\text{KL}\left\{f_{\text{heat(s)}}\left(I_a^k\right)\middle\|f_{\text{heat(s)}}\left[\text{OMix}_\lambda(I_a,I_b)^k\right]\right\}\\ l_1^b=\text{KL}\left\{f_{\text{heat(s)}}\left(I_a^k\right)\middle\|f_{\text{heat(s)}}\left[\text{OMix}_\lambda(I_b,I_a)^k\right]\right\}\\ l_1=\left(l_1^a+l_1^b\right)\end{cases}。\tag{13}$$

式中：l_1^a、l_1^b分别为图I_a、I_b的第1种混合类型的损失函数；$\text{KL}\{\cdot\}$代表相对熵；$f_{\text{heat(s)}}$为学生网络热力图的特征。考虑到图像中的所有区域，第1种混合类型损失函数的期望$L_1=E\left[l_1\right]$。

当两张图片在指定区域均为有高置信度的前景目标，并且其交并比(IoU)大于指定的阈值λ_{IoU}时，在OMix操作中被视为第2种混合类型。在不考虑一组图片中指定区域目标对应的大小与位置时，第2种混合类型的损失函数可以使用JS散度进行度量。JS散度有更好的度量性能，因为其对于不同目标种类(如前景、背景)有不同的权重计算。第2种混合类型的损失函数可以表示为

$$l_2^a=\text{JS}\left\{\text{OMix}_\lambda\left[f_{\text{heat(s)}}\left(I_a^k\right),f_{\text{heat(s)}}\left(I_b^k\right)\right]\middle\|f_{\text{heat(s)}}\left[\text{OMix}_\lambda(I_a,I_b)^k\right]\right\},\tag{14}$$

式中：$f_{\text{heat(s)}}\left[\text{OMix}_\lambda(I_a,I_b)^k\right]$表示经过OMix增强区域$k$通过学生网络产生的推断结果；$\text{OMix}_\lambda\left[f_{\text{heat(s)}}\left(I_a^k\right),f_{\text{heat(s)}}\left(I_b^k\right)\right]$表示两张原始图片$I_a$和$I_b$对应区域生成的混合伪标签。第2种混合类型的损失函数即为两者之间JS散度的距离。

考虑到两张图像混合时，一般不同目标的位置和大小不可能完全相同。第1种情况下，当一张图像中指定区域为一个小目标时，同组另一张图像的相应位置可能存在一个较大的目标。这样虽然会导致混入目标区域为前景目标，但实际对于该小目标而言，并不是混入了一个完全的前景目标，其产生的效果与混入了一部分背景相同。第2种情况，尽管两个区域中有相同大小的目标，但由于位置不同，也将对混入目标的一致性损失函数产生影响。为此，引入指定区域中两个目标的IoU作为一致性的额外参数进行约束。引入IoU的第2种混合类型可以定义为

$$l_2^a=\text{JS}\left\{\text{OMix}_{\frac{\lambda}{\lambda_{\text{IoU}}}}\left[f_{\text{heat(s)}}\left(I_a^k\right),f_{\text{heat(s)}}\left(I_b^k\right)\right]\middle\|f_{\text{heat(s)}}\left[\text{OMix}_\lambda(I_a,I_b)^k\right]\right\},\tag{15}$$

$$\lambda_{\text{IoU}}=\text{IoU}\left[\text{obj}\left(I_a^k\right),\text{obj}\left(I_b^k\right)\right],\lambda\leqslant\lambda_{\text{IoU}}<1,\tag{16}$$

