DOI:10.19651/j.cnki.emt.2517997

特征软融合与正负样本对比的弱监督目标定位*

阮皓皓[1,2] 李冰锋[1,2] 李新伟[1,2] 冀得魁[1,2]
(1.河南理工大学电气工程与自动化学院 焦作 454000;2.河南省煤矿装备智能检测与控制重点实验室 焦作 454003)

摘 要:针对弱监督目标定位任务中,使用硬融合方式来融合深浅层特征导致网络过度关注区分性强区域或误将背景识别为目标的问题,本文提出了一种基于深浅层特征软融合和正负样本对比的弱监督目标定位方法。首先,提出的深浅层特征软融合策略通过设计前景生成器,分别从浅层特征和深层特征中生成前景预测图,然后采取反向监督操作,引导网络逐步学习多层细粒度特征,实现深浅层特征之间的相互优化。其次,本文基于对比学习思想提出了正负样本对比损失函数,通过构造正负样本,以引导网络在训练过程中更专注于前景区域,抑制背景噪声的干扰。本文在 CUB-200-2011 和 ILSVRC-2012 数据集上以验证本文方法的有效性,在两个数据集上的定位准确率分别达到了 95.77%和 72.90%。实验结果表明,本文方法在弱监督目标定位任务场景下的有效性和适用性。

关键词:弱监督目标定位;深浅层特征软融合;正负样本对比;前景生成器
中图分类号:TP391.4;TN791 **文献标识码**:A **国家标准学科分类代码**:520.60

Feature soft fusion and positive-negative sample contrast for weakly supervised object localization

Ruan Haohao[1,2] Li Bingfeng[1,2] Li Xinwei[1,2] Ji Dekui[1,2]
(1. School of Electrical Engineering and Automation, Henan University of Technology, Jiaozuo 454000, China; 2. Henan Province Key Laboratory for Intelligent Detection and Control of Coal Mine Equipment, Jiaozuo 454003, China)

Abstract: In weakly supervised object localization tasks, using hard fusion to combine deep and shallow features can cause the network to overly focus on discriminative regions or mistakenly identify the background as the object. To address this issue, this paper proposes a weakly supervised object localization method based on soft fusion of deep and shallow features and positive-negative sample contrast. First, the proposed soft fusion strategy for shallow and deep features generates foreground prediction maps from both shallow and deep features by designing a foreground generator. Then, a reverse supervision operation is applied to guide the network in gradually learning multi-level fine-grained features, achieving mutual optimization between shallow and deep features. Second, based on the concept of contrastive learning, a positive and negative sample contrastive loss function is proposed. By constructing positive and negative samples, the network is guided to focus more on the foreground regions during training while suppressing background noise interference. The effectiveness of the proposed method is validated on the CUB-200-2011 and ILSVRC-2012 datasets, achieving localization accuracies of 95.77% and 72.90%, respectively. The experimental results demonstrate the effectiveness and applicability of the proposed method in weakly supervised object localization tasks.

Keywords: weakly supervised object localization; soft fusion of deep and shallow features; positive-negative sample contrast; foreground generator

0 引 言

在人工智能和机器学习领域,弱监督学习作为一种重要的方法,近年来在多个研究方向中得到了广泛应用,例如语义分割、目标检测以及细粒度图像分类等[1-2]。与完全监督学习依赖于大量高质量、精确标注的数据不同,弱监督学习通过利用有限、有噪声或不精确的标注,开发出性能优越的模型,从而有效降低了对大规模标注数据的依赖。随着

收稿日期:2025-01-24
* 基金项目:河南理工大学博士基金(B2018-33)项目资助

弱监督学习的迅速发展，弱监督目标定位（weakly supervised object localization，WSOL）逐渐成为计算机视觉领域的研究热点。WSOL 通过仅使用包含图像级标签的训练数据，训练模型以精确地定位图像中的目标对象。其核心目标是在粗粒度标注的基础上生成细粒度的目标定位信息，为数据标注成本较高或细粒度标注稀缺的场景提供了有效解决方案。

作为一种应用价值显著的技术，WSOL 在应对复杂和多样化视觉任务时展现了强大的潜力，为弱监督学习方法在实际场景中的应用提供了重要支撑。其中，Zhou 等[3]提出的类激活图（class activation map，CAM）是 WSOL 领域最具代表性的方法之一。CAM 利用分类网络的内部特征生成目标的定位图，这种能力使其成为众多后续技术改进的基础。在 CAM 的基础上，研究者提出了多种改进方法。例如，Mai 等[4]提出的对抗擦除方法，通过在训练过程中有选择性地去除目标图像中最具辨识性的区域，迫使模型关注目标的更大范围，从而提高定位效果。Zhang 等[5]提出一种多阶段方法，通过分别优化分类任务和定位任务，解决了这两者之间的任务冲突问题。这些改进方法不断提升 WSOL 的性能和适用范围，为其在实际应用中解决更加复杂的目标定位问题奠定了坚实的技术基础。

