计算机科学
COMPUTER SCIENCE

基于类别标签引导的协同显著性目标检测方法

李芳芳，孔雨秋，刘洋，李朋玥

引用本文

李芳芳，孔雨秋，刘洋，李朋玥. 基于类别标签引导的协同显著性目标检测方法[J]. 计算机科学，2026，53(1)：163-172.

LI Fangfang, KONG Yuqiu, LIU Yang , LI Pengyue. Co-salient Object Detection Guided by Category Labels [J]. Computer Science, 2026, 53(1): 163-172.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250100071

基于类别标签引导的协同显著性目标检测方法

李芳芳[1] 孔雨秋[2] 刘 洋[3] 李朋玥[1]
1 大连理工大学计算机科学与技术学院 辽宁 大连 116000
2 大连民族大学信息与通信工程学院 辽宁 大连 116000
3 大连理工大学未来技术学院/人工智能学院 辽宁 大连 116000
(lff@mail.dlut.edu.cn)

摘 要 像素级标签的获取耗时耗力,而图像级标签的获取要容易得多。目前,使用图像级标签解决协同显著性目标检测任务尚未得到深入探索。对此,运用一种两阶段方法解决弱监督协同显著分割任务,仅依赖图像级标签(即类别标签)进行模型训练。利用类别标签的语义信息,实现对协同显著目标的定位和分割。在第一阶段,提出了伪标签生成网络,利用类别标签作为监督信号,生成输入图像的显著图;在第二阶段,提出了协同显著分割网络,用上一阶段得到的显著图作为伪标签进行监督训练。此外,在训练过程中还引入了自我校正学习策略,以提升模型的性能。文中首次提出使用图像级标签来解决协同显著性目标检测问题,并在3个具有代表性的数据集上进行了实验验证,得到的结果证实了所提方法的有效性和可行性。

关键词:协同显著性目标检测;弱监督;自我校正学习策略;类别标签;两阶段方法
中图分类号 TP391

Co-salient Object Detection Guided by Category Labels

LI Fangfang[1],KONG Yuqiu[2],LIU Yang[3] and LI Pengyue[1]
1 School of Computer Science and Technology,Dalian University of Technology,Dalian,Liaoning 116000,China
2 School of Information and Communication Engineering,Dalian Minzu University,Dalian,Liaoning 116000,China
3 School of Future Technology,School of Artificial Intelligence,Dalian University of Technology,Dalian,Liaoning 116000,China

Abstract Acquiring pixel-level labels is laborious and time-consuming,whereas image-level labels can be obtained much more easily. However,the use of image-level labels for co-salient object detection(CoSOD) remains underexplored. This paper presents a two-stage approach for weakly supervised CoSOD,relying solely on image-level labels(class labels) for model training. By utilizing the semantic information of class labels,this approach enables the localization and segmentation of co-salient objects. In the first stage,a pseudo-label generation network is proposed to generate saliency maps for input images,supervised by class labels. In the second stage,a co-salient object segmentation network is trained using these saliency maps as pseudo-labels. A self-corrective learning strategy is also incorporated to enhance model performance. For the first time,this paper proposes using image-level labels based training approach for CoSOD. Experiments on three representative datasets demonstrate the effectiveness and feasibility of the proposed method.

Keywords Co-salient object detection,Weakly supervised,Self-corrective learning strategy,Class labels,Two-stage approach

1 引言

显著性目标检测(Salient Object Detection,SOD)旨在从单张图像中精确定位并分割出最显著、最吸引人注意的物体[1-2]。协同显著性目标检测(Co-salient Object Detection,CoSOD)进一步扩展了这一任务,目标是从一组相关图像中检测并分割出共同显著的目标[3-4]。换言之,CoSOD需要在多张图像中识别出共有目标,同时排除其他干扰物体的影响。因此,CoSOD比单图像的SOD任务更具挑战性。

现有的全监督CoSOD方法高度依赖于像素级真值标注,这种标注方式不仅耗费大量时间和人力,还显著增加了模型训练成本,同时也在一定程度上限制了模型在新场景中的应用和泛化能力。

在计算机视觉领域的诸多任务中,如CoSOD、语义分割和指称图像分割等,高昂的人工标注成本已成为一个亟待解决的问题。因此,许多研究者积极探索弱监督学习技术,以降低对精确标注的依赖。与全监督学习方法相比,弱监督学习方法的核心优势在于能够灵活利用多种形式的弱标签作为训

到稿日期:2025-01-13 返修日期:2025-05-08
基金项目:国家自然科学基金(62406055);辽宁省科技计划联合计划(技术攻关计划项目)(2024JH2/102600090)
This work was supported by the National Natural Science Foundation of China(62406055) and Liaoning Provincial Science and Technology Plan Joint Program(Technology R&D Program Project)(2024JH2/102600090).
通信作者:孔雨秋(yqkong@dlnu.edu.cn)

练监督信号，从而减少对大规模精确标注的依赖。同时，与无监督方法相比，弱监督学习在处理复杂和精细任务时表现出更高的效率和准确性。通过利用弱标签生成伪标签的形式，弱监督方法能够在缺乏大规模真实标注的情况下，有效训练和优化网络。这不仅大幅降低了标注成本，还显著提升了模型的泛化能力和鲁棒性。

尽管当前的研究取得了一定进展，但仍面临诸多挑战。一方面，部分现有方法[5-6]通过弱监督或无监督学习来解决协同显著性目标检测任务，但这些方法通常依赖于人为手动设计的特征来直接提取显著信息，这在一定程度上限制了模型的性能提升。另一方面，针对无监督显著性目标检测任务，当前研究多聚焦于优化传统 SOD 方法所提取的较为粗糙的显著性线索。例如，Nguyen 等[7]将这些显著性线索作为标签，用于训练多个深度网络，从而提升检测精度。Zhou 等[8]则通过挖掘高质量的显著性线索，训练出更为稳健的显著性检测器。此外，Xie 等[9]提出了一种基于对比学习的方法，利用类别不可知的激活映射网络生成类别无关的激活图，有效解决了弱监督语义分割任务中的相关问题。Xie 等[10]引入自然语言监督机制，通过自然语言描述来探索目标对象的完整内容，并排除无关背景区域的干扰，为显著性检测提供了新的思路。

针对上述问题，本文首次提出了一种基于类别标签引导的协同显著性目标检测模型。该方法的核心创新在于，引入类别标签作为监督信号，有效解决了现有方法在监督信号利用方面的局限性。与传统的像素级标注相比，类别标签的获取成本更低，且能够显著减少标注工作量和降低模型训练成本。同时，本文方法通过优化类别标签的使用方式，进一步训练出更为稳健的协同显著分割网络，从而实现更精准的检测。

本文设计了一种两阶段弱监督协同显著性目标检测模型框架。该框架充分利用了预训练的 CLIP[11]模型在图像与文本匹配以及共性语义信息提取方面的优势，通过文本标签生成像素级伪标签，实现高效的弱监督学习。

在第一阶段，利用 CLIP[11]模型强大的图像与文本对齐能力，监督模型训练并生成初始像素级伪标签。这些伪标签虽然不完整，但已包含关键的显著性信息，为后续学习提供了重要指导。例如，初始结果可能仅正确标记了部分显著区域，但这些正确部分为模型指明了学习方向。

具体而言，首先通过参数冻结的自监督的视觉 Transformer[12]对输入图像进行编码，将特征输入到分割头，使用类别标签进行监督。通过优化跨模态对比损失和单模态对比损失，得到初始的显著图，作为下一阶段的监督信号。