式中：λ_{IoU}表示图片I_a中第k个区域中目标与图片I_b对应区域中目标锚框的交并比。当两张前景图片中目标完全重合($\lambda_{\text{IoU}}=1$)时，退化为式(14)中的JS散度。I_a^k为小目标、I_b^k为大目标时，尽管图片混合的对应区域存在检测目标，但由于两个目标的IoU较小，此时I_b^k中目标的特征将会弱化。同一组训练图片中两个目标在同一区域中处于不同位置时，较小的IoU也会因位置偏差弱化混入I_b^k中的特征。考虑图片I_a中所有目标区域，则其第2种类型损失函数$L_2^a=E\left[l_2^a\right]$。在该组训练样本中另一张图片$I_b$的损失函数$l_2^b$也可以表示为

$$\begin{cases}l_2^b=\text{JS}\left\{\text{OMix}_{\frac{\lambda}{\lambda_{\text{IoU}}}}\left[f_{\text{heat(s)}}\left(I_b^k\right),f_{\text{heat(s)}}\left(I_a^k\right)\right]\middle\|f_{\text{heat(s)}}\left[\text{OMix}_\lambda(I_b,I_a)^k\right]\right\}\\ \lambda_{\text{IoU}}=\text{IoU}\left[\text{obj}\left(I_b^k\right),\text{obj}\left(I_a^k\right)\right],\lambda\leqslant\lambda_{\text{IoU}}<1\end{cases}。\tag{17}$$

图I_b的第2种类型损失函数$L_2^b=E\left[l_2^b\right]$。则OMix操作第2种混合类型的损失函数$L_2=L_2^a+L_2^b$。

结合两种不同混合类型的损失函数，OMix增强操作的损失函数定义为

$$L_{\text{OMix}}=\gamma_{o1}L_1+\gamma_{o2}L_2,\tag{18}$$

式中：γ_{o1}和γ_{o2}为OMix操作中两类损失函数的平衡

系数。

综上所述，IRCC-OMix模型的损失函数为

$$L_{\text{OMix-IRCC}}=L_{\text{Sup}}+\lambda_{\text{u}}\left(L_{\text{Unsup}}+L_{\text{OMix}}\right)。\tag{19}$$

4 实验验证

4.1 实验环境设置

为验证上述基于CenterNet的一致性半监督目标检测模型与OMix增强方法的效果，基于PyTorch深度学习框架使用Intel(R) Xeon(R) Gold 5222 CPU、64 GB memory、NVIDIA Quadro RTX 4000 GPU的计算机对模型进行训练和测试。实验数据集采用开源红外图像数据集FLIR，FLIR红外图像数据集由10228张红外图像组成，其中用于训练的红外图像有8862张，用于测试的有1366张。图像中的标签包含汽车、行人、自行车和其他车辆等，采用其中最主要的汽车(car)、行人(person)和自行车(bicycle)作为实验识别与检测的目标。与其他半监督目标检测的任务相同[20]，将所有用于训练的红外图像随机分为有标签的训练数据集(train-3k，3000张图像)和无标签的训练数据集(train-5.8k，5862张图像)，并且保证所有用于检测的目标都会出现在有标签的训练数据集。最后将训练好的模型在测试集(test)中进行测试，获得评估模型参数。采用平均精度均值(mAP)和IoU分别为0.5和0.75时的平均精度($AP^{0.5}$，$AP^{0.75}$)以及推断帧率(FPS)作为评估模型目标分类能力和检测速度的评估参数。

实验采用预训练的ResNet-50作为特征提取网络获取图像的高维语义特征[5]，使用随机梯度下降(SGD)对模型进行优化，动量系数为0.9，权重衰减系数为0.0002，训练总共需要迭代33750次。在训练过程中，模型初始学习率γ设置为0.001，并且在第27600次和第30000次迭代时学习率权重衰减系数分别为0.1和0.1。对彩色图像系数进行微调，将基于CenterNet红外图像目标检测模型中的超参数分别设置为：$\lambda_{\text{size}}=0.1$、$\lambda_{\text{off}}=0.8$、$\alpha_{\text{c}}=0.25$、$\beta_{\text{c}}=2$。在基于CenterNet的一致性半监督目标检测模型中，学生网络更新教师网络的EMA系数$\alpha_{\text{t}}=0.99$，无监督损失函数的权重$\lambda_{\text{u}}=3$。OMix增强操作过程中，两张图片的混合系数λ从Beta(α,α)中采样而来，$\alpha=80$，产生的两种类型损失函数的权重$\gamma_{\text{o1}}=1$、$\gamma_{\text{o2}}=0.15$。

