基于稀疏对比自蒸馏的 RGB-D 显著性物体检测

于洋洋[1]，吴敦全[1*]，陈程立诏[1,2]

（1.青岛大学计算机科学技术学院，山东 青岛 266071；2.中国石油大学（华东）计算机科学与技术学院，山东 青岛 257061）

摘　要：近年来，红绿蓝-深度（RGB-D）显著性目标检测技术取得了巨大进展，性能得到显著提高。然而，该技术依赖于复杂且资源密集的网络架构，无法应用于资源受限环境。虽然，轻量级网络在尺寸和速度上有所改善，但往往以牺牲性能为代价。为了克服上述限制，提出了一种新颖的轻量化解决方案，以实现网络参数的精简和性能的提升。本文提供了一种有效的通用训练策略，提出稀疏对比自蒸馏技术。该技术旨在对现有的 RGB-D 显著性检测模型进行压缩和加速，同时增强模型性能。本文方法由两个关键技术组成：稀疏自蒸馏和对抗性对比学习。稀疏自蒸馏排除显著性检测模型中的非必要参数，同时保留关键参数，从而实现更高效和有效的显著性预测。而对抗性对比学习通过纠正潜在错误，进一步完善自蒸馏过程，以提高模型的整体性能。在 NJUD、NLPR、LFSD、ReDWeb-S 和 COME15K 等基准数据集上的实验结果显示，与现有 SOTA（State-of-The-Art）方法相比，本文方法能够产生更为准确的显著性检测结果。此外，本文方法与现有 SOTA 轻量级 RGB-D 显著性检测模型的比较结果进一步证实了本文方法在不牺牲性能的前提下能够在模型尺寸减小和性能提升之间实现平衡。

关键词：红绿蓝-深度（RGB-D）显著性目标检测；稀疏自蒸馏；对比学习

中图分类号：TP391.4　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0069648

RGB-D Saliency Object Detection Based on Sparse Contrastive Self-Distillation

YU Yangyang[1], WU Dunquan[1*], CHEN Chenglizhao[1,2]

(1.College of Computer Science and Technology, Qingdao University, Qingdao 266071, Shandong, China;
2.College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 257061, Shandong, China)

【Abstract】 In recent years, Red Green Blue-Depth (RGB-D) saliency object detection technology has made significant progress with a notable improvement in performance. However, the dependency of the RGB-D saliency object detection technology on complex and resource-intensive architectures has limited its application in resource-constrained environments. Although lightweight networks have improved in terms of size and speed, they often come at the cost of sacrificing performance. To address this challenge, an innovative lightweight solution is proposed, which overcomes these limitations by streamlining network parameters and enhancing performance. An effective and universal training strategy and a sparse contrastive self-distillation technology are proposed, in order to compress and accelerate existing RGB-D saliency detection models while improving model performance. This strategy is primarily composed of two key technologies: sparse self-distillation and adversarial contrastive learning. Sparse self-distillation focuses on eliminating unnecessary parameters in saliency detection models while retaining key parameters, thereby achieving more efficient and effective saliency prediction. Adversarial contrastive learning, on the other hand, aims to correct potential errors, further refining the self-distillation process and improving the overall performance of the model. Experimental results on benchmark datasets such as NJUD, NLPR, LFSD, ReDWeb-S, and COME15K demonstrate that, compared to other State of The Art (SOTA) methods, our method can produce more accurate saliency detection results. Furthermore, comparison results of the proposed method with existing SOTA lightweight RGB-D saliency detection models further confirm that our method can achieve a balance between model size reduction and performance enhancement, without sacrificing performance.

【Key words】 Red Green Blue-Depth (RGB-D) saliency object detection; sparse self-distillation; contrast learning

0　引言

显著性目标检测是一项核心的计算机视觉技术，旨在识别图像或视频中的显著目标，在图像检索、目标跟踪、场景理解、图像分割等应用中发挥了关键作用。随着深度学习和计算机视觉技术的快速发展，显著性目标检测技术不断取得进步。传统的显著性目标检测通常基于红绿蓝（RGB）通道信息，通过对比纹理、形状等特征来识别显著目标。然而，在复杂背景或低对比度场景下，单纯依靠 RGB 通道

收稿日期：2024-03-25　　**修回日期：**2024-06-13

基金项目：国家自然科学基金（62172246）；山东省高等学校青年创新团队发展计划（2021KJ062）。

通信作者 E-mail：[*] wudunquan@163.com

的信息可能不足以准确检测显著目标。因此，越来越多的研究者开始关注 RGB-深度(RGB-D)显著性目标检测，即在 RGB 通道的基础上，加入深度信息，以提高检测的准确性。

在深度学习的大背景下，利用前沿技术如 Transformer[1]和图卷积网络[2]，研究者可以更好地融合 RGB 和深度信息，从而提高检测的准确性。在 RGB-D 显著性目标检测的研究中，出现了许多高性能模型。这些模型通常使用多模态融合技术，将 RGB 和深度信息结合起来，以提高显著性目标检测的准确性。为了获得更好的检测性能，许多模型采用了复杂的网络架构，如多层卷积神经网络、Transformer 等。这些复杂架构通常需要更多的计算资源和存储空间。然而，这些高性能模型[3-5]的复杂性也成为它们的弊端，因为它们往往需要庞大的参数量和巨大的计算资源。为了克服资源密集型模型的局限性，一些研究者提出了轻量级模型。部分研究方法[6]尝试采用如 MobileNet 这类较浅的网络结构作为基础，旨在构建更加轻量化的模型。这种做法往往会导致模型大小与性能之间的失衡。此外，一些轻量级解决方案[7-9]通过知识蒸馏技术，让简单的“学生”模型从复杂的“教师”模型学习，虽然这样做可以在减小模型尺寸和提高运行速度上取得成效，但性能却有所下降。