在第二阶段，采用标签自我校正学习策略，进一步优化协同显著分割网络的训练。图像特征依次通过图像内特征聚合模块和组特征增强模块后，得到增强的视觉特征；利用 CLIP[11]文本编码器得到类别标签的共性文本特征，使用对比损失将文本特征与目标特征对齐，突出协同显著目标表示。通过迭代优化，模型能够利用第一阶段生成的正确标记部分，逐步补全缺失信息，从而实现更精准的协同显著性目标检测。

实验结果表明，该模型在多个 CoSOD 任务的代表性数据集上取得了优异的性能。这表明即使像素级伪标签存在不完整性，其关键信息也足以引导模型向正确的方向学习，最终实现高效的协同显著性检测。

综上所述，为解决弱监督 COSOD 的问题，本文提出了一个两阶段模型框架，包括伪标签生成网络和协同显著分割网络。本文的主要贡献如下：

1）伪标签生成网络利用自监督视觉 Transformer 提取的特征，结合类别标签监督生成像素级伪标签，为第二阶段的分割网络提供监督信号。

2）在第二阶段协同显著分割网络模型的监督训练过程中，使用自我校正学习策略动态更新伪标签，进一步优化模型训练，提升检测精度。

3）在基准数据集 CoSOD3k[13]，Cosal2015[14]和 CoCA[15]上的测试结果表明，本文方法在性能上具有显著优势，有效提升了弱监督 CoSOD 任务的检测精度。

2 相关工作

2.1 协同显著性目标检测

CoSOD 是计算机视觉领域的一个重要任务分支，旨在从一组图像中识别出共同显著的目标。现有的方法主要分为以下几类。一些工作通过构建图结构来建模一组图像中像素之间的关系[16-17]，然后利用一致特征来挖掘协同显著目标。另一些工作则先采用额外的显著性目标检测提取显著目标，再进一步分析协同显著性信息[18-19]。此外，SAEF[20]首次提出采用无监督深度学习模型生成显著候选区域。其他一些工作[21-22]尝试在输入图像之间建立共享属性，以发现协同显著像素，并使用分类信息作为语义信息的补充。

尽管这些方法在协同显著性目标检测任务中表现出色，但它们大多依靠像素级监督或额外信息来学习协同显著特征。本文通过分析显著目标特征和背景特征之间的关系，以及学习显著目标与标签特征的对应关系，解决了 CoSOD 问题，无需额外使用 SOD 数据集或像素级标签。

2.2 对比学习

对比学习是一种通过对比正、负实例来学习有意义表征的方法。其核心假设是：在学习到的嵌入空间中，相似的实例特征应更接近，而不相似的实例特征应更远离。Ye 等[23]和 He 等[24]基于对比学习，用来自同一类的样本创建正对，不同类的样本构建负对，以进行模型的训练。

对比学习利用各种损失函数来建立学习过程的优化目标。这些损失函数在指导模型捕获重要表示和区分相似/不相似实例方面发挥着至关重要的作用。损失函数的设计需要结合具体的任务需求和数据特征，每个损失函数都旨在促进模型学习数据中的相似性表示和差异性表示。

2.3 弱监督学习

弱监督可以分为 3 类。第一类是不完全监督，即训练集中仅部分数据被标记，其余数据未标注。这种情况在多种任务中普遍存在。例如，在图像分类任务中，只有部分图像提供了人工标注的真值标签。第二类是不确切监督，即图像只有粗粒度的标签，而非精确的标注。为了减少标注工作量，许多计算机视觉任务尝试使用不同类型的弱标签，如边界框[25]、涂鸦[26]、点[27]等。另外，Shaharabany 等[28]使用图像类别标签提供的高级语义信息，将语义表示与视觉表示相结合，以解决开放词汇语义分割任务。第三类是不准确监督，即提供的标签并非总是准确的真值。

在本文模型的训练过程中，仅依赖类别标签进行监督，无需像素级标注。通过对比学习来生成初始的伪标签，这种方法不仅显著减少了标注工作量，还有效提升了模型的泛化能力。

2.4 预训练的视觉和语言模型

近年来，预训练模型在计算机视觉和自然语言处理领域取得了显著进展。这些模型通过对比学习[29]、度量学习[30]和自蒸馏[12]等技术，学习通用的特征表示，为下游任务提供了强大的特征提取器。

DINO[12]是基于自监督学习的视觉 Transformer 模型，通过在大规模无标签数据集上进行对比学习，能够学习具有通用性的视觉表征。Caron 等[12]的研究表明，DINO[12]的注意力机制能关注图像中的显著目标区域，这使其在图像检索和目标分割等任务中表现出色。近期的研究工作也利用 DINO[12]的最后一层输出特征作为全局感知表示，应用于目标检测等任务[31]。

CLIP[11]是一个具有里程碑意义的对比语言-图像预训练模型。该模型由图像编码器和文本编码器组成，利用简单的对比学习，在 400 万个图像-文本对的大型数据集上进行训练，展现了其在嵌入空间中对齐图像和文本模态的强大能力。CLIP[11]的成功激发了将预训练知识转移到多种下游任务的研究，如目标检测[32]、参考图像分割[33]等。

3 基于类别标签引导的协同显著性目标检测模型

3.1 概述

如图 1 所示，模型分为两个阶段。第一阶段是伪标签生成网络，利用类别标签作为监督训练，训练生成像素级伪标签；第二阶段是协同显著分割网络，使用第一阶段生成的伪标签进行监督训练，以实现协同显著性目标检测。

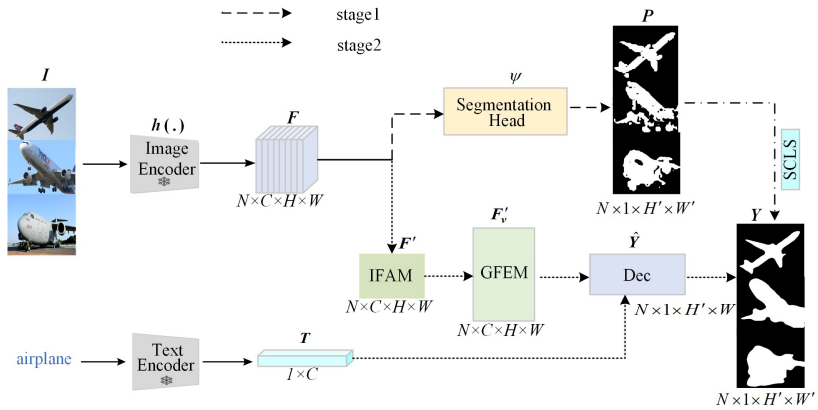


图 1 模型框架图

Fig. 1 Model structure diagram

在伪标签生成阶段，模型利用类别标签作为监督信号，通过跨模态对比损失和单模态对比损失学习生成初始的像素级伪标签。将图像输入伪标签生成网络，使用类别标签信息，得到初始的显著图。显著图作为伪标签监督第二阶段训练。

协同显著分割网络由特征编码器、图像内特征聚合模块(Intra-image Feature Aggregation Module，IFAM)、组特征增强模块(Group Feature Enhancement Module，GFEM)以及解码器(Decoder，Dec)组成，使用自我校正学习策略(Self-corrective Learning Strategy，SCLS)进行训练。在第二阶段将图像特征依次输入 IFAM 和 GFEM，增强图像视觉特征，然后对齐目标视觉特征和标签特征，得到最终的分割结果。

CoSOD 数据集包括带有类别标签的图像组。每组表示为(I,t)，其中 $I=\{I_i\}_{i=1}^n$，I_i是输入图像，t 是这组图像对应的标签。每组内的所有图像都包含类别相同、对应同一标签的显著目标。模型的目标是检测同一组每幅图像中共有的显著目标。