尽管基于 CAM 的 WSOL 方法取得了成功，但这些方法通常侧重于捕捉目标最具区分性的部分，而非准确地勾画整个目标区域。此外，网络分类器在同时承担定位和分类任务时，必然会影响网络的性能，并带来优化上的挑战。为了解决这些问题，最近有研究提出了可学习的激活图生成器。这些方法通过直接从特征提取网络生成前景预测图来实现目标定位。例如，Xie 等[6]设计了一种激活图生成器，通过分类网络初步生成粗略的前景预测图，然后通过多重损失函数和区域擦除引导学习，从而生成更精确的前景预测图。Meng 等[7]则使用一种特殊的前景记忆机制，从像素特征中生成前景预测图，并通过部件感知注意力模块进一步优化，以增强定位效果。最近，Zhai 等[8]引入了一种激活图约束模块，通过减少背景区域的激活值来提升前景预测图的质量。这些方法主要通过深度特征图提取前景预测，确保在定位过程中，网络能够利用丰富的语义信息维持分类性能。

在定位任务中，单纯依赖深层特征往往会导致结果不理想，因为缺乏关于目标位置的详细信息。卷积神经网络的早期层提取的浅层特征虽然语义信息较少，但包含了更多的细节，如更清晰的边缘和较少的失真。文献[9]使用逐元素相乘方法来融合浅层特征和深层特征，然而使用这种硬融合方法来融合不同层的特征图，常常导致网络过于关注目标的最具区分性区域，或误将背景区域识别为目标区域。

为了解决上述问题，本文提出了一种基于深浅特征软融合和正负样本对比（positive-negative sample contrast，PNSC）的 WSOL 方法，在实现更加精确和鲁棒的目标定位效果。该方法通过精心设计的前景生成器，分别从深层特征和浅层特征中生成前景预测图，并将其中一组前景预测图作为伪标签，反向监督另一组预测图，以实现特征间的相互优化与协同。作为一种软融合策略，该方法在训练过程中能够有效引导网络逐步学习多层细粒度特征，从而提升网络对目标细节的捕捉能力。同时，考虑到在复杂场景中，软融合方法可能引入背景噪声的问题，这种背景噪声往往会干扰定位的准确性并导致结果出现偏差，本文通过构建前景预测图作为正样本、构建背景预测图作为负样本，然后基于对比学习思想设计正负样本对比损失函数。该方法能够有效引导网络更加专注于前景区域的特征表示，同时抑制其对背景区域的关注，从而进一步提升定位的精确性和可靠性。最后，本文在多个公开数据集上进行了广泛的实验验证，实验结果全面表明本文提出的方法在定位性能上具有显著优势，不仅提升了目标定位的准确率，还展现出较强的通用性和适应性，从而证明了该方法的有效性和应用潜力。

1　本文方法

1.1　网络架构

本文方法的网络架构如图 1 所示，主要由深浅层特征软融合模块和正负样本对比模块构成，关于深浅层特征软融合模块的详细内容将在第 1.2 节中介绍，关于 PNSC 模块的详细内容将在第 1.3 节中解释。

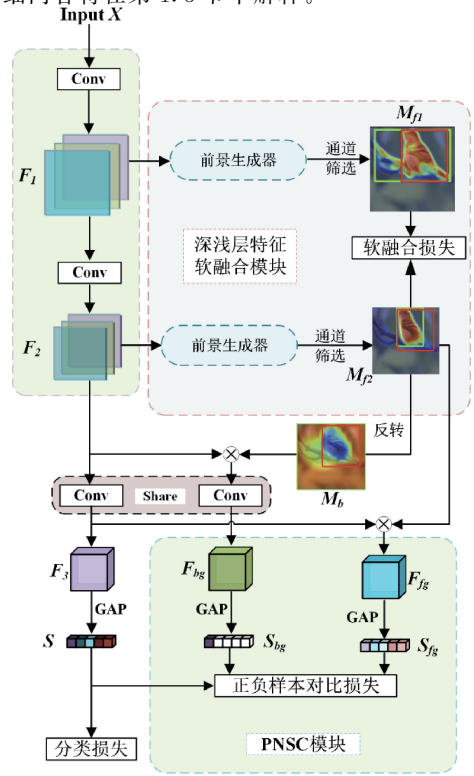

图 1　本文方法的网络架构图

Fig. 1　The network architecture diagram of our method

本文方法使用常见的卷积神经网络[10-12]作为特征提取网络。对于输入图像 X,首先从特征提取网络的浅层提取浅层特征,记为 $F_1 \in R^{H_1 \times W_1 \times N_1}$,其中 H_1、W_1 和 N_1 分别表示 F_1 的高度、宽度和通道数。接着,在特征提取网络的最后一个卷积层之前,提取深层特征,记为 $F_2 \in R^{H_2 \times W_2 \times N_2}$,其中 H_2、W_2 和 N_2 分别表示 F_2 的的高度、宽度和通道数。最后,F_2 被传递到主干网络的最后一个卷积层,得到 $F_3 \in R^{H_2 \times W_2 \times C}$。在对特征 F_3 进行全局平均池化后,能够获得原始分类输出,即类别预测输出 $S \in R^C$,其中 C 表示数据集中类别的数量。类别预测输出 S 通过交叉熵分类损失函数 L_{cls} 进行监督,如式(1)所示。

$$L_{cls} = -\sum_{i=1}^{C} y_i \log\left(\frac{\exp(S^i)}{\sum_{j}^{C}\exp(S^j)}\right) \tag{1}$$

式中:y_i 表示图像级的独热编码标签。

1.2 深浅层特征软融合模块

深浅层特征软融合模块的主要思想是通过提高浅层特征在定位任务中的利用率,从而提升定位性能。在卷积神经网络中,随着网络层数的加深,通过池化或下采样等操作,特征图的尺寸会被缩小,导致目标细节特征的退化或丢失,对于目标定位任务来说,深层特征显得略微粗糙。而浅层特征虽然其包含的语义信息较弱,但在其中包含丰富的目标结构信息与细节信息。因此,提高浅层特征对定位的贡献是必要的。