基于 RGB-D 显著性目标检测技术面临的挑战，本文提出了一种创新的稀疏对比自蒸馏技术，旨在减少 RGB-D 显著性目标检测模型的参数，同时显著提升模型性能。该技术通过稀疏自蒸馏的方式，优化了学生模型中的参数分布，去除非必要的参数，同时保留了对提高检测性能至关重要的参数。此外，为了避免自蒸馏过程中知识传递不准确，本文提出了对抗性对比学习策略，通过与负样本的对比学习，有效分辨出可能导致误差的因素，进而纠正蒸馏过程中的偏差，进一步提升模型性能。本文方法的一个重要创新之处在于保持学生模型与教师模型在蒸馏阶段的模型深度一致，这避免了知识稀释，确保了信息的全面有效传递。

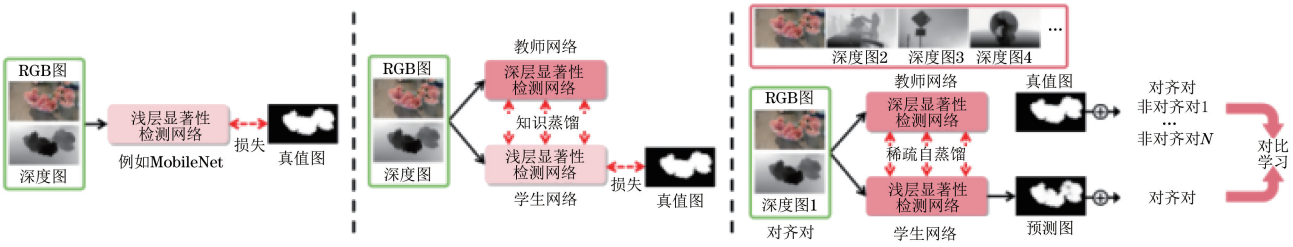
总结本文的贡献，可以归纳为以下几点：1)提出了稀疏对比自蒸馏算法，使得显著性目标检测模型更加轻量化，同时显著提升了模型性能；2)通过选择性地保留关键参数并消除冗余参数，本文算法不仅简化了模型结构，同时实现了效率与性能的最佳平衡；3)利用对抗性对比学习策略，有效预防和纠正自蒸馏阶段可能的不准确性，从而进一步提高了显著性检测的准确性。

1 相关工作

1.1 RGB-D 显著性目标检测

传统的 RGB-D 显著性目标检测方法[10-11]主要依赖手工制作的特征，并利用各种显著性先验，如对比度、图像背景和目标信息。这些方法通常结合来自不同层的深度线索，或使用中心暗通道先验。然而，它们忽略了 RGB 和深度模态之间的固有差异，这可能导致结果不够可靠。相比之下，随着深度学习的兴起，基于卷积神经网络(CNN)的方法成为图像显著性检测领域的主流方法。在这些方法中，基于融合的方法[2,12]对 RGB-D 显著性检测做出了重大贡献，显著提高了性能。这些方法大致可以分为输入融合[13]、后期融合[14]和中期融合[15-16]方法。输入融合方法直接结合 RGB 图像和深度图，形成四通道 RGB-D 输入。后期融合方法使用传播技术迭代细化初始显著图。中期融合方法采用双流结构转换跨模态特征并融合跨层次特征。

然而，这些工作也存在一些缺点。高性能模型通常具有复杂的网络架构，对计算资源的要求较高，这限制了其在资源受限环境中的应用。复杂模型通常需要大量参数，增加了训练和推理的成本。本文引入了稀疏自蒸馏技术，通过精简模型的参数量来降低计算资源需求，同时保持高性能。为了纠正自蒸馏过程中的潜在错误，本文采用了对抗性对比学习策略，进一步提升了显著性检测的准确性，如图 1 所示(彩色效果见《计算机工程》官网 HTML 版，下同)。



(a)传统的显著性检测方法 (b)基于知识蒸馏的显著性检测方法 (c)稀疏对比自蒸馏的显著性物体检测方法

图 1 现有 RGB-D 显著性目标检测的不同轻量化方法

Fig. 1 Different lightweight methods for existing RGB-D saliency object detection

1.2 自蒸馏

知识蒸馏[17]是一种将信息从一个较大的“教师”网络转移到一个较小的“学生”网络的技术。图 2 阐述了本文提出的稀疏对比自蒸馏法的整体框架。LIN 等[18]提出的方法通过低秩分解结合知识转移来实现网络的加速与压缩。近年来，自蒸馏方法的提出，开辟了在网络内部进行知识蒸馏的新路径。例如，LIU 等[19]提出的 MetaDistiller 采用阶段式自蒸馏，将知识从网络的深层阶段转移到较浅层阶段。自蒸馏技术可以从复杂的“教师”模型向简单的“学生”模型传递知识，达到模型压缩和加速的效果。尽管自蒸馏为多种任务提供了新的视角，但在 RGB-D 显著性目标检测领域，这一方法的探索还相对有限。传统的蒸馏方法在蒸馏过程中可能会导致知识流失，进而影响模型性能。一些蒸馏方法依赖强大的教师模型，这可能限制学生模型的性能提升。不同于 MetaDistiller，本文方法在教师和学生网络间采用相同的步骤确保深度一致性，减少知识稀释的风险，并通过对比损失来明确构建知识，而非依赖标签生成器产生软目标。为了减少对教师模型的依赖，本文通过稀疏编码策略减少模型的冗余参数，同时保持关键参数，优化了学生模型的结构。