3.2 伪标签生成网络

如图 2 所示，给定图像组 $\boldsymbol{I}$，其中包含 n 张图像，使用自监督视觉 Transformer[12]作为编码器 I，记作 $h(\cdot)$。编码器 $h(\cdot)$将 $\boldsymbol{I}$ 映射为特征图 $\boldsymbol{F}=\{\boldsymbol{F}_i\}_{i=1}^n$，其中 $\boldsymbol{F}_i\in\mathbb{R}^{C\times H\times W}$，$C$ 表示通道数，H 和 W 表示空间维度。

首先，使用编码器[8] $h(\cdot)$从图像中提取特征。具体过程如下：将图像分成 $s\times s$ 个正方形块，每个方块对应一个 token，同时添加一个额外的可学习 token(Class Token，cls)；编码器对这些 token 进行处理，输出最后一个自注意力层的 key 特征作为图像特征 $\boldsymbol{F}$；基于提取的特征图 $\boldsymbol{F}$，通过分割头 $\psi(\cdot)$生成训练集初始的显著图 $\boldsymbol{P}=\{\boldsymbol{P}_i\}_{i=1}^n$，其中$\boldsymbol{P}_i\in\mathbb{R}^{H\times W}$。

$$\boldsymbol{F}=h(\boldsymbol{I}) \tag{1}$$

$$\boldsymbol{P}=\psi(\boldsymbol{F}) \tag{2}$$

具体来说，$\psi(\cdot)$由一个 3×3 卷积层、批量归一化层和 sigmoid 层组成。$\boldsymbol{P}_i$指示第 i 个样本的前景区域，相应的背景区域可以表示为$(1-\boldsymbol{P}_i)$。

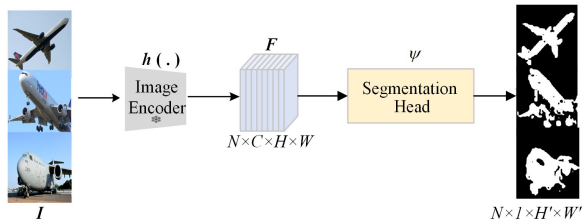


图 2 伪标签生成网络

Fig. 2 Pseudo-label generation network

3.2.1 跨模态对比学习

CLIP[11]模型的核心目标是通过对比学习，将图像和文本映射到同一个特征空间中，使得匹配的对的特征向量更接近，而不匹配的图像和文本对的特征向量更远离，从而实现图像和文本之间的跨模态对齐。因此，在训练过程中，利用类别标

签作为监督信息，结合跨模态对比学习损失，能够促使生成的显著图 $\boldsymbol{P}_i$ 有效地区分开目标区域和背景区域。

借助 CLIP[11] 模型卓越的跨模态特性，以类别标签作为文本线索，将其与对应图像在 CLIP[11] 构建的特征空间中建立联系。通过跨模态对比学习，生成能初步划分目标与背景区域的初始显著图。

跨模态对比损失计算基于 CLIP[11] 模型的图像编码器 $f_i(\cdot)$ 和文本编码器 $f_t(\cdot)$。具体步骤如下：首先，将 $\boldsymbol{P}_i$ 和 $(1-\boldsymbol{P}_i)$ 乘以 $\boldsymbol{I}_i$ 来分别屏蔽图像背景区域和目标区域；然后将屏蔽后的图像通过图像编码器分别映射为表示向量 $\boldsymbol{v}_i^{if}$ 和 $\boldsymbol{v}_i^{ib}$。

$$\boldsymbol{v}_i^{if}=f_i(\boldsymbol{P}_i\cdot\boldsymbol{I}_i) \tag{3}$$

$$\boldsymbol{v}_i^{ib}=f_i((1-\boldsymbol{P}_i)\cdot\boldsymbol{I}_i) \tag{4}$$

$(\boldsymbol{P}_i\cdot\boldsymbol{I}_i)$ 对应类别标签的文本为 t，通过文本编码得到文本表示向量为：

$$\boldsymbol{T}=f_t(\boldsymbol{t}) \tag{5}$$

如图 3 所示，给定第 i 个目标对象的表示 $\boldsymbol{v}_i^{if}$，其对应的文本表示为 $\boldsymbol{T}$。计算目标表示和文本表示之间的相似度，然后最小化目标区域表示和对应文本标签表示之间的损失 L_{ft}：

$$s_i^{ft}=sim(\boldsymbol{v}_i^{if},\boldsymbol{T}) \tag{6}$$

$$L_{ft}=-\sum_{i=1}^{n}\ln(s_i^{ft}) \tag{7}$$

其中，s_i^{ft} 代表 $\boldsymbol{v}_i^{if}$ 和 $\boldsymbol{T}$ 之间的余弦相似度。生成的显著图在 L_{ft} 的监督下会逐渐接近目标对象。

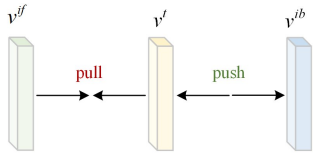


图 3 跨模态对比学习

Fig. 3 Cross-modal comparative learning

此外，为了抑制显著图 $\boldsymbol{P}_i$ 中的背景区域，使用背景区域和标签表示之间的对比损失 L_{bt} 来让 $\boldsymbol{P}_i$ 包含更多的目标内容。具体为：给定第 i 个目标对象的背景表示 $\boldsymbol{v}_i^{ib}$，其目标对应的文本表示为 $\boldsymbol{T}$，计算图像背景表示和文本表示之间的相似度，然后最小化背景区域和对应文本标签之间的损失 L_{bt}：

$$s_i^{bt}=sim(\boldsymbol{v}_i^{ib},\boldsymbol{T}) \tag{8}$$

$$L_{bt}=-\sum_{i=1}^{n}\ln(1-s_i^{bt}) \tag{9}$$

其中，s_i^{bt} 代表 $\boldsymbol{v}_i^{ib}$ 和 $\boldsymbol{T}$ 之间的余弦相似度。当 L_{bt} 最小化时，$\boldsymbol{I}_i\cdot(1-\boldsymbol{P}_i)$ 中保留了较少的目标像素，则 $\boldsymbol{I}_i\cdot\boldsymbol{P}_i$ 中包含了更多的目标像素。这使得显著图 $\boldsymbol{P}_i$ 中包含更完整的对象内容。

因此，跨模态对比学习损失记为：

$$L_1=L_{ft}+L_{bt} \tag{10}$$

3.2.2 单模态对比学习

由于训练过程中只有粗粒度图像级标签信息，仅利用文本驱动损失分割得到的显著图是不够准确的。考虑到对于给定图像，其前景和背景表示包含不同的语义信息，因此它们在特征空间中的距离应该很远。同时，对于属于同一图像组的图像之间，前景表示在特征空间中应该距离近，前景和组内其他背景表示应该距离远。因此，可以利用前景和背景表示之间的语义信息来区分前景对象区域。

如图 4 所示，$\boldsymbol{P}_i$ 可以将特征映射 $\boldsymbol{F}$ 分解为前景和背景特征表示，即 $\boldsymbol{v}_i^f$ 和 $\boldsymbol{v}_i^b$。对于第 i 个样本，$\boldsymbol{v}_i^f$ 和 $\boldsymbol{v}_i^b$ 可以表示为：

$$\boldsymbol{v}_i^f=\boldsymbol{P}_i\otimes\boldsymbol{F}_i^{\mathrm{T}} \tag{11}$$

$$\boldsymbol{v}_i^b=(1-\boldsymbol{P}_i)\otimes\boldsymbol{F}_i^{\mathrm{T}} \tag{12}$$