深浅层特征软融合模块主要由两部分构成。首先,将浅层特征 F_1 和深层特征 F_2 分别通过前景生成器,用以生成两组不同的前景预测图集合。每个前景生成器均由一个卷积核大小为 3×3 的卷积层和一个 *sigmoid* 归一化层构成,用以提取前景信息。接下来,从浅层前景预测图集合中,利用通道筛选操作选择出一个特定通道的浅层前景预测图 $M_{f1} \in R^{H_1 \times W_1}$,上述过程可用如式(2)所示。

$$M_{f1} = Select\{Sig(Conv(F_1))\} \tag{2}$$

式中:$Conv()$ 表示使用卷积操作,$Sig()$ 表示使用 *sigmoid* 函数进行归一化操作,$Select\{\}$ 表示通道筛选操作。

类似地,对深层特征 F_2 应用相同的操作以得到深层前景预测图 $M_{f2} \in R^{H_2 \times W_2}$,如式(3)所示。

$$M_{f2} = Select\{Sig(Conv(F_2))\} \tag{3}$$

在训练过程中,先将浅层前景预测图 M_{f1} 下采样到与深层前景预测图 M_{f2} 同样的维度大小 (H_2, W_2),并将其作为伪标签监督深层前景预测图 M_{f2},但不参与反向传播过程,如图 2 所示。

同样地,将深层前景预测图 M_{f2} 上采样到与浅层前景预测图 M_{f1} 同样的维度大小 (H_1, W_1),并将其作为伪标签监督浅层前景预测图 M_{f1},同样不参与反向传播过程,如图 3 所示。监督过程中采用软融合损失 L_{sf},以确保浅层和深层特征之间的协同优化。损失 L_{sf} 的具体定义如式(4)所示。

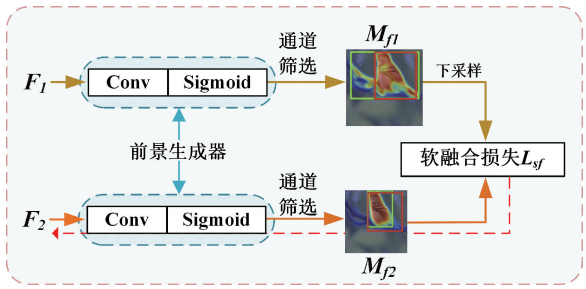

图 2 深浅层特征软融合模块:M_{f1} 为伪标签

Fig. 2 Multi-Scale deformable grouped residual module: M_{f1} as the pseudo-label

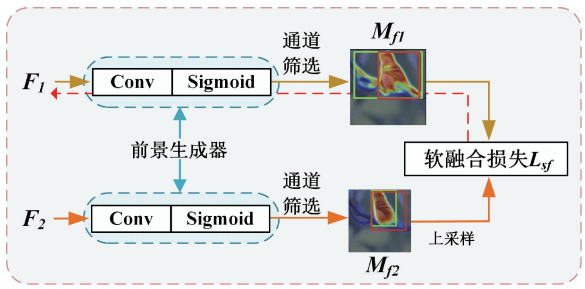

图 3 深浅层特征软融合模块:M_{f2} 为伪标签

Fig. 3 Multi-Scale deformable grouped residual module: M_{f2} as the pseudo-label

$$L_{sf} = \omega_1 \frac{1}{H_1 W_1}\sum_{i=1}^{H_1}\sum_{j=1}^{W_1}(M_{f1}(i,j) - M_{f2}^{*}(i,j))^2 + \omega_2 \frac{1}{H_2 W_2}\sum_{i=1}^{H_2}\sum_{j=1}^{W_2}(M_{f1}^{*}(i,j) - M_{f2}(i,j))^2 + \frac{1}{\chi}\sum_{x\in\chi}M_{f1}[x] + \frac{1}{\gamma}\sum_{y\in\gamma}M_{f2}[y] \tag{4}$$

式中:$*$ 表示在训练阶段的反向传播过程中被冻结的部分。ω_1 和 ω_2 表示加权系数,用于平衡软融合损失 L_{sf}。$\frac{1}{\chi}\sum_{x\in\chi}M_{f1}[x] + \frac{1}{\gamma}\sum_{y\in\gamma}M_{f2}[y]$ 作为两个正则化项,用于限制前景区域的范围,以避免因前景区域过大而导致的次优结果。

在图 4 中,前 3 行从上至下每一行分别表示输入图像、浅层前景预测图 M_{f1} 和深层前景预测图 M_{f2}。图 3 最后一行展示了通过深浅层特征软融合模块之后,生成的前景预测图为网络提供更精确的目标区域信息。

1.3 正负样本对比模块

尽管经过深浅层特征软融合模块处理后的前景预测图定位精度有所提高,但在某些复杂场景下,浅层前景预测图 $\boldsymbol{M}_{f1}$ 和深层前景预测图 $\boldsymbol{M}_{f2}$ 中可能同时存在背景噪声,如图 5 所示,从上至下每一行分别表示输入图像、浅层前景预测图 $\boldsymbol{M}_{f1}$ 和深层前景预测图 $\boldsymbol{M}_{f2}$。