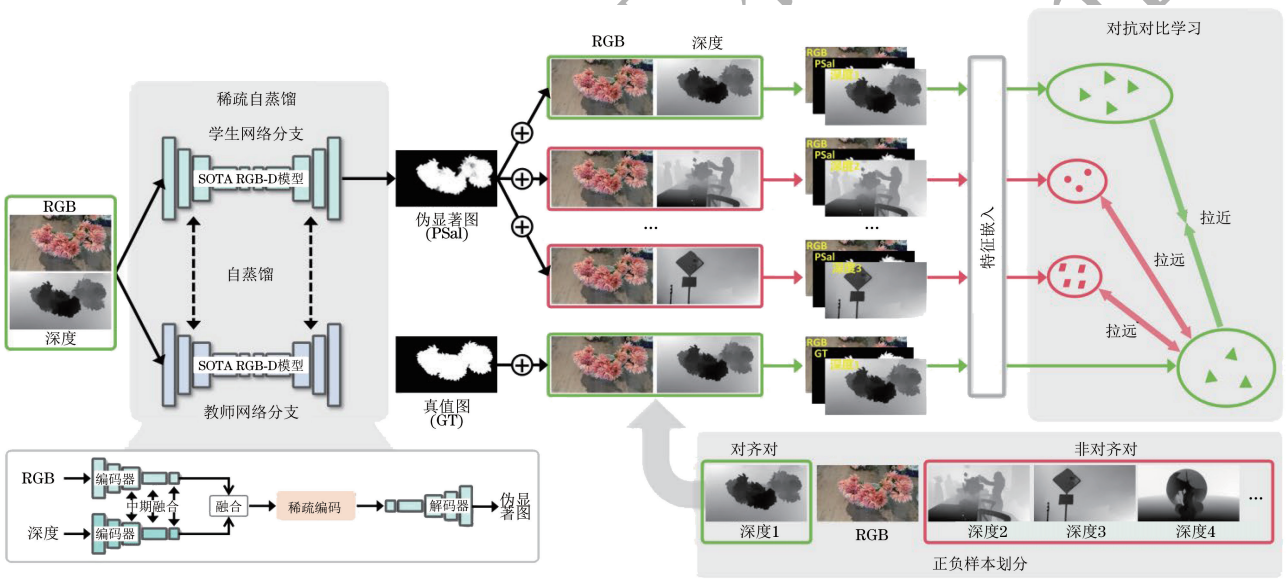


图 2 稀疏对比自蒸馏方法的整体流程结构

Fig. 2 Whole process structure of sparse contrastive self-distillation method

1.3 对比学习

对比损失在自监督学习领域中应用广泛[20-21]，主要目的是在表示空间中拉近锚点与正样本之间的距离，同时推远锚点与负样本之间的距离。先前的研究如文献[22]已经将对比学习应用于更为高级的任务中。例如，TIAN 等[23]提出的对比蒸馏框架旨在捕捉图像分类中结构化表示间的相关性。对比学习通过拉近正样本与锚点之间的距离，同时推远负样本，优化了特征学习过程。然而，对比学习依赖正负样本的划分，可能存在过拟合的风险，尤其是在样本分布不均衡的情况下。此外，对比学习的计算复杂度较高，可能需要更多的计算资源。

受到 WU 等[24]工作的启发，本文方法采用了对抗性对比学习策略，通过构建正负样本来指导伪显著图的生成过程，进一步提高显著性检测的准确性。为了降低过拟合风险，本文采用了稀疏自蒸馏和对抗性对比学习相结合的方法，确保模型的稳健性和准确性。

2 本文方法

2.1 整体结构

本文提出稀疏对比自蒸馏法的整体框架，该方法的核心在于两项技术革新：稀疏自蒸馏与对抗性对比学习。稀疏自蒸馏可实现在训练过程中剔除冗余参数的同时保持关键参数，降低了模型的复杂度。这种优化不仅提升了模型的效率，还增强了其对潜在显著目标（即伪显著图）的预测能力。基于上述的伪显著图，对抗性对比学习进一步解决了稀疏自蒸馏可能引入的问题，显著提高了显著性检测的性能。

2.2 稀疏自蒸馏

目前，RGB-D 显著性检测方法[3-5]普遍采用复杂的深度学习架构，这些架构包含大量的冗余参数，导致模型较为复杂。虽然知识蒸馏技术已被尝试用于模型简化，但传统的模型简化方法[7-9]常会导致模型性能降低。这主要是因为：在使用一个复杂的“教师”模型（如 ResNet50）来训练一个简单的“学生”显