4.2 基于CenterNet的半监督目标检测模型

利用FLIR数据集对当前主流的监督目标检测模型进行测试，结果如表1所示。在使用全部数据集的情况下(train3k+train5.8k)，基于关键点的目标检测模型CenterNet的全类平均精度达到60.5%，FPS达到64，与其他基于目标框目标检测模型(如Faster RCNN、YOLOv3、SSD)等相比，在速度和准确度上都有一定的优势。虽然CenterNet与YOLOv3的检测精度与检测速度相差不大，但基于锚框的目标检测算法性能容易受到锚框信息的影响，使用CenterNet更容易训练出相对鲁棒的模型。并且，在仅使用部分数据集(train-3k)训练的情况下，CenterNet的mAP相较于YOLOv3有0.5个百分点的提升，说明在红外图像目标检测过程中，CenterNet在较小的数据集之下仍有较好的鲁棒性。

表1 采用监督学习-使用有标签的数据进行训练时的测试结果

Table 1 Test results when using supervised learning and training on labeled data

Method	Labeled data	$AP^{0.5:0.95}$	$AP^{0.5}$	$AP^{0.75}$	FPS
Faster-RCNN[4]	train-3k	55.1	77.6	58.7	6
	train-3k+train-5.8k	**61.9**	**81.4**	**66.7**	
SSD300[5]	train-3k	51.7	74.2	55.9	**67**
	train-3k+train-5.8k	58.3	78.9	61.8	
YOLOv3[3]	train-3k	52.9	72.7	55.3	60
	train-3k+train-5.8k	60.2	79.1	63	
CenterNet	train-3k	53.4	72.1	57	64
	train-3k+train-5.8k	60.5	79.3	64.3	

如表2所示，对IRCC模型进行测试，在使用train-3k作为有标签数据、train-5.8k作为无标签数据的情况下，IRCC模型的mAP为55.3%，与仅使用部分训练数据集train-3k相比，其检测精度mAP提升了1.9个百分点。将IRCC模型与同类半监督目标检测模型进行对比，IRCC模型相对CSD模型mAP提升了1.4个百分点，说明基于TSML半监督学习方法在红外图像中能更好地利用无标签数据，具有良好的泛化性能。与STAC模型相比，IRCC模型的mAP虽然只提升了0.2个百分点，但是基于关键点的IRCC模型不需要统计锚框的先验信息，影响训练结果的超参数较少，具有更好的鲁棒性。

表2 采用半监督学习-且同时使用有标签和无标签数据进行训练的结果

Table 2 Test results when using semi-supervised learning and training on labeled and unlabeled data

Method	Labeled data	Unlabeled data	mAP /%
CSD[20]	train-3k	train-5.8k	53.9
STAC[21]	train-3k	train-5.8k	55.1
IRCC+cutout	train-3k	train-5.8k	55.3
IRCC+OMix(type 1)			55.6
IRCC+OMix(type 2)	train-3k	train-5.8k	56
IRCC+OMix(type 1,2)			**56.8**

4.3　基于OMix增强的半监督目标检测模型

强图像增强在基于一致性的半监督目标检测模型中具有重要作用，基于OMix增强的图像能更好利用红外图像中的像素点信息，本小节通过实验验证其对半监督目标检测模型的影响。与其他半监督目标检测中图像的增强方法相同，首先对无标签数据进行增强操作，当需要对弱增强图片中的目标应用几何变换时，强图像增强图片中对应区域也需要进行相应操作[21]。实验使用随机增强[22]的策略对几何变换（geometric transformation）和插值变换（cutout或OMix）产生的搜索空间进行随机搜索，并且每一次增强操作都选取随机不同等级的增强强度。对于一组图片中的相同位置，进行相同的几何变换。首先从几何变换中随机选择一种几何变换应用于图像，之后选择实验设置的插值变换（cutout或OMix）进行增强。通过表2可知，基于OMix增强的目标检测模型相比使用cutout增强的模型mAP提高了1.5个百分点。OMix增强同时使用两种类型损失函数的精度达到56.8%，充分说明OMix增强在半监督目标检测中的有效性。与全监督目标检测情况下（train-3k）的CenterNet目标检测模型进行对比，使用额外无标签数据的半监督模型（IRCC-OMix）取得了3.4个百分点的性能提升，这说明基于CenterNet和OMix增强的半监督红外图像目标检测模型可以充分挖掘无标签图像的特征，获得更可靠的识别准确率。