其中，$\boldsymbol{P}_i$ 和 $\boldsymbol{F}_i$ 被展平，即 $\boldsymbol{P}_i\in\mathbb{R}^{HW}$ 和 $\boldsymbol{F}_i\in\mathbb{R}^{C\times HW}$。其中 $\boldsymbol{v}_i^f\in\mathbb{R}^C$，$\boldsymbol{v}_i^b\in\mathbb{R}^C$；$\otimes$ 表示矩阵乘法；T 表示矩阵转置。

前景表示 $(\boldsymbol{v}_i^f,\boldsymbol{v}_j^f)$ 被视为正对，背景表示 $(\boldsymbol{v}_i^f,\boldsymbol{v}_j^b)$ 被视为负对。因此，单模态对比损失记为：

$$s_{ij}^{\mathrm{pos}}=sim(\boldsymbol{v}_i^f,\boldsymbol{v}_j^f) \tag{13}$$

$$s_{ij}^{\mathrm{neg}}=sim(\boldsymbol{v}_i^f,\boldsymbol{v}_j^b) \tag{14}$$

$$L_{\mathrm{pos}}=-\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\ln(s_{ij}^{\mathrm{pos}}) \tag{15}$$

$$L_{\mathrm{neg}}=-\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\ln(1-s_{ij}^{\mathrm{neg}}) \tag{16}$$

$$L_2=L_{\mathrm{pos}}+L_{\mathrm{neg}} \tag{17}$$

其中，s_{ij}^{pos} 和 s_{ij}^{neg} 分别是 $\boldsymbol{v}_i^f$ 和 $\boldsymbol{v}_j^f$，$\boldsymbol{v}_i^f$ 和 $\boldsymbol{v}_j^b$ 之间的相似度。

第一阶段整体训练损失为：

$$L_{\mathrm{stage1}}=L_1+L_2 \tag{18}$$

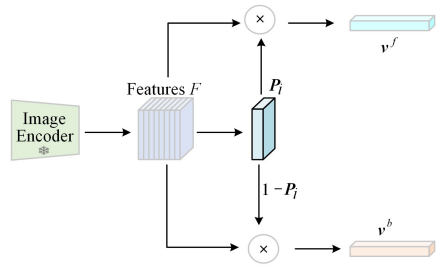


图 4 单模态对比学习

Fig. 4 Single-modal comparative learning

3.3 协同显著分割网络

本节将详细介绍第二阶段协同显著分割网络的结构，如图 5 所示，它包括 IFAM，GFEM，Dec，SCLS。

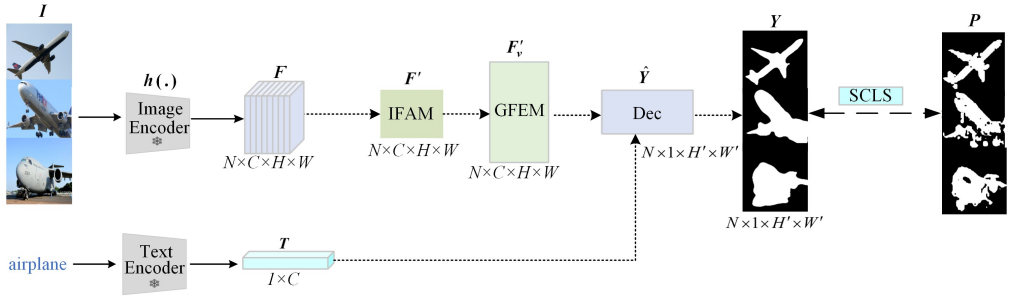


图 5 协同显著分割网络

Fig. 5 Co-salient segmentation network

3.3.1 IFAM

对于输入的图像特征，IFAM 通过自注意力机制增强特征的表示能力。如图 6 所示，对于输入的一组图像，其特征表示为 $\boldsymbol{F}\in\mathbb{R}^{N\times C\times W\times H}$，键(Key)、查询(Query)和值(Value)分别定义为 $\boldsymbol{K}=\boldsymbol{F},\boldsymbol{Q}=\boldsymbol{F},\boldsymbol{V}=\boldsymbol{F}$。

首先，对键 $\boldsymbol{K}$ 使用卷积核大小为 3×3 的组卷积，得到聚合键 $\boldsymbol{K}_1$。这一操作聚合了 $\boldsymbol{K}$ 的局部相邻特征之间的信息，提高了键捕获和代表信息的能力。然后，将聚合键 $\boldsymbol{K}_1$ 和查询特征 $\boldsymbol{Q}$ 进行级联，通过两次连续的 1×1 卷积 $\boldsymbol{W}_\sigma$ 和 $\boldsymbol{W}_\theta$ 得到注意力矩阵：

$$\boldsymbol{A}=\boldsymbol{W}_\sigma\boldsymbol{W}_\theta[\boldsymbol{K}_1,\boldsymbol{Q}] \tag{19}$$

具体来说，$\boldsymbol{A}$ 在每个空间位置的局部注意力矩阵是基于查询特征和聚合相邻信息的聚合键的特征来学习的，而不是基于孤立的查询、键对。通过这种方式，利用相关信息 $\boldsymbol{K}_1$ 的指导来增强自注意力学习。然后，根据注意力矩阵 $\boldsymbol{A}$，通过汇总自注意力中的所有值特征 $\boldsymbol{V}$ 来计算增强的图像特征 $\boldsymbol{F}'$：

$$\boldsymbol{F}'=\boldsymbol{W}_v[\boldsymbol{A}\times\boldsymbol{V}] \tag{20}$$

通过重复上述操作多次，特征表示能力逐步增强，最终输出的增强图像特征能够更好地捕捉图像中的关键信息。

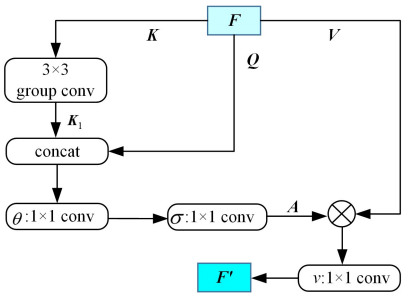


图 6 IFAM 框架

Fig. 6 Framework of IFAM

3.3.2 GFEM

同一类别的目标在其外观和特征上有相似之处，这一特性已被广泛应用于许多相关任务。

Transformer 的自注意力机制通过计算查询、键和值之间的关系来捕捉全局依赖。尽管它能够捕捉到共性特征，但这种机制较为通用，可能会引入一些与共性特征无关的信息。相比之下，GFEM 的设计目标明确，旨在处理具有相似外观和特征的图像组。通过显式建模图像组之间的共性关系，GFEM 专注于提取共性信息，因此在 CoSOD 任务中表现得更加高效和直接。

GFEM 通过计算同一组中所有图像之间的全局关系来为 CoSOD 任务引入先验信息。如图 7 所示，对于任意两个图像的特征 $\{\boldsymbol{F}_n',\boldsymbol{F}_m'\}\in\boldsymbol{F}'$，计算内积得到它们的像素级相关性：

$$\boldsymbol{S}_{(n,m)}=\omega(\boldsymbol{F}_n')^{\mathrm{T}}\tau(\boldsymbol{F}_m') \tag{21}$$

其中，ω,τ 表示卷积层。亲和矩阵 $\boldsymbol{S}_{(n,m)}\in\mathbb{R}^{HW\times HW}$ 捕获了图像对 (n,m) 的特征间关系。