著性模型(如 AlexNet-16)时,蒸馏过程中知识的流失往往会导致模型性能损失。此外,传统轻量化技术的效果很大程度上依赖于选用一个强大的教师模型来引导学生模型,这限制了优化模型性能的潜力。为此,本文提出了一种创新的稀疏自蒸馏方法。该方法旨在通过结合自蒸馏和稀疏编码技术,对现有的显著性检测模型进行精炼和增强。稀疏自蒸馏通过更高效地利用模型自身的知识,减少了对外部教师模型的依赖,降低了性能退化的风险,形成了更轻量且高效的显著性检测模型。

2.2.1 自蒸馏

在自蒸馏的训练阶段,教师显著性模型专注于保留对模型性能至关重要的参数,从而将关键的显著性检测知识传递给学生模型。如图 2 所示,在自蒸馏过程中,一个经过充分训练的最优 RGB-D 显著性模型(如 SEFF[25])扮演教师的角色。学生模型在结构上与教师模型相同,并从零开始训练。这一结构等价性地确保了教师与学生模型在蒸馏过程中的深度一致性,最小化了知识稀释的风险。

教师模型通过处理 RGB 和深度图像,将显著性相关的知识通过不同的特征层 $F_i(i\in\{1,2,3,4,5\})$传递给学生模型。这种知识转移采用 Kullback-Leibler(KL)散度损失来高效地传递信息。在自蒸馏过程中,针对每个特征层 F_i,学生模型利用深度感知的显著性洞见进行优化,从而提升了其预测准确性。这一过程不仅针对模型的整体结构进行优化,而且通过精确地调整和优化关键特征的传递,确保了学生模型能够更准确地预测显著图。这一过程可表示为:

$$L_{SD}=L_K(F_i^{(T)},F_i^{(S)}) \tag{1}$$

$$F_i^{(T)}=T(I_{RGB},I_D) \tag{2}$$

$$F_i^{(S)}=S(I_{RGB},I_D) \tag{3}$$

式中:T 和 S 分别表示教师显著性检测器和学生显著性检测器;L_K 为 Kullback-Leibler 损失;L_{SD} 为提出的自蒸馏损失;I_{RGB} 和 I_D 分别表示 RGB 图像和深度图像。

2.2.2 稀疏编码

自蒸馏过程使得学生显著性模型能够保留那些对性能有显著影响的参数,但同时也暴露出了模型中存在大量冗余参数的问题,这些参数增加了模型的复杂度。为了解决这一问题,本文采用了稀疏编码策略,以在学生模型中去除这些冗余参数。如图 2 所展示的,稀疏编码策略通过选择性减少显著性检测器的融合特征来实现对学生模型的剪枝,而这些融合特征通常构成了模型参数的主要部分。

稀疏编码策略针对 RGB 和深度图像组合得到的融合特征 $F_f\in\mathbb{R}^{H\times W\times C}$($H$、$W$、$C$ 分别代表特征的高度、宽度和通道数)进行操作。通过一个随机函数 $R(\cdot)$在这些特征中随机选择位置,并利用零化函数 $Z_n(\cdot)$在选定位置将特征权重设置为零。此过程在每个特征层级 i($i\in\{1,2,3,4,5\}$)上进行,从而产生更加稀疏的融合特征 F'_f。稀疏编码在每一层提供了一个清晰的框架来减少冗余,同时保持显著性模型的完整性和效率。这一过程可表示为:

$$F'_f=Z_n[R(F_f)]\to 0,\ n=\alpha\times H\times W\times C \tag{4}$$

式中:$Z_n(\cdot)$是零化函数,用于将权重设置为 0;α 是比例因子;n 是设置为零的权重位置数,由 α 控制。

2.3 对抗性对比学习

在稀疏自蒸馏的基础上,本文进一步介绍了对抗性对比学习这一新颖的方法。尽管稀疏自蒸馏依赖于使用现有的 SOTA 显著性模型作为教师来指导学生显著性检测器,但在蒸馏过程中,来自教师网络的误差有可能被传递给学生网络,从而影响学生网络的性能。受对比学习的启发,该方法旨在通过使网络输出接近正样本和远离负样本来改进特征学习过程。因此,为了降低在稀疏自蒸馏过程中传递错误信息的风险,本文采用了对抗性对比学习策略。

表 1 所展示的是目标基线模型在其原始训练数据集(包括 NJUD、NLPR)上的重新训练。其中,表中参数 S、F、M 分别表示 S-measure、F-measure、平均绝对误差,FPS 表示每秒帧率。训练过程中,这一策略主要利用学生显著性检测器生成的伪显著图来精炼显著性检测。通过构造正确的正样本和误导的负样本,以对比学习的方式,引导伪显著图朝向更准确的显著图转变。正负样本的划分基于训练数据集,其中对齐的 RGB 和深度图像对作为正样本,不匹配的深度图像视为负样本。例如,一个对齐的正样本可能包括 RGB 图像 1 和相应的深度图像 1,而非对齐的负样本可能包括 RGB 图像 1 和其他任何深度图像 2、3、…、$N-1$ 的组合。这些样本经过通道连接过程与伪显著图结合,进而通过一个预训练的深度网络(如 VGG 或 Transformer)的特征嵌入模块的处理,提取特征嵌入。随后,三元组损失函数被应用于这些嵌入,以优化伪显著图的生成过程。本文的对比学习过程如下:

$$L_{CL}=\mathrm{Tri}(e_1,e_2,e_3) \tag{5}$$

$$e_1=\mathrm{FE}[\mathbb{C}(A_{PS},P_{sal})] \tag{6}$$

$$e_2=\mathrm{FE}[\mathbb{C}(M_{NS},P_{sal})] \tag{7}$$

$$e_3=\mathrm{FE}[\mathbb{C}(A_{\mathrm{PS}},G_{\mathrm{T}})] \tag{8}$$

式中：L_{CL} 表示对比学习损失；Tri(·)为三元组损失函数；FE(·)为特征嵌入模块；$\mathbb{C}$ 代表通道连接操作；A_{PS} 代表对齐的正样本；M_{NS} 代表非对齐的负样本；P_{sal} 代表伪显著图；e_1、e_2、e_3 表示进行特征嵌入之后的中间向量。