为进一步分析OMix提升模型检测性能的原因，实验分别统计了对不同种类目标的测试结果，如表3所示。在FLIR数据集中，汽车一般为像素点较多的大目标，而行人和自行车多为小目标，检测难度也更大。OMix增强过程中单独使用类型1的损失函数，与使用cutout增强相比对于汽车和行人的检测准确率分别提升了0.4个百分点和0.8个百分点。这是因为类型1增强为前景目标与背景混合，混入背景的过程中增大了目标的识别难度，提高了模型对较大目标的识别准确度。而单独使用类型2的OMix增强方式，与使用cutout增强相比能够显著增强对小目标的识别能力，对行人和自行车的识别准确率分别达到了61.7%和39.9%。对比使用cutout增强操作的半监督模型，同时使用两种损失函数类型的OMix增强操作对3种目标的识别能力都达到了最优，mAP为56.8%。与基于cutout增强的半监督目标检测模型相比，使用OMix增强模型的mAP总体提升1.5个百分点，对汽车、行人以及自行车的检测性能分别提升了0.6个百分点、1.0个百分点和2.9个百分点，充分说明OMix增强对于mAP的提升主要来自数据集中小目标的贡献。这主要是由于半监督学习中的cutout操作可能会使红外图片中的小目标失去过多的特征点，导致深度互学习过程中的性能退化，而OMix增强方式能够充分利用图像中的像素点，提升模型对不同目标的识别能力，尤其对小目标的作用更为显著。

表3　不同种类目标的测试结果

Table 3　Test results for different kinds of object

Method	AP /%			mAP /%
	car	person	bicycle	
IRCC+cutout	67.4	61.3	37.2	55.3
IRCC+OMix(type 1)	67.8	62.1	36.9	55.6
IRCC+OMix(type 2)	66.4	61.7	39.9	56
IRCC+OMix(type 1,2)	**68**	**62.3**	**40.1**	**56.8**

4.4　模型效果对比

所提IRCC-OMix算法的实验效果与其他模型的对比如图4所示。使用CenterNet的全监督模型存在部分漏检的情况，使用半监督训练的IRCC算法能够充分利用无标签的数据，提升对红外图像中目标的检测性能。基于OMix增强的IRCC半监督模型对边界框的定位更加准确，能够检测出图4(b)、(c)中不同场景下像素点较少的行人，相较于其他模型，对红外图像中较小目标的检测能力有显著提升。

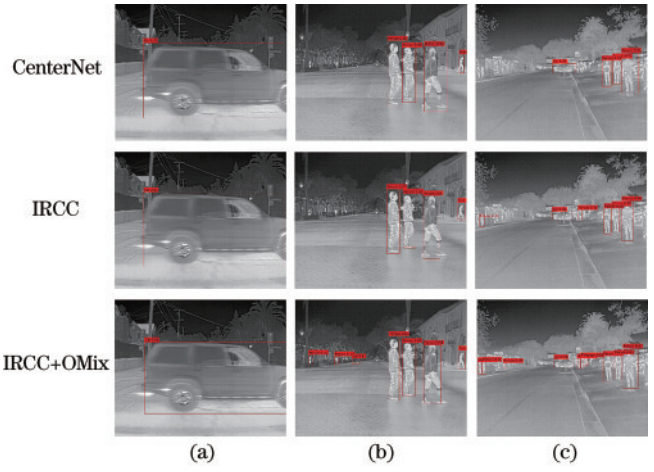

图4　算法效果展示。(a)大目标单一场景；(b)中小目标单一场景；(c)中小目标稠密场景

Fig. 4　Display of algorithm effect. (a) Large target single scene; (b) medium and small target single scene;(c) medium and small target dense scene