以此类推，计算输入的一组图像间特征的亲和矩阵，并重塑成 $\boldsymbol{S}\in\mathbb{R}^{NHW\times N\times HW}$。从 $\boldsymbol{S}$ 中找到每幅图像的最大值 $\boldsymbol{A}\in\mathbb{R}^{NHW\times N}$，对 N 幅图像的所有最大值取平均，生成全局亲和注意力图 $\boldsymbol{A}_f\in\mathbb{R}^{NHW}$。使用 Softmax 操作对 $\boldsymbol{A}_f$ 进行归一化，并将其重塑为注意力图 $\boldsymbol{A}_s\in\mathbb{R}^{N\times 1\times H\times W}$。注意力图 $\boldsymbol{A}_s$ 与输入特征相乘，生成全局关联特征 $\boldsymbol{F}_e'=\boldsymbol{F}'\times\boldsymbol{A}_s$。最后，将输入的原始特征与全局关联特征进行融合，形成最终的增强特征表示：

$$\boldsymbol{F}_v'=\boldsymbol{F}'+\boldsymbol{F}_e' \tag{22}$$

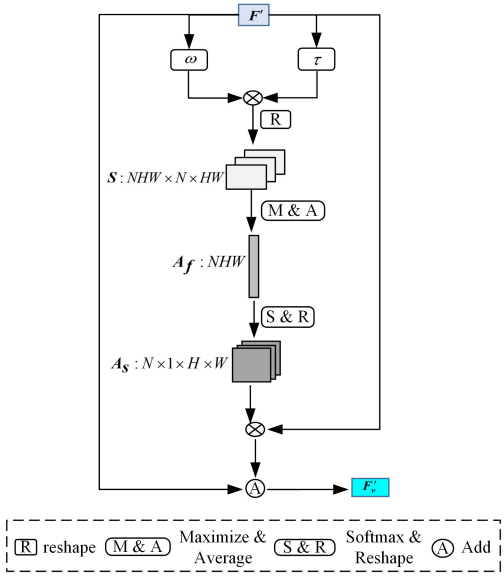


图 7 GFEM 框架

Fig. 7 Framework of GFEM

3.3.3 Dec

解码部分的具体操作如下：将视觉特征 $\boldsymbol{F}_v'$ 逐渐上采样到原始图像分辨率的二分之一，获得高分辨率的特征图 $\boldsymbol{F}_v'\in\mathbb{R}^{H/2\times W/2\times C}$。

采用图像投影和文本投影分别对增强的图像特征 $\boldsymbol{F}_v'$ 和标签特征 $\boldsymbol{T}$ 进行变换：

$$\boldsymbol{z}_v=\boldsymbol{F}_v'\boldsymbol{W}_v+\boldsymbol{b}_v \tag{23}$$

$$\boldsymbol{z}_t=\boldsymbol{T}\boldsymbol{W}_t+\boldsymbol{b}_t \tag{24}$$

其中，$\boldsymbol{z}_t\in\mathbb{R}^D,\boldsymbol{z}_v\in\mathbb{R}^{N\times D}$；$\boldsymbol{W}_v$ 和 $\boldsymbol{W}_t$ 是两个可学习的矩阵，用于将 $\boldsymbol{F}_v'$ 和 $\boldsymbol{T}$ 转换为相同的特征维度；$\boldsymbol{b}_v$ 和 $\boldsymbol{b}_t$ 是两个可学习的偏置。

给定一个转换后的标签特征和一组转换后的图像特征，使用一个文本到像素的对比损失，显式地将文本特征与相应的目标像素视觉特征对齐。该损失鼓励 $\boldsymbol{z}_t$ 与目标对应的 $\boldsymbol{z}_v$ 相似，与非目标对应的特征 $\boldsymbol{z}_v$ 不相似。通过点积运算计算相似度，因此，文本到像素的对比损失可以表示为：

$$\hat{Y}_i=\mathrm{sigmoid}(\boldsymbol{z}_t\cdot\boldsymbol{z}_v^i) \tag{25}$$

$$L_{vt}^i(\boldsymbol{z}_t,\boldsymbol{z}_v^i)=\begin{cases}-\log\hat{Y}_i, & i\in P\\ -\log(1-\hat{Y}_i), & i\in N\end{cases} \tag{26}$$

$$L_{vt}(\boldsymbol{z}_t,\boldsymbol{z}_v)=\frac{1}{|P\cup N|}\sum_{i\in P\cup N}L_{vt}^i(\boldsymbol{z}_t,\boldsymbol{z}_v^i) \tag{27}$$

其中，P 和 N 表示标签 Y 中分别为“1”和“0”的集合，$|P\cup N|$ 是基数，最终预测结果为 $\hat{Y}$。

第二阶段模型的总损失函数为：

$$L=L_{vt}(\boldsymbol{z}_t,\boldsymbol{z}_v)+L_2 \tag{28}$$

其中 L_2 为单模态对比损失。

3.3.4 SCLS

在测试结果中观察到，过度拟合第一阶段得到的伪标签

中的噪声会对训练稳健的协同显著分割网络产生负面影响。在训练的初期，模型准确率会逐渐升高，但到达一定阶段时，由于模型不断拟合不完全准确的伪标签，模型准确率会下降。

为了缓解伪标签噪声对模型训练的负面影响，SCLS 通过在训练过程中动态更新伪标签来提高模型的鲁棒性，具体如下。

初始阶段(前 50 轮训练)：在前 50 轮训练过程中，保持伪标签不变。这些伪标签是基于初始模型生成的，目的是在训练初期提供稳定的监督信号，使模型能够快速收敛并学习基本的特征和目标的语义信息。

后续阶段(50 轮之后)：从第 51 轮开始，使用模型在训练过程中得到的预测结果 $\hat{Y}$ 动态更新伪标签。通过这种方式，模型可以逐步纠正初始伪标签中的错误，提高伪标签的质量，进一步提升模型的性能。动态更新伪标签的公式如下：

$$\hat{\boldsymbol{Y}}^j = Det^j(\boldsymbol{F}, \boldsymbol{T}) \tag{29}$$

$$\boldsymbol{Y}^{j+1} = \begin{cases} \boldsymbol{P}, & j \leqslant 50 \\ \hat{\boldsymbol{Y}}^j, & j > 50 \end{cases} \tag{30}$$

其中，j 表示当前训练轮数；$Det(\cdot)$ 为协同显著分割网络中的所有操作。

4 实验细节与结果分析

4.1 数据集

本文使用 COCO-9k[34] 和 DUTS[35] 作为训练集。COCO-9k[34] 是 COCO[36] 数据集的一个子集，包含来自 COCO[36] 数据集的 65 组，共 9213 张图像。DUTS[35] 数据集分为 291 组，共包含 8250 张图像。

在测试阶段，本文使用 CoSOD3k[13]，Cosal2015[14] 和 CoCA[15] 3 个数据集。CoSOD3k[13] 数据集由 160 个组的 3316 张图像组成。Cosal2015[14] 数据集由 50 个组的 2015 张图像组成。CoCA[15] 由 80 个组的 1295 张图像组成。

这 3 个数据集广泛应用于协同显著性目标检测任务的评估，且在某些图像中包含多个显著目标，其目标的外观和尺度变化大，为模型的性能评估带来了巨大的挑战。

4.2 评价指标

采用以下 4 个评价指标来评估模型的性能：平均绝对误差(MAE)[37]、最大 F-measure($F_\beta^{\max}$)[38]、最大 E-measure($E_\phi^{\max}$)[39] 和 S-measure(S_α)[40]。

MAE[37]：像素级评估指标，用于测量预测图和真值图之间的绝对差异，无需二值化。其定义为：

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\hat{Y}(x,y) - GT(x,y)| \tag{31}$$