表 1 与现有 SOTA RGB-D 显著性检测模型的定量比较

Table 1 Quantitative comparison with existing SOTA RGB-D saliency detection models

模型	NJUD			NLPR			LFSD			ReDWeb-S			COME15K			FPS/(帧·s^{-1})	参数量/MB
	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$		
A2dele	0.871	0.874	0.051	0.898	0.878	0.028	0.834	0.832	0.077	0.641	0.603	0.160	0.787	0.795	0.092	120	57.32
BTSNet	0.952	0.961	0.023	0.919	0.898	0.029	0.824	0.803	0.089	0.701	0.708	0.135	0.848	0.832	0.067	23	23.22
SPNet	0.958	0.965	0.016	0.930	0.923	0.020	0.771	0.771	0.118	0.709	0.712	0.129	0.852	0.838	0.063	12	12.42
SSLSOD	0.950	0.962	0.026	0.900	0.862	0.032	0.838	0.821	0.083	0.710	0.706	0.136	0.850	0.832	0.068	52	52.41
CIRNet	0.945	0.953	0.029	0.931	0.922	0.024	0.795	0.789	0.118	0.703	0.708	0.132	0.849	0.835	0.064	31	30.91
HiDANet	0.952	0.962	0.018	0.931	0.925	0.021	0.771	0.760	0.121	0.736	0.745	0.132	0.846	0.832	0.067	9	9.42
PopNet	0.961	0.970	0.015	0.928	0.921	0.020	0.844	0.828	0.079	0.776	0.784	0.125	0.851	0.834	0.063	10	10.06
SEFF	0.962	0.972	0.017	0.934	0.927	0.019	0.870	0.869	0.062	0.802	0.815	0.121	0.855	0.841	0.061	12	12.41
Ours(CIRNet)	0.948	0.956	0.027	0.933	0.926	0.021	0.797	0.792	0.121	0.715	0.721	0.129	0.856	0.848	0.061	45	22.56
Ours(HiDANet)	0.956	0.965	0.017	0.934	0.926	0.022	0.774	0.762	0.115	0.750	0.759	0.126	0.851	0.837	0.066	21	7.73
Ours (SEFF)	0.964	0.969	0.016	0.937	0.932	0.017	0.876	0.873	0.058	0.824	0.829	0.118	0.863	0.855	0.057	26	8.67

2.4 损失函数

本文方法的整体损失 L 由对比学习损失 L_{CL}、自蒸馏损失 L_{SD} 和二元交叉熵损失 L_{BCE} 构成，可表示为：

$$L=L_{\mathrm{BCE}}+\lambda_1 L_{\mathrm{SD}}+\lambda_2 L_{\mathrm{CL}} \tag{9}$$

式中：λ_1 和 λ_2 为平衡 L_{BCE} 和 L_{CL} 的超参数，按照经验分别设置为 0.9 和 0.9。

3 实验

3.1 数据集和评估指标

在 5 个广泛使用的公共基准数据集(NJUD[26]、NLPR[27]、LFSD[28]、COME15K[29] 和 ReDWeb-S[30])上评估了本文模型的有效性。

NJUD 包括 2 003 对不同分辨率的立体图像对。在这些图像对中，1 400 张图像作为训练集，100 张图像作为验证集，其余作为测试集。

NLPR 包括 11 种室内和室外场景的 1 000 张图片。其中，650 张图像作为训练集，50 张图像作为验证集，其余 300 张图像作为测试集。

LFSD 是一个相对较小的测试数据集，包含 100 张带有深度信息的图像，这些图像是通过 Lytro 光场相机捕获的。

COME15K 是最新提出的目前规模最大的数据集，包含了 15 625 对样本，原始图像收集自 Holopix 平台，深度图像通过光流估计算法得到，给出了很多挑战性场景下的显著性检测场景。

ReDWeb-S 是基于 ReDWeb 构建的数据集，ReDWeb 中的深度图像是通过最新的光流估计算法 FlowNet2.0 以及后处理得到。

采用 S、F、M 进行定量评价[31-32]。

$$S=\alpha\times S_0+(1-\alpha)S_{\mathrm{r}} \tag{10}$$

式中：S_0 和 S_{r} 分别为对象感知结构相似性和区域感知结构相似性；$\alpha=0.5$ 为协调参数。

$$F=(1+\beta^2)\frac{P\times R}{\beta^2 P+R} \tag{11}$$

式中：P 和 R 分别为准确率和召回率；$\beta^2=0.3$，用来权衡准确率和召回率。

$$M=\frac{1}{H\times W}\sum_{x=1}^{H}\sum_{y=1}^{W}|S(x,y)G(x,y)| \tag{12}$$

式中：S 和 G 分别是预测的显著性图和地面真值。

3.2 实验设置

使用 PyTorch 和 NVIDIA GeForce 3090 GPU 实现了本文方法。与文献[2,33]相同，所提方法在一个综合训练集上进行训练，包括 NJUD[26] 数据集的 1 400 个样本和 NLPR[27] 数据集的 650 个样本，输入分辨率调整为 352×352 像素，学习率设置为 10^{-4}。本文采用二元交叉熵损失进行监督。本文模型的批处理大小和总训练轮数与目标基线模型相同。