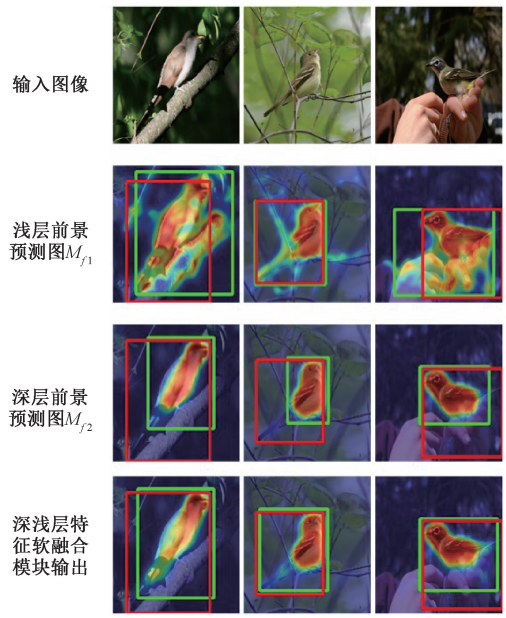

图 4　深浅层特征软融合模块效果展示

Fig. 4　Illustration of the effect of the soft fusion module for deep and shallow features

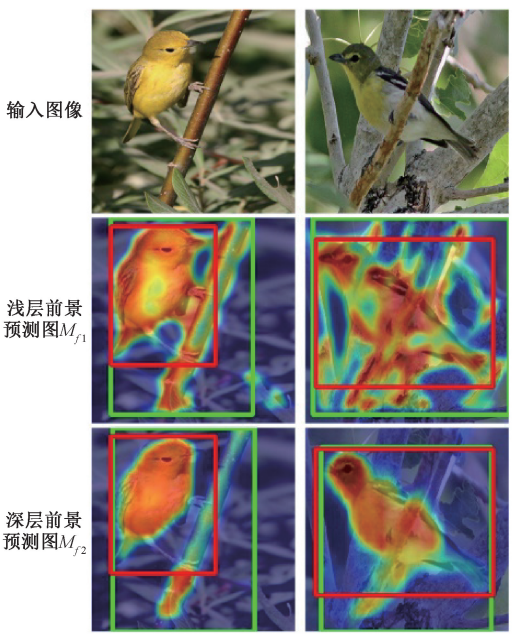

图 5　复杂场景下背景噪声对定位结果的影响

Fig. 5　The impact of background noise on localization results in complex scenarios

为了解决这一问题，单纯依赖深浅层特征软融合模块是不够的，因此本文进一步提出一种 PNSC 模块，该模块的核心思想是利用深层前景预测图 $\boldsymbol{M}_{f2}$，分别构造出正样本分支和负样本分支，并基于对比学习思想设计了一种正负样本对比损失函数 $\boldsymbol{L}_{PNSC}$，其目标是最大化正样本之间的相似性和最小化正负样本之间的相似性，从而强化网络对目标区域的精确定位能力。

具体来说，PNSC 模块主要分为正样本分支和负样本分支。通过对深层前景预测图 $\boldsymbol{M}_{f2}$ 进行逐元素反转操作，可以得到背景预测图 $\boldsymbol{M}_b \in \boldsymbol{R}^{H_2 \times W_2}$，上述过程如式(5)所示。

$$M_b = 1 - M_{f2} \tag{5}$$

随后将 M_b 与 F_2 相乘然后送入最后一层卷积层中，生成背景特征 $F_{bg} \in R^{H_2 \times W_2 \times C}$，上述过程如式(6)所示。

$$F_{bg} = Conv(M_b \times F_2) \tag{6}$$

通过对背景特征 F_{bg} 进行全局平均池化操作，可以将其转换为一个负样本预测输出 S_{bg}。

对于正样本分支，将 M_{f2} 与 F_3 相乘得到前景特征 $F_{fg} \in R^{H_2 \times W_2 \times C}$，上述过程如式(7)所示。

$$F_{fg} = M_{f2} \times F_3 \tag{7}$$

同样对前景特征 F_{fg} 进行全局平均池化操作，可以将其转换为一个正样本预测输出 S_{fg}。

最后将类别预测输出 S、正样本预测输出 S_{fg}、负样本预测输出 S_{bg} 一起送入正负样本对比损失函数 L_{PNSC} 中，具体过程如式(8)所示。

$$L_{PNSC} = \frac{S \cdot S_{fg}}{\| S \|_2 \| S_{fg} \|_2} + \left(1 - \frac{S \cdot S_{bg}}{\| S \|_2 \| S_{bg} \|_2}\right) \tag{8}$$

在正负样本对比损失函数 L_{PNSC} 中，将类别预测输出 S 和正样本预测输出 S_{fg} 视为一对正样本，通过最大化正样本之间的相似性，增强网络对目标区域特征的一致性学习，从而提高目标定位的精确性。将类别预测输出 S 和负样本预测输出 S_{bg} 视为一对正负样本，通过最小化正负样本之间的相似性，引导网络有效区分目标与背景特征，进一步提升定位边界的清晰度和准确性。

2　实验设计与结果分析

2.1　实验数据集和评价指标

本文在两个公共数据集上评估了本文的方法：CUB-200-2011[13] 和 ILSVRC-2012[14]。CUB-200-2011 数据集包含 200 个鸟类类别的 11 788 张图像，其中 5 994 张用于训练，5 794 张用于测试。除了类别标签外，CUB-200-2011 数据集中的每张图像还包括位置信息注释，这些注释仅用于评估目标定位预测结果。ILSVRC-2012 数据集包含约 120 万张跨越 1 000 个不同类别的训练图像以及 50 000 张验证图像。