F-measure[38]：用于评估精度和召回率的加权调和平均值。显著图的输出使用不同的阈值进行二值化，以获得一组显著预测。对预测的显著图和真值图的精确度和召回率进行比较，将使用整个数据集的最佳阈值获得的最佳 F-measure 分数定义为 $F_\beta^{\max}$。F-measure[38] 的计算式为：

$$F_\beta = \frac{(1+\beta^2) Precision \times Recall}{\beta^2 (Precision + Recall)} \tag{32}$$

E-measure[39]：一种感知度量，用于从局部和全局角度评估预测图和真值图之间的相似性。其定义为：

$$E_\xi = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \boldsymbol{\phi}_\xi(x,y) \tag{33}$$

其中，$\boldsymbol{\phi}_\xi$ 表示增强的对齐矩阵。与 F-measure 类似，采用最大 E-measure($E_\phi^{\max}$)作为本文实验的评估指标。

S-measure[40]：预测图与其对应的真值图之间的结构相似性度量，可以在不进行二值化的情况下获得。S-measure 计算式为：

$$S_\alpha = \alpha \times S_o + (1-\alpha) \times S_r \tag{34}$$

其中，S_o 和 S_r 表示对象感知和区域感知的结构相似性，α 默认设置为 0.5。

在这些指标中，MAE 的值越小，表示预测图与真值图之间的差异越小；另外 3 个指标的值越大，表示预测图与真值图之间的相似度越高。

4.3 实施细节

本文采用基于 PyTorch 的深度学习框架进行实验，所有实验均在配备一块 NVIDIA GeForce RTX 4090 Ti 的服务器上完成。为了确保输入数据的一致性，训练图像的分辨率统一调整为 320×320 像素。

在第一阶段，使用 SGD 优化器训练模型 50 轮。初始学习率设置为 2×10^{-3}。权重衰减设为 5×10^{-4}。

在第二阶段，同样使用 SGD 优化器来训练模型 100 轮。初始学习率设置为 2×10^{-3}。权重衰减设为 5×10^{-4}。

在每个训练批次中，随机选择每组中的 N 个样本($N=8$)。

4.4 与其他方法的比较

本文主要与之前在公共 CoSOD 训练数据集上训练的先进无监督和弱监督方法进行比较，包括 UCCDGO[6]，SegSwap[41]，SAM CSD[42]，TokenCut[43]，DVFDVD[44]，FOUND[45]，TSD[8]，A2S-v3[46]，以及弱监督方法 WSLNet[47]，PSOD[27]，VRF[48]。

本文方法与其他方法的评估结果对比如表 1－表 3 所列。结果表明，本文方法仅利用类别标签即可在性能指标上取得显著提升，相较于其他现有方法，展现出更优越的性能表现。

表 1 在 CoSOD3k 数据集上与其他方法的比较

Table 1 Comparison with other methods on the CoSOD3k dataset

Algorithm	$F_\beta^{\max}$	$E_\phi^{\max}$	S^α	MAE
UCCDGO[6]	—	—	—	—
SegSwap[41]	0.560	0.705	0.608	0.177
SAM CSD[42]	0.703	0.810	0.723	0.108
TokenCut[43]	0.720	0.811	0.744	0.151
DVFDVD[44]	0.722	0.819	0.773	0.104
FOUND[45]	0.715	0.816	0.758	0.105
TSD[8]	0.653	0.755	0.664	0.136
A2S-v3[46]	0.645	0.753	0.666	0.138
WSLNet[47]	0.716	0.811	0.764	0.110
PSOD[27]	0.739	0.821	0.783	0.101
VRF[48]	0.754	0.830	0.789	0.096
Our Method	**0.785**	**0.859**	**0.801**	**0.072**

注：粗体表示最优值。

如表 1 所列，与 VRF[48] 相比，本文方法在 CoSOD3k[13] 数据集上的性能提升如下：MAE 指标下降了 25%，$F_\beta^{\max}$ 指标增长了 4.11%，$E_\phi^{\max}$ 指标增长了 3.49%，S_α 指标增长了 1.52%。

表 2 在 CoCA 数据集上与其他方法的比较

Table 2 Comparison with other methods on the CoCA dataset

Algorithm	$F_\beta^{\max}$	$E_\phi^{\max}$	S^α	MAE
UCCDGO[6]	—	—	—	—
SegSwap[41]	0.422	0.666	0.567	0.165
SAM CSD[42]	—	—	—	—
TokenCut[43]	0.476	0.704	0.627	0.167
DVFDVD[44]	0.422	0.592	0.581	0.223
FOUND[45]	0.424	0.624	0.596	0.171
TSD[8]	0.411	0.714	0.584	0.133
A2S-v3[46]	0.399	0.710	0.575	0.140
WSLNet[47]	0.424	0.639	0.597	0.181
PSOD[27]	0.447	0.680	0.606	0.184
VRF[48]	0.440	0.638	0.611	0.178
Our Method	**0.503**	**0.715**	**0.652**	**0.123**

注：粗体表示最优值。

表 3 在 Cosal2015 数据集上与其他方法的比较

Table 3 Comparison with other methods on the Cosal2015 dataset

Algorithm	$F_\beta^{\max}$	$E_\phi^{\max}$	S^α	MAE
UCCDGO[6]	0.758	—	0.751	—
SegSwap[41]	0.618	0.720	0.632	0.178
SAM CSD[42]	0.782	0.847	0.782	0.092
TokenCut[43]	0.805	0.857	0.793	0.139
DVFDVD[44]	0.777	0.842	0.809	0.091
FOUND[45]	0.781	0.851	0.800	0.091
TSD[8]	0.712	0.767	0.691	0.151
A2S-v3[46]	0.711	0.769	0.699	0.148
WSLNet[47]	0.781	0.844	0.805	0.098
PSOD[27]	0.822	0.874	0.836	0.081
VRF[48]	0.831	0.878	0.839	0.076
Our Method	**0.845**	**0.889**	**0.840**	**0.072**

注：粗体表示最优值。

如表 2 所列，与 TokenCut[43] 相比，本文方法在 CoCA[15] 数据集上的性能提升如下：$F_\beta^{\max}$ 指标增长了 5.67%，S_α 指标增长了 3.99%。与 TSD[8] 相比，本文方法的 MAE 指标下降了 7.52%。

如表 3 所列，与 VRF[48] 相比，本文方法在 Cosal2015[14] 数据集上的性能提升如下：MAE 指标下降了 5.26%，$F_\beta^{\max}$ 指标增长了 1.68%，$E_\phi^{\max}$ 指标增长了 1.25%。

综上所述，本文方法在各项评估指标上均优于其他现有方法，这表明本文方法在分割性能上具有显著优势。实验结果充分验证了本文方法的有效性和可行性。

图 8 展示了基于 CoSOD3k[13] 数据集的不同方法生成的显著性图，用于进行定性比较。“axe”组用于评估模型检测小物体的能力。本文方法在检测精度方面优于其他方法，能够准确识别小的协同显著目标，其他方法可能会遗漏或专注于其他类别的物体。在“elephant”组中，图像包含复杂的显著目标边缘。本文方法能够较为精准地检测并分割协同显著目标，相对完整地保留了目标的轮廓，而其他方法则难以实现。在“crutch”组中，纤细目标的分割是关键挑战。本文方法能够较为准确地分离出纤细目标与背景，而其他方法甚至可能无法正确识别，将纤细目标与周围区域混淆。总体而言，本文方法能够更有效地发现组内公共信息，并准确区分组间信息，从而在复杂场景下表现出色。