3.3 实验结果与分析

为了证明本文方法在常见显著性目标检测场景中的有效性，将本文方法与 3 个目标基线模型在 6 个广泛使用的 RGB-D 显著性目标检测数据集上进行对比，定性结果展示在图 3 中，其中包括使用本文方法重新训练的 CIRNet、HiDANet 和 SEFF 版本。

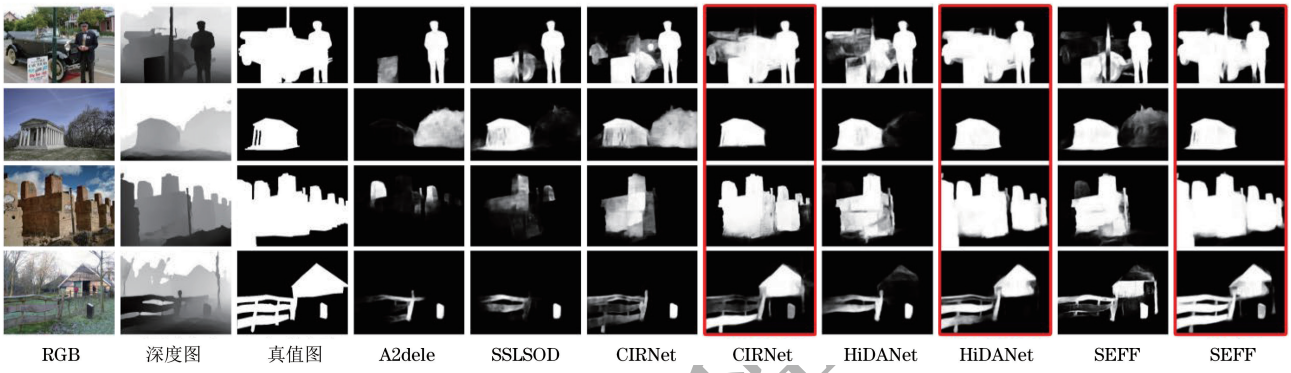


图 3 本文方法与最具代表性的 SOTA 模型的可视化比较

Fig. 3 Visualization comparison between method presented in this paper and most representative SOTA models

将本文方法与最新的 8 个 SOTA RGB-D 显著性目标检测模型进行了比较，如图 3 所示。这些 SOTA 模型包括 A2dele[7]、BTSNet[3]、SPNet[5]、SSLSOD[4]、CIRNet[34]、HiDANet[35]、PopNet[36]和 SEFF[25]。为了进行公平的比较，本研究采用了代码实现中的默认参数设置或者是作者提供的显著图。此外，选取了 3 个最新的顶尖 SOTA 模型，即 CIRNet、HiDANet 和 SEFF，作为本文研究的目标教师-学生基线模型。本文将所提出的方法应用于这些模型，旨在实现模型性能的进一步提升。

表 1 实验结果充分显示，采用本文方法的所有目标SOTA模型均实现了在原始版本基础上的显著性能提升。具体来说，本文方法使 CIRNet、HiDANet 和 SEFF 的 S 指标分别平均提高了 1.4、1.7 和 2.1 百分点，这一结果强有力地证明了所提稀疏对比自蒸馏方法在提升 SOTA 模型性能方面的有效性。

图 3 证明与其他 SOTA 方法相比，本文方法能够在保持较低速度的同时生成更为准确的显著性预测结果。此外，将本文方法与当前 SOTA 的轻量级 RGB-D 显著性目标检测模型进行了比较。表 2 的结果显示，尽管本文方法在速度上略低于其他轻量级模型，但它成功地在速度与性能之间实现了良好的平衡。

表 2 其他 SOTA 轻量化 RGB-D 显著性检测模型的定量比较(基于 SEFF)

Table 2 Quantitative comparison of other SOTA lightweight RGB-D saliency detection models (based on SEFF)

模型	LFSD			ReDWeb-S			COME15K			FPS/(帧·s^{-1})	参数量/MB
	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$		
A2dele	0.834	0.832	0.077	0.641	0.603	0.160	0.787	0.795	0.092	120	57.32
MobileSal	0.847	0.841	0.078	0.652	0.610	0.162	0.814	0.804	0.087	65	26.00
PGAR	0.833	0.831	0.093	0.656	0.632	0.161	0.836	0.819	0.079	48	62.00
DFM-Net	0.865	0.864	0.072	0.705	0.786	0.136	0.851	0.826	0.071	70	8.50
Ours	0.876	0.873	0.058	0.824	0.829	0.118	0.863	0.855	0.057	26	8.67