同时，本文采用了 3 种常用指标来评估本文的方法。这些指标包括 Top-1 定位准确率(Top-1 Loc)、Top-5 定位准确率(Top-5 Loc)以及 GT-Known 定位准确率(GT-known Loc)。这些指标在评估 WSOL 方法时被广泛接受。此外，参考文献[15-17]，本文还采用了最大框准确率(MaxBoxAccV2)。这是一个新兴的指标，用于评估网络的

定位性能,能够更直观地反映网络在不同 IoU 阈值($\delta \in 0.3, 0.5, 0.7$)下的定位能力。更高的 MaxBoxAccV2 值表明目标定位的准确性更高。

2.2 实验细节

本文在多个基础网络上评估了本文的方法,包括 VGG16、ResNet50 和 InceptionV3。在训练过程中,本文将输入图像的大小调整为 256×256,然后随机裁剪为 224×224。对于模型优化,本文使用了随机梯度下降优化器,学习率设为 0.001,并且采用 10 倍衰减率。在 CUB-200-2011 和 ILSVRC-2012 数据集上,批量大小设置为 32。本文的所有实验均基于 PyTorch 深度学习框架,且所有实验均在配置有 Ubuntu18.04.6 LTS 操作系统的服务器上进行,该服务器搭载了一张拥有 24 GB 显存的 Nvidia GeForce RTX 4090 GPU 和一款 Intel Core i7-12700 处理器。

2.3 消融实验

为了充分探索深浅层特征软融合模块的作用,在 CUB-200-2011 数据集上,本文使用 VGG16 作为特征提取网络,对软融合损失 L_{sf} 中的加权系数 ω_1 和 ω_2 进行了消融实验研究,实验结果如表 1 所示。

表 1 软融合损失中加权系数消融实验分析

Table 1 Ablation study analysis of the weighted coefficients in the soft fusion loss %

加权系数	CUB-200-2011		
(ω_1,ω_2)	Top-1 Loc	Top-5 Loc	GT-known Loc
(0,1)	63.23	79.12	82.47
(5,1)	68.93	83.28	88.90
(2,1)	69.36	83.74	89.23
(1,1.5)	**70.68**	**85.33**	**90.82**
(1,4)	69.66	83.97	90.32
(1,8)	68.24	83.31	89.87
(1,0)	65.57	80.64	85.16

通过逐步调整两个加权系数的值,实验分析了它们对网络性能的影响,从而揭示了不同层次特征的融合效果及其对定位精度的贡献。从表 1 中可以看出,当权重系数(ω_1,ω_2)的取值为(1, 1.5)时,网络取得了最高的定位精度。这表明,通过适当放大加权系数 ω_2,可以有效地提升网络在目标区域定位上的表现。然而,当 ω_1 或 ω_2 的值设置过大时,定位性能会受到不利影响,可能导致模型过度关注某一特征层,进而影响整体的定位效果。

为验证本文所提出的深浅层特征软融合模块和 PNSC 模块的有效性,本文在 CUB-200-2011 数据集上,使用 VGG16 作为特征提取网络进行了消融实验。实验结果如表 2 所示。通过逐步引入这些模块并进行比较,可以看到各个模块的贡献,以及它们对目标定位精度的提升效果。从表 2 中可以看出,与基线方法(Baseline)对比,分别引入深浅层特征软融合模块和 PNSC 模块均能提升网络的定位性能。此外,同时引入两个模块,即本文的方法分别在 Top-1 Loc、Top-5 Loc、GT-known Loc 指标上达到了 70.68%、85.33%、90.82%的定位准确率。

表 2 本文方法中各个模块消融实验

Table 2 Ablation experiments on each module of the proposed method %

方法	CUB-200-2011		
	Top-1 Loc	Top-5 Loc	GT-known Loc
Baseline	53.45	65.46	70.39
+软融合	68.93	83.28	88.90
+PNSC	67.85	81.83	86.97
本文方法	**70.68**	**85.33**	**90.82**

2.4 对比实验

本文基于多种特征提取网络,在 CUB-200-2011 和 ILSVRC-2012 数据集上,与其他 WSOL 方法进行了比较。

表 3 展示了在 CUB-200-2011 数据集上,且基于 VGG16 特征提取网络,本文提出的方法与其他 WSOL 方法的比较结果。从表中可以看出,本文提出的方法在 Top-1 Loc 指标上达到了 70.68%,在所有评估的方法中取得了最佳结果(在表中加粗标注的数据)。

表 3 在 CUB-200-2011 数据集上,基于 VGG16 网络的本文方法与其他方法对比实验分析

Table 3 Comparative analysis of the proposed method with other methods based on the VGG16 on the CUB-200-2011 dataset

方法	特征提取网络	Top-1 Loc/%	Top-5 Loc/%	GT Loc/%
CAM[3]	VGG16	44.15	52.16	58.00
ADL[18]	VGG16	52.36	N/A	75.40
EIL[4]	VGG16	56.21	69.51	72.37
RCAM[19]	VGG16	58.96	N/A	76.30
PSOL[5]	VGG16	60.27	72.45	77.29
DPM[17]	VGG16	67.30	82.20	82.20
CREAM[20]	VGG16	70.44	85.67	90.98
Ours	VGG16	**70.68**	85.33	90.82