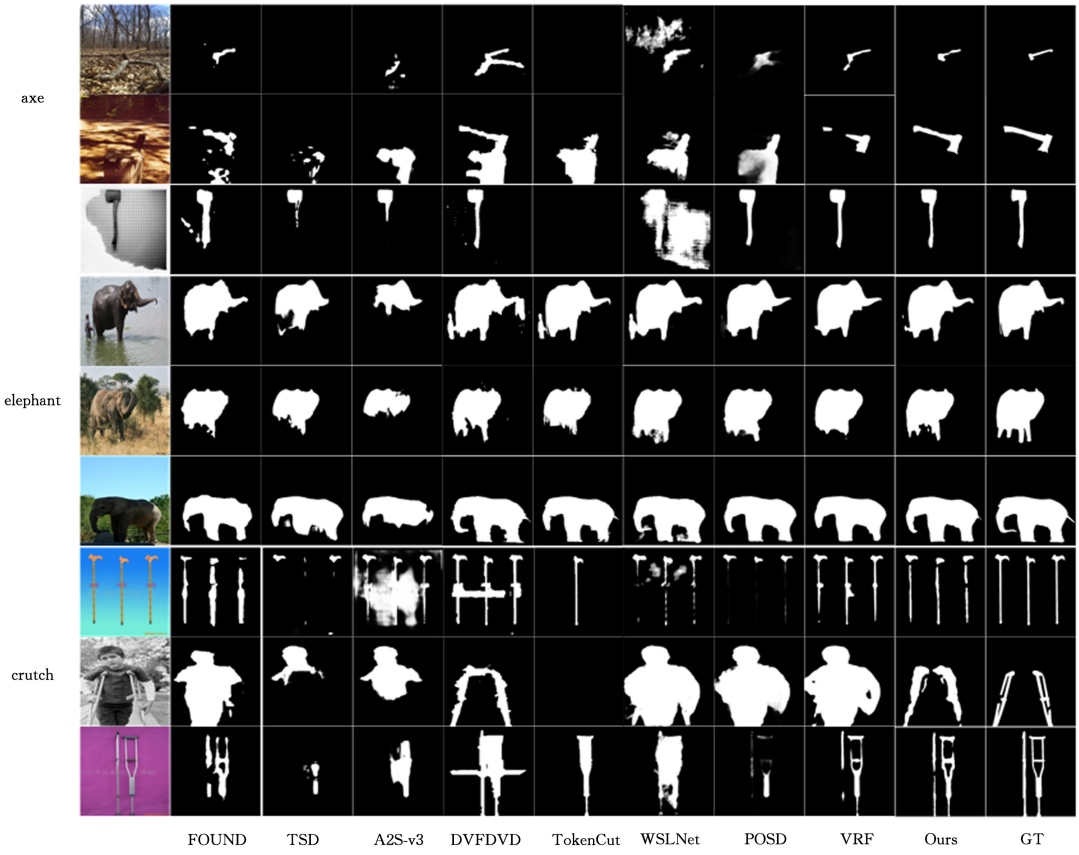


图 8 可视化结果示例

Fig. 8 Example of visualization results

4.5 消融实验

本文采用两阶段网络结构。通过对其进行一系列消融实验，验证了模型的有效性和合理性。

表 4 列出了在 CoSOD3k[13] 数据集上对第一阶段的损失函数进行的消融实验结果，可以看出，当使用 L_1 和 L_2 两个损失的组合时，得到的评估结果最优。

表 4 CoSOD3k 数据集上损失函数的消融实验

Table 4 Ablation study of the loss function on CoSOD3k dataset

Loss		$F_\beta^{\max}$	$E_\phi^{\max}$	S_α	MAE
L_1	L_2				
√		0.440	0.545	0.520	0.301
	√	0.608	0.692	0.611	0.132
√	√	**0.707**	**0.799**	**0.721**	**0.113**

注：粗体表示最优值。

表 5 列出了在 CoSOD3k[13] 数据集上对第二阶段网络进行的消融实验结果。实验以仅包含编码器和解码器的网络作为基线模型，每次向该基线模型中添加一个模块，以评估每个模块对模型性能的影响。消融实验的结果证明了模型设计的有效性。具体来说：

与基线网络相比，添加 IFAM 后，MAE 指标下降 6.32%，S_α 提高 1.31%。

与基线网络相比，添加 GFEM 后，MAE 指标下降 6.32%，$E_\varphi^{\max}$ 提高 1.09%，S_α 提高 1.57%。

与基线网络相比，使用 IFAM 和 GFEM 的组合后，MAE 指标降低 11.58%，$F_\beta^{\max}$ 提高 2.01%，$E_\phi^{\max}$ 提高 1.93%，S_α 提高 2.88%。

在使用 IFAM 和 GFEM 的基础上，进一步使用 SCLS，MAE 下降 14.29%，$F_\beta^{\max}$ 提高 3.02%，$E_\phi^{\max}$ 和 S_α 分别提高 1.78% 和 1.78%。

为了进一步验证第二阶段网络中各个模块的有效性，进行了额外的消融实验。具体来说，在第二阶段网络中分别去掉 IFAM 和 GFEM，观察各个指标的变化。

与完整模型相比，去掉 GFEM 后，MAE 指标上升了 8.33%，$F_\beta^{\max}$ 下降了 1.15%。

与完整模型相比，去掉 IFAM 后，MAE 指标上升了 8.33%，$F_\beta^{\max}$ 下降了 0.76%，$E_\phi^{\max}$ 下降了 0.47%。

表 5 不同模块在 CoSOD3k 数据集上的消融实验

Table 5 Ablation study of different modules on CoSOD3k dataset

Algorithm	$F_\beta^{\max}$	$E_\phi^{\max}$	S^α	MAE
Base	0.747	0.828	0.765	0.095
Base+IFAM	0.751	0.835	0.775	0.089
Base+GFEM	0.754	0.837	0.777	0.089
Base+IFAM +GFEM	0.762	0.844	0.787	0.084
Base+IFAM +SCLS	0.776	0.853	0.794	0.078
Base+GFEM +SCLS	0.779	0.855	0.795	0.078
Base+IFAM +GFEM +SCLS	**0.785**	**0.859**	**0.801**	**0.072**

注：粗体表示最优值。

综上，消融实验结果表明，模型中的每个模块均对性能有提升作用。特别是，IFAM 和 GFEM 的组合显著降低了 MAE 指标，提高了 $F_\beta^{\max}$，$E_\phi^{\max}$ 和 S_α 指标。进一步引入 SCLS 后，模型性能得到了进一步优化，这充分证明了本文方法的有效性。

结束语 本文提出了一种新颖的两阶段弱监督协同显著性目标检测模型，旨在解决 CoSOD 任务。模型仅使用图像级标注作为监督，首先生成初始伪标签，并将其作为监督信号训练第二阶段网络，以实现更精确的协同显著性目标分割。实验结果表明，即便仅依赖低成本、易获取的类别标签，本文方法在分割性能上相较于之前的方法仍有显著提升。这些结果验证了本文方法的可行性和有效性。

未来的研究将集中在以下几个主要方向。1)提升模型的泛化能力：通过引入数据增强技术，如随机裁剪、旋转、颜色抖动等，增强模型对不同场景和变化的适应性，从而提高其鲁棒性。2)优化网络结构：将现有的两阶段网络改进为单阶段网络，进一步简化模型架构，从而提高训练效率，同时确保模型性能保持不变或进一步提升。3)模型轻量化：通过剪枝技术等方法去除模型中冗余的参数，减少模型的大小和计算量，以提高模型的效率和可扩展性。

参 考 文 献

[1] YU S, ZHANG B, XIAO J, et al. Structure-Consistent Weakly Supervised Salient Object Detection with Local Saliency Coherence[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2021: 3234-3242.

[2] TANG L, LI B, ZHONG Y, et al. Disentangled High Quality Salient Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2021: 3580-3590.