3.4 组件评估

本文对方法中使用的主要组件进行了全面评估，以验证其有效性，并在表 3 中展示了相应的定量结果，基线模型(SEFF)位于第一行。其中，组件

表 3 主要组件的定量评估

Table 3 Quantitative evaluation of major components

Components					LFSD			ReDWeb-S			COME15K		
SE	SD	NSD	FE	ACL	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$
×	×	×	×	×	0.870	0.869	0.062	0.802	0.815	0.121	0.855	0.841	0.061
×	√	√	√	√	0.871	0.871	0.062	0.806	0.819	0.122	0.857	0.842	0.061
√	×	√	√	√	0.871	0.870	0.061	0.809	0.821	0.121	0.858	0.844	0.058
√	√	×	√	×	0.873	0.872	0.063	0.814	0.824	0.120	0.860	0.846	0.062
√	√	√	√	×	0.875	0.872	0.060	0.819	0.828	0.119	0.862	0.853	0.059
√	√	√	√	√	0.876	0.873	0.058	0.824	0.829	0.118	0.863	0.855	0.057

SE、SD、NSD、FE 以及 ACL 分别表示稀疏编码、自蒸馏、负样本划分策略、前端特征提取模块以及对抗性对比学习策略。通过比较第 2 行与第 6 行，明显看出稀疏编码在提升 RGB-D 显著性目标检测模型训练效果中的重要作用。例如，在 COME15K 测试集上，S 指标从 0.857 提升至 0.863，这强调了本文自蒸馏技术的有效性。自蒸馏对于 RGB-D 显著性目标检测模型训练的贡献可通过比较第 3 行与第 6 行看出，其中 COME15K 数据集的 M 指标从 0.058 降至 0.057，彰显了本文自蒸馏方法的功效。第 4 行和第 5 行的比较进一步证明了负样本划分策略的有效性，COME15K 数据集中的 M 指标从 0.062 降至 0.059。第 5 行与第 6 行的对比展现了本文对抗性对比学习策略的优势，例如，对于 ReDWeb-S 数据集，S 指标从 0.819 提升至 0.824，突显了本文对抗性对比学习技术的有效性。

3.5 消融实验

3.5.1 负样本比例的不同选择

通过广泛的消融研究，本文确定了训练集中负样本（不对齐的 RGB 图和深度图像对）的最优比例。实验结果表明，100%比例在所有度量标准中均呈现出最佳结果，因此本文选择使用训练集中所有剩余的深度图构建负样本。

3.5.2 不同的特征嵌入方法

为评估特征嵌入[式(6)～式(8)]的效果，本文使用了 ResNet、VGG、Transformer 和 GNN 等不同方法进行了详尽的消融研究。如表 4 和表 5 结果显示，与 ResNet 相比，VGG 模型在速度上略显劣势。而尽管 Transformer 在某些数据集（如 LFSD 和 ReDWeb-S）上的表现优于 ResNet，但这是以显著降低的 FPS 为代价。此外，GNN 在所有测试方法中表现最差。因此，综合考量效率和效果，本文选择 ResNet 作为特征嵌入架构。

表 4 选择不同负样本比例的结果比较(基于 SEFF)

Table 4 Comparison of results for different negative sample proportions (based on SEFF)

负样本比例/%	LFSD			COME15K		
	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$
10	0.872	0.870	0.061	0.857	0.842	0.060
40	0.873	0.872	0.020	0.858	0.843	0.060
70	0.875	0.873	0.018	0.861	0.849	0.059
100	0.876	0.873	0.058	0.863	0.855	0.057

表 5 不同特征嵌入方法的结果比较(基于 SEFF)

Table 5 Comparison of choice of different feature embedding methods (based on SEFF)

模型	LFSD			COME15K			FPS/(帧·s^{-1})
	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$	
ResNet	0.876	0.873	0.058	0.863	0.855	0.057	26
VGG	0.878	0.877	0.057	0.864	0.856	0.057	23
Transformer	0.878	0.873	0.057	0.861	0.853	0.055	7
GNN	0.873	0.872	0.059	0.860	0.851	0.059	10

3.5.3 比例因子 α 的有效性

比例因子 α 的大小对模型速度和参数量有显著影响。根据表 6，当 α 值较小时，FPS 和参数量较多，这意味着剪枝较少，模型计算速度较小且参数量较大。当 α 值增加到 0.3，FPS 上升，参数量减少，表明模型变得更加轻量化，计算效率提高。然而，α 值过大（超过 0.3）会导致剪枝过多，FPS 虽稍有上升，但参数量显著减少，降低了模型性能。因此，通

表 6 比例因子 α 的消融研究(基于 SEFF)

Table 6 Ablation of scale factor α (based on SEFF)

指标	LFSD			COME15K			FPS/(帧·s^{-1})	参数量/MB
	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$		
0.1	0.874	0.871	0.059	0.863	0.853	0.058	23	11.15
0.3	0.876	0.873	0.058	0.863	0.855	0.057	26	8.67
0.5	0.873	0.872	0.059	0.861	0.850	0.059	27	6.19
0.7	0.871	0.869	0.057	0.860	0.849	0.059	39	3.71