表 4 展示了在 CUB-200-2011 数据集上,且基于 ResNet50 特征提取网络,本文提出的方法与其他 WSOL 方法的比较结果。从表中可以看出,本文提出的方法在 Top-1 Loc、Top-5 Loc、GT-known Loc 指标上分别达到了 76.34%、89.80%、95.77%的定位准确率,在所有评估的方法中取得了最佳结果(在表中加粗标注的数据)。与

CREAM 方法相比，本文提出的方法在 GT-known Loc 指标上提升了 5.89%。

表 4　在 CUB-200-2011 数据集上，基于 ResNet50 网络，本文方法与其他方法对比实验分析

Table 4　Comparative analysis of the proposed method with other methods based on the ResNet50 on the CUB-200-2011 dataset

方法	特征提取网络	Top-1 Loc/%	Top-5 Loc/%	GT Loc/%
CAM[3]	ResNet50	46.71	54.44	57.35
RCAM[19]	ResNet50	59.53	N/A	77.58
PSOL[5]	ResNet50	70.68	86.64	90.00
FAM[7]	ResNet50	73.74	N/A	85.73
DPM[17]	ResNet50	71.20	83.60	82.30
CREAM[20]	ResNet50	76.03	N/A	89.88
Ours	ResNet50	**76.34**	**89.80**	**95.77**

表 5 展示了在 CUB-200-2011 数据集上，且基于 InceptionV3 特征提取网络时，本文提出的方法全面优于所有其他的 WSOL 方法，分别达到了 75.04%的 Top-1 Loc、89.40%的 Top-5 Loc 和 95.63%的 GT-known Loc 定位性能。与最新 BAS 方法相比，本文提出的方法在 Top-1 Loc、Top-5 Loc、GT-known Loc 指标上分别提升了 2.95%、1.29%、1.00%。

表 5　在 CUB-200-2011 数据集上，基于 InceptionV3 网络的本文方法与其他方法对比实验分析

Table 5　Comparative analysis of the proposed method with other methods based on the inceptionV3 on the CUB-200-2011 dataset　%

方法	特征提取网络	Top-1 Loc	Top-5 Loc	GT Loc
CAM[3]	InceptionV3	41.06	50.66	55.10
RCAM[19]	InceptionV3	51.05	N/A	65.10
PSOL[5]	InceptionV3	65.51	83.44	N/A
FAM[7]	InceptionV3	70.67	N/A	87.25
DPM[17]	InceptionV3	64.30	69.60	79.60
CREAM[20]	InceptionV3	71.76	86.37	90.43
BAS[8]	InceptionV3	72.09	88.11	94.63
Ours	InceptionV3	**75.04**	**89.40**	**95.63**

表 6 展示了在 ILSVRC-2012 数据集上，且基于 VGG16 特征提取网络，本文提出的方法与其他 WSOL 方法的比较结果。从表中可以看出，本文提出的方法在 Top-1 Loc 指标上达到了 52.23%，在 Top-5 Loc 指标上达到了 64.93%，在 GT-known Loc 指标上达到了 69.47%。

表 6　在 ILSVRC-2012 数据集上，基于 VGG16 网络的本文方法与其他方法对比实验分析

Table 6　Comparative analysis of the proposed method with other methods based on the VGG16 on the ILSVRC-2012 dataset　%

方法	特征提取网络	Top-1 Loc	Top-5 Loc	GT Loc
CAM[3]	VGG16	42.80	54.86	59.00
ADL[18]	VGG16	44.92	N/A	N/A
EIL[4]	VGG16	46.81	N/A	N/A
RCAM[19]	VGG16	44.62	57.92	60.73
PSOL[5]	VGG16	50.89	60.90	64.03
DPM[17]	VGG16	51.10	63.80	69.30
CREAM[20]	VGG16	52.37	64.20	68.32
Ours	VGG16	52.23	**64.93**	**69.47**

表 7 展示了在 ILSVRC-2012 数据集上，且基于 ResNet50 特征提取网络，本文提出的方法与其他 WSOL 方法的比较结果。从表中可以看出，本文提出的方法在 Top-1 Loc 指标上达到了 57.72%，在 Top-5 Loc 指标上达到了 69.29%，在 GT-known Loc 指标上达到了 72.90%，在所有评估的方法中取得了最佳结果(在表中加粗标注的数据)。与 CREAM 方法相比，本文提出的方法在 Top-1 Loc 和 GT-known Loc 指标上分别提升了 2.06%、3.59%。

表 7　ILSVRC-2012 数据集上，基于 ResNet50 网络的本文方法与其他方法对比实验分析

Table 7　Comparative analysis of the proposed method with other methods based on the ResNet50 on the ILSVRC-2012 dataset　%

方法	特征提取网络	Top-1 Loc	Top-5 Loc	GT Loc
CAM[3]	ResNet50	38.99	49.47	51.86
RCAM[19]	ResNet50	49.42	N/A	62.20
PSOL[5]	ResNet50	53.98	63.08	65.44
FAM[7]	ResNet50	54.46	N/A	64.56
DPM[17]	ResNet50	54.40	65.50	69.60
CREAM[20]	ResNet50	55.66	N/A	69.31
Ours	ResNet50	**57.72**	**69.29**	**72.90**

在表 8 中，展示了基于 InceptionV3 特征提取网络，本文提出的方法与其他 WSOL 方法在 ILSVRC-2012 数据集上的比较结果。从表中可以看出，本文提出的方法在三种指标上超越了所有其他方法。与最近的 BAS 方法相比，本文提出的方法在 Top-1 Loc 指标上提高了 0.54%，在 Top-5 Loc 指标上提高了 0.7%，在 GT-known Loc 指标上提高了 0.67%。