[3] ZHENG P, FU H, FAN D P, et al. GCoNet+: A Stronger Group Collaborative Co-Salient Object Detector [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10929-10946.

[4] LI L, HAN J, ZHANG N, et al. Discriminative Co-Saliency and Background Mining Transformer for Co-Salient Object Detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2023: 7247-7256.

[5] TSAI C C, LI W, HSU K J, et al. Image Co-Saliency Detection and Co-Segmentation via Progressive Joint Optimization [J]. IEEE Transactions on Image Processing, 2018, 28(1): 56-71.

[6] HSU K J, TSAI C C, LIN Y Y, et al. Unsupervised CNN-based co-saliency detection with graphical optimization[C]// Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 485-501.

[7] NGUYEN T, DAX M, MUMMADI C K, et al. Deepusps: Deep robust unsupervised saliency prediction via self-supervision[J]. Advances in Neural Information Processing Systems, 2019, 32: 204-214.

[8] ZHOU H, QIAO B, YANG L, et al. Texture-guided saliency distilling for unsupervised salient object detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2023: 7257-7267.

[9] XIE J, XIANG J, CHEN J, et al. C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2022: 989-998.

[10] XIE J, HOU X, YE K, et al. Clims: Cross language image matching for weakly supervised semantic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition. Los Alamitos: IEEE Computer Society, 2022:4483-4492.

[11] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the International Conference on Machine Learning. New York: PMLR, 2021:8748-8763.

[12] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society, 2021:9650-9660.

[13] FAN D P, LI T, LIN Z, et al. Re-thinking co-salient object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(8):4339-4354.

[14] ZHANG D, HAN J, LI C, et al. Detection of co-salient objects by looking deep and wide[J]. International Journal of Computer Vision, 2016, 120:215-232.

[15] ZHANG Z, JIN W, XU J, et al. Gradient-Induced Co-Saliency Detection[C]// Proceedings of the Computer Vision — ECCV 2020: 16th European Conference. Berlin: Springer, 2020: 455-472.

[16] WEI L, ZHAO S, BOURAHLA O E F, et al. Deep group-wise fully convolutional network for co-saliency detection with graph propagation[J]. IEEE Transactions on Image Processing, 2019, 28(10):5052-5063.

[17] HU R, DENG Z, ZHU X. Multi-scale graph fusion for co-saliency detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2021: 7789-7796.

[18] JIN W, XU J, CHENG M M, et al. Icnet: Intra-saliency correlation network for co-saliency detection[J]. Advances in Neural Information Processing Systems, 2020, 33:18749-18759.

[19] ZHANG N, HAN J, LIU N, et al. Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society, 2021:4167-4176.

[20] TSAI C C, HSU K J, LIN Y Y, et al. Deep co-saliency detection via stacked autoencoder-enabled fusion and self-trained cnns[J]. IEEE Transactions on Multimedia, 2019, 22(4):1016-1031.

[21] FAN Q, FAN D P, FU H, et al. Group collaborative learning for co-salient object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2021:12288-12298.

[22] ZHANG K, DONG M, LIU B, et al. DeepACG: Co-Saliency Detection via Semantic-aware Contrast Gromov-Wasserstein Distance[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2021:13703-13712.

[23] YE M, ZHANG X, YUEN P C, et al. Unsupervised Embedding Learning via Invariant and Spreading Instance Feature[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2019:6210-6219.

[24] HE K, FAN H, WU Y, et al. Momentum Contrast for Unsupervised Visual Representation Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2020:9729-9738.

[25] DAI J, HE K, SUN J. BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation [C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society, 2015: 1635-1643.

[26] ZHANG J, YU X, LI A, et al. Weakly-Supervised Salient Object Detection via Scribble Annotations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2020: 12546-12555.

[27] GAO S, ZHANG W, WANG Y, et al. Weakly-supervised salient object detection using point supervision[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022:670-678.

[28] SHAHARABANY T, TEWEL Y, WOLF L. What is Where by Looking: Weakly-Supervised Open-World Phrase-Grounding without Text Inputs[J]. Advances in Neural Information Processing Systems, 2022, 35:28222-28237.

[29] CHEN T, KORNBLITH S, NOROUZI M, et al. A Simple Framework for Contrastive Learning of Visual Representations [C]// Proceedings of the International Conference on Machine Learning. New York: PMLR, 2020:1597-1607.

[30] CHEN X, HE K. Exploring Simple Siamese Representation Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2021:15750-15758.

[31] SIMEONI O, PUY G, VO H V, et al. Localizing objects with self-supervised transformers and no labels[C]// Proceedings of the 32nd British Machine Vision Conference 2021. London: BMVA, 2021:310.

[32] FENG C, ZHONG Y, JIE Z, et al. Promptdet: Expand your detector vocabulary with uncurated images[C]// Computer Vision-ECCV 2022: 17th European Conference. Berlin: Springer-Verlag, 2022:701-717.

[33] WANG Z, LU Y, LI Q, et al. CRIS: CLIP-Driven Referring Image Segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2022:11686-11695.

[34] WEI L, ZHAO S, EL FAROUK BOURAHLA O, et al. Group-wise deep co-saliency detection[C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. AAAI, 2017:3041-3047.

[35] WANG L, LU H, WANG Y, et al. Learning to Detect Salient Objects with Image-Level Supervision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017:136-145.

[36] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// Proceedings of the Computer Vision-ECCV 2014: 13th European Conference. Berlin: Springer, 2014:740-755.

[37] CHENG M M, WARRELL J, LIN W Y, et al. Efficient Salient Region Detection with Soft Image Abstraction[C]// Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society, 2013: 1529-1536.

[38] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2009: 1597-1604.

[39] FAN D P, GONG C, CAO Y, et al. Enhanced-alignment measure for binary foreground map evaluation[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. AAAI, 2018: 698-704.

[40] FAN D P, CHENG M M, LIU Y, et al. Structure-Measure: A New Way to Evaluate Foreground Maps[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society, 2017: 4548-4557.

[41] SHEN X, EFROS A A, JOULIN A, et al. Learning Co-segmentation by Segment Swapping for Retrieval and Discovery[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2022: 5082-5092.

[42] LIU Y, LI T, WU Y, et al. Self-supervised image co-saliency detection[J]. Computers and Electrical Engineering, 2023, 105: 108533.

[43] WANG Y, SHEN X, HU S X, et al. Self-Supervised Transformers for Unsupervised Object Discovery using Normalized Cut [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2022: 14543-14553.

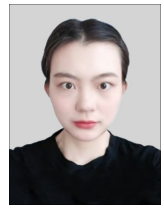
[44] AMIR S, GANDELSMAN Y, BAGON S, et al. Deep vit features as dense visual descriptors[J]. arXiv: 2112.05814, 2021.

[45] SIMÉONI O, SEKKAT C, PUY G, et al. Unsupervised object localization: Observing the background to discover objects[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2023: 3176-3186.

[46] YUAN Y, LIU W, GAO P, et al. Unified unsupervised salient object detection via knowledge transfer[C]// Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. 2024: 1616-1624.

[47] CONG R, QIN Q, ZHANG C, et al. A weakly supervised learning framework for salient object detection via hybrid labels[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(2): 534-548.

[48] XU B, LIANG H, GONG W, et al. A visual representation-guided framework with global affinity for weakly supervised salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(1): 248-259.



LI Fangfang, born in 2000, postgraduate, is a member of CCF (No. Z3690G). Her main research interests include deep learning and computer vision.

KONG Yuqiu, born in 1992, Ph.D, associate professor, is a member of CCF (No. C0194M). Her main research interests include computer vision and cross-media analysis.

（责任编辑：喻藜）