5　结　　论

设计了一种基于CenterNet与OMix增强的半监督红外图像目标检测算法（IRCC-OMix），通过基于关键点的目标检测模型提升模型的泛化性能，利用半监督学习使用无标签数据提高模型的检测准确度，设计OMix图像增强操作，增强模型对小目标的检测能力。与仅使用有标签数据进行训练的模型相比，IRCC-OMix模型能够充分利用无标签图像中的信息，提升检测的平均精度。同时，模型主干网络采用单阶段的检测算法CenterNet，其检测速度在服务器上达到64 frame·s^{-1}，能够更好适应嵌入式移动设备。然而，考

虑到模型在嵌入式设备上的应用，其在实时性上还有一定的提升空间。未来考虑对主干网络进一步轻量化，减少模型的参数，缩短模型在移动设备上的推断时间。

参 考 文 献

[1] Wang X, Lü G F, Xu L Z. Infrared dim target detection based on visual attention[J]. Infrared Physics & Technology, 2012, 55(6): 513-521.

[2] 宋子壮, 杨嘉伟, 张东方, 等. 基于无监督域适应的低空海面红外目标检测[J]. 光学学报, 2022, 42(4): 0415001.
Song Z Z, Yang J W, Zhang D F, et al. Low-altitude Sea surface infrared object detection based on unsupervised domain adaptation[J]. Acta Optica Sinica, 2022, 42(4): 0415001.

[3] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[4] Tian Y N, Yang G D, Wang Z, et al. Apple detection during different growth stages in orchards using the improved YOLO-V3 model[J]. Computers and Electronics in Agriculture, 2019, 157: 417-426.

[5] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[6] 许延雷, 梁继然, 董国军, 等. 基于改进CenterNet的航拍图像目标检测算法[J]. 激光与光电子学进展, 2021, 58(20): 2010013.
Xu Y L, Liang J R, Dong G J, et al. Aerial image target detection algorithm based on improved CenterNet[J]. Laser & Optoelectronics Progress, 2021, 58(20): 2010013.

[7] Wang Z Y, Li Y L, Guo Y, et al. Data-uncertainty guided multi-phase learning for semi-supervised object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 4566-4575.

[8] Berthelot D, Carlini N, Goodfellow I, et al. MixMatch: a holistic approach to semi-supervised learning[EB/OL]. (2019-05-06)[2021-05-03]. https://arxiv.org/abs/1905.02249.

[9] Xie Q Z, Luong M T, Hovy E, et al. Self-training with noisy student improves ImageNet classification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10684-10695.

[10] Xie Q Z, Dai Z H, Hovy E, et al. Unsupervised data augmentation for consistency training[EB/OL]. (2019-04-29)[2022-02-04]. https://arxiv.org/abs/1904.12848.

[11] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[12] Yun S, Han D, Chun S, et al. CutMix: regularization strategy to train strong classifiers with localizable features [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 6022-6031.

[13] Kim J M, Jang J, Seo S, et al. MUM: mix image tiles and UnMix feature tiles for semi-supervised object detection[EB/OL]. (2021-11-22) [2022-02-05]. https://arxiv.org/abs/2111.10958.

[14] Zhou Z W, Siddiquee M M R, Tajbakhsh N, et al. UNet: redesigning skip connections to exploit multiscale features in image segmentation[J]. IEEE Transactions on Medical Imaging, 2020, 39(6): 1856-1867.

[15] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[16] Law H, Deng J. CornerNet: detecting objects as paired keypoints[J]. International Journal of Computer Vision, 2020, 128(3): 642-656.

[17] Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 483-499.

[18] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2999-3007.

[19] Sohn K, Berthelot D, Li C L, et al. FixMatch: simplifying semi-supervised learning with consistency and confidence[EB/OL]. (2020-01-21)[2022-02-04]. https://arxiv.org/abs/2001.07685.

[20] Jeong J, Verma V, Hyun M, et al. Interpolation-based semi-supervised learning for object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 11597-11606.

[21] Sohn K, Zhang Z Z, Li C L, et al. A simple semi-supervised learning framework for object detection[EB/OL]. (2020-05-10) [2022-02-04]. https://arxiv. org/abs/2005.04757.

[22] Cubuk E D, Zoph B, Shlens J, et al. Randaugment: Practical automated data augmentation with a reduced search space[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 14-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 3008-3017.