表 8 ILSVRC-2012 数据集上,基于 InceptionV3 网络的本文方法与其他方法对比实验分析

Table 8 Comparative analysis of the proposed method with other methods based on the InceptionV3 on the ILSVRC-2012 dataset

方法	特征提取网络	Top-1 Loc	Top-5 Loc	GT Loc
CAM[3]	InceptionV3	46.29	58.19	62.68
RCAM[19]	InceptionV3	47.70	N/A	62.76
PSOL[5]	InceptionV3	54.82	63.25	65.21
FAM[7]	InceptionV3	55.24	N/A	68.62
CREAM[20]	InceptionV3	56.07	66.19	69.03
BAS[8]	InceptionV3	58.50	69.03	72.07
Ours	InceptionV3	**59.04**	**69.73**	**72.74**

与 GT-known Loc 相比,MaxBoxAccV2 指标能够评估在不同阈值下($\delta \in 0.3, 0.5, 0.7$)的定位性能。图 6 展示了在 CUB-200-2011 和 ILSVRC-2012 数据集上,MaxBoxAccV2 指标在使用不同特征提取网络时,本文方法与其他方法进行比较的结果。

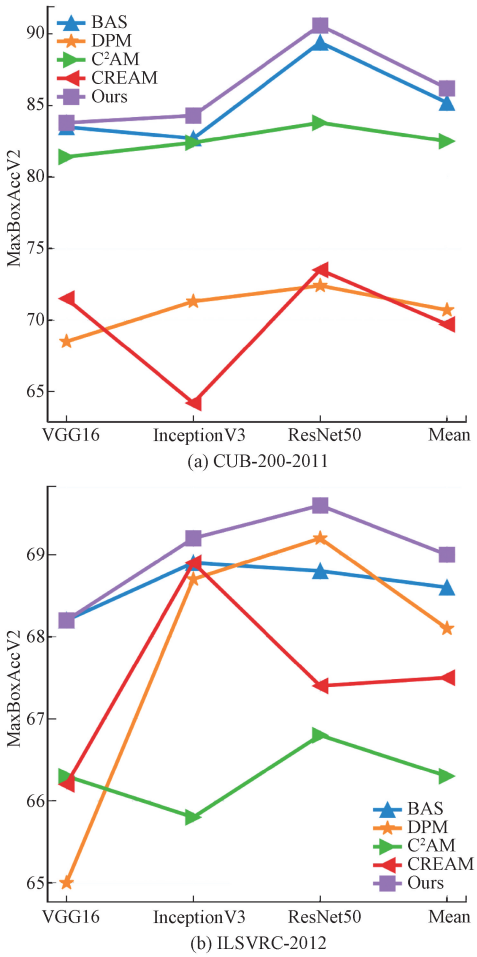

图 6 最大框准确率(MaxBoxAccV2)对比实验分析

Fig. 6 Comparative experimental analysis of MaxBoxAccV2

实验数据表明,在 MaxBoxAccV2 指标上,本文方法在使用不同特征提取网络时始终优于其他方法。这突显了本文方法生成的边界框的优越性,并验证了其有效性和广泛适用性。

2.5 可视化分析

为进一步验证本文所提方法的显著效果,本文展示了 CUB-200-2011 和 ILSVRC-2012 数据集中的选定样本的预测结果,如图 7 和 8 所示。图 7 中从上至下每一行分别表示输入图像、CAM 方法预测结果、本文方法预测结果。在 CUB-200-2011 数据集上,可以观察到,CAM 方法主要强调物体的最具区分性的区域,而本文方法则能够将注意力分布在多个物体区域,从而实现更精确的定位。

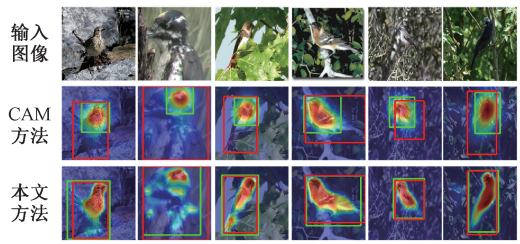

图 7 CUB-200-2011 数据集选定样本的可视化对比

Fig. 7 Visual comparison of selected samples from the CUB-200-2011 dataset

图 8 中,在 ILSVRC-2012 数据集上,显而易见的是,本文方法能够有效消除非目标类别的干扰,并且能够在图像中存在多个相同类别或不同类别目标的情况下,完成精确的定位任务。

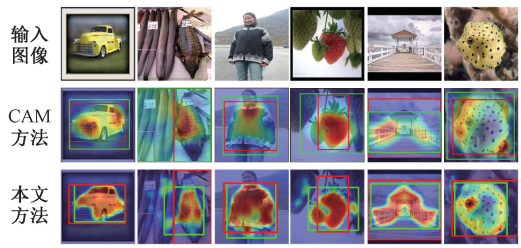

图 8 ILSVRC-2012 数据集选定样本的可视化对比

Fig. 8 Visual comparison of selected samples from the ILSVRC-2012 dataset

3 结 论

本文提出了一种基于深浅特征软融合和正负样本对比的弱监督目标定位方法。通过设计前景生成器,本文分别从深层特征和浅层特征中生成前景预测图,并利用软融合策略进行协同优化,进而提升了网络对目标细节的捕捉能力。此外,本文针对复杂场景中可能出现的背景噪声,提出了正负样本对比模块,通过正负样本对比损失函数引导网络更专注于前景区域的特征表示,抑制背景区域的干扰,进一步提高了定位精度。实验结果表明,本文提出的方法在多个公开数据集上的实验中均取得了显著的性能提升,尤

其在定位精度和鲁棒性方面表现突出，证明了该方法的有效性和广泛的应用潜力。本文的下一步工作将集中于进一步优化模型和改善方法。同时积极探索如何提升模型在复杂场景下（如图像中存在同类多目标、图像中白噪声过多等场景）的定位能力。

参考文献

[1] SHAO F F, CHEN L, SHAO J, et al. Deep learning for weakly-supervised object detection and localization: A survey[J]. Neurocomputing, 2022, 496: 192-207.