过调整比例因子 α，能在模型性能和计算速度之间取得平衡，确保模型既轻量化又高效。表 6 的数据支持了这一观点，显示出 α 值的变化对模型速度和参数量的影响。

3.5.4 与传统知识蒸馏方法的比较

为了进一步证明所提方法的有效性，分别对比了传统知识蒸馏（学生网络基于 FasterNet-Tiny 和 FasterNet-Small[37]）、无稀疏编码和稀疏自蒸馏方法的模型性能（教师网络均基于 SEFF），具体实验结果如表 7 所示。实验结果表明，稀疏自蒸馏方法在所有基准数据集上均表现出了更优异的性能，且参数量显著减少。例如，在 LFSD 数据集上，相比基线模型，稀疏自蒸馏方法的 S 指标从 0.870 提升至 0.876，参数量减少了 30%以上，同时在其他评估指标上也有显著提升。虽然相比传统蒸馏方法（Tiny），所提稀疏自蒸馏方法的参数量减少的幅度略小，但模型性能有大幅提升。因此，相比传统蒸馏方法，稀疏自蒸馏能够在减少参数量的同时有效地增强模型性能。

表 7 与传统蒸馏方法的比较（基于 SEFF）

Table 7 Comparison with traditional distillation methods (based on SEFF)

模型	LFSD			COME15K			参数量/MB
	$S\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$F\uparrow$	$M\downarrow$	
基线模型	0.870	0.869	0.062	0.855	0.841	0.061	12.41
传统蒸馏（Tiny）	0.845	0.851	0.080	0.843	0.826	0.072	7.64
传统蒸馏（Small）	0.852	0.859	0.078	0.846	0.830	0.070	9.78
无稀疏编码	0.871	0.871	0.062	0.857	0.842	0.061	12.41
有稀疏编码	0.876	0.873	0.058	0.863	0.855	0.057	8.67

3.6 局限性

本文方法存在两个潜在的局限性：一是稀疏自蒸馏和对抗性对比学习的有效性极度依赖于 RGB-D 数据集的质量和图像对齐度，不一致或低质量的数据集可能会削弱模型学习效果，导致模型性能下降；二是对抗性对比学习组件对正负样本选择的敏感度可能引入偏差或过拟合，尤其是在样本分布无法代表实际场景多样性时。

4 结束语

本研究通过引入稀疏对比自蒸馏方法，为 RGB-D 显著性检测领域提供了一种轻量级且高效的方案，使得在资源受限的设备上部署先进显著性检测模型成为可能。通过巧妙结合稀疏自蒸馏与对抗性对比学习，本文方法不仅实现了模型简化，即通过减少非必要参数，并通过减小知识传递过程中的潜在误差，提升了模型准确性。广泛的定量评估结果证明本文方法能在大幅减小模型尺寸的同时，维持乃至提升模型性能，本文方法在某些案例中效果尤为显著。

未来，计划探索更为复杂的参数优化与模型压缩策略，旨在进一步缩小模型体积，而不牺牲准确度，以在边缘设备上实现即时性能，拓展实时显著性检测在资源有限环境下的应用前景。此外，未来将致力于提高模型的通用性和鲁棒性，以适应更广泛的应用场景，为后续研究和应用提供坚实基础。

参考文献

[1] LIANG J, CAO J, FAN Y, et al. VRT: a video restoration transformer[J]. IEEE Transactions on Image Processing, 2024, 33: 2171-2182.

[2] SONG M, SONG W, YANG G, et al. Improving RGB-D salient object detection via modality-aware decoder[J]. IEEE Transactions on Image Processing, 2022, 31: 6124-6138.

[3] ZHANG W, JIANG Y, FU K, et al. BTS-Net: bi-directional transfer-and-selection network for RGB-D salient object detection[C]//Proceedings of 2021 IEEE International Conference on Multimedia and Expo. Washington D. C., USA: IEEE Press, 2021: 1-6.

[4] ZHAO X, PANG Y, ZHANG L, et al. Self-supervised pretraining for RGB-D salient object detection[EB/OL]. (2021-01-29)[2024-02-25]. https://arxiv.org/abs/2101.12482.

[5] ZHOU T, FU H, CHEN G, et al. Specificity-preserving RGB-D saliency detection[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2021: 4681-4691.

[6] WU Y H, LIU Y, XU J, et al. MobileSal: extremely efficient RGB-D salient object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 10261-10269.

[7] PIAO Y, RONG Z, ZHANG M, et al. A2dele: adaptive and attentive depth distiller for efficient RGB-D salient object detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 9060-9069.

[8] ZHANG J, LIANG Q, SHI Y. KD-SCFNet: towards more accurate and efficient salient object detection via knowledge distillation[EB/OL]. (2022-09-21)[2024-02-25]. https://arxiv.org/abs/2208.02178.

[9] ZHOU H, QIAO B, YANG L, et al. Texture-guided saliency distilling for unsupervised salient object detection[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2023:

7257-7267.
[10] 李沼洁, 朱恒亮, 毛国君, 等. 渐进式特征增强的弱监督显著性目标检测[J]. 计算机工程, 2024, 50(12): 233-244.
LI Z J, ZHU H L, MAO G J, et al. Progressively feature-enhanced weakly supervised for salient object detection[J]. Computer Engineering, 2024, 50(12): 233-244. (in Chinese)
[11] 李军侠, 王星驰, 殷梓, 等. 边缘深度挖掘的弱监督显著性目标检测[J]. 计算机工程, 2023, 49(7): 169-178.
LI J X, WANG X C, YIN Z, et al. Weakly supervised salient object detection via edge depth mining[J]. Computer Engineering, 2023, 49(7): 169-178. (in Chinese)
[12] CHEN C, WANG G, PENG C, et al. Improved robust video saliency detection based on