[2] 朱阳光，刘瑞敏，黄琼桃. 基于深度神经网络的弱监督信息细粒度图像识别[J]. 电子测量与仪器学报，2020,34(2):115-122.
ZHU Y G, LIU R M, HUANG Q T. Fine-grained image recognition of weak supervisory information based on deep neural network[J]. Journal of Electronic Measurement and Instrumentation, 2020,34(2): 115-122.

[3] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas: IEEE,2016: 2921-2929.

[4] MAI J, YANG M, LUO W. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle: IEEE,2020: 8766-8775.

[5] ZHANG C L, CAO Y H, WU J. Rethinking the route towards weakly supervised object localization[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle: IEEE,2020: 13460-13469.

[6] XIE J, LUO C, ZHU X, et al. Online refinement of low-level feature based activation map for weakly supervised object localization [C]. IEEE/CVF International Conference on Computer Vision, Montreal: IEEE,2021: 132-141.

[7] MENG M, ZHANG T, TIAN Q, et al. Foreground activation maps for weakly supervised object localization[C]. IEEE/CVF International Conference on Computer Vision, Montreal: IEEE, 2021: 3385-3395.

[8] ZHAI W, WU P Y, ZHU K, et al. Background activation suppression for weakly supervised object localization and semantic segmentation [J]. International Journal of Computer Vision, 2024, 132(3): 750-775.

[9] WEI J, WANG Q, LI ZH, et al. Shallow feature matters for weakly supervised object localization[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal: IEEE,2021: 5993-6001.

[10] 何其霖，穆平安. VGG 网络与多特征融合的遮挡人脸检测[J]. 电子测量技术,2021,44(18):150-154.
HE Q L, MU P AN. Occlusion face detection based on VGG network and multi-feature fusion[J]. Electronic Measurement Technology, 2021, 44(18): 150-154.

[11] 周璇，易剑平. 基于优化 CBAM 改进 ResNet50 的异常行为识别方法[J]. 国外电子测量技术,2024,43(5):36-41.
ZHOU X, YI J P. Improved abnormal behavior recognition method of ResNet50 based on optimized CBAM [J]. Foreign Electronic Measurement Technology, 2024, 43(5): 36-41.

[12] 张志伟，武杰，边云，等. 基于迁移学习 Inception V3 网络模型鉴别胰腺浆液性囊腺瘤与黏液性囊腺瘤[J]. 中国医学影像技术,2023,39(6):876-879.
ZHANG ZH W, WU J, BIAN Y, et al. Transfer learning-based InceptionV3 network model for differentiating pancreatic serous cystic neoplasm and mucinous cystic neoplasm [J]. Chinese Medical Imaging Technology, 2023, 39(6): 876-879.

[13] 李冰锋，冀得魁，杨艺. 基于改进 MMAL 的细粒度图像分类研究[J]. 电子测量技术，2024，47(17)：172-179.
LI B F, JI D K, YANG Y, Analysis of fine-grained image classification through improved MMAL [J]. Electronic Measurement Technology, 2024, 47(17): 172-179.

[14] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115: 211-252.

[15] XIE J, XIANG J, CHEN J, et al. C^2AM: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Louisiana: IEEE, 2022: 989-998.

[16] WANG CH W, XU R T, XU SH B, et al. Exploring intrinsic discrimination and consistency for weakly supervised object localization[J]. IEEE Transactions on Image Processing, 2024, 33: 1045-1058.

[17] MENG M, ZHANG T, YANG W, et al. Diverse complementary part mining for weakly supervised object localization[J]. IEEE Transactions on Image Processing, 2022, 31: 1774-1788.

[18] CHOE J, SHIM H. Attention-based dropout layer for weakly supervised object localization[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition,Long Beach: IEEE, 2019: 2219-2228.

[19] BAE W, NOH J, KIM G. Rethinking class activation mapping for weakly supervised object localization[C]. European Conference on Computer Vision, Seattle: IEEE,2020:618-634.

[20] XU J, HOU J, ZHANG Y, et al. Cream: Weakly supervised object localization via class re-activation mapping[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Louisiana: IEEE, 2022: 9437-9446.

作者简介

阮皓皓，硕士研究生，主要研究方向为计算机视觉与目标检测。
E-mail:212207020023@home.hpu.edu.cn

李冰锋，博士，讲师，主要研究方向为迁移学习、计算机视觉与目标检测。
E-mail:libingfeng@hpu.edu.cn

李新伟（通信作者），博士，副教授，主要研究方向为对比学习、计算机视觉与目标检测。
E-mail:lixinwei@hpu.edu.cn

冀得魁，硕士研究生，主要研究方向为计算机视觉与目标检测。
E-mail:j1326553@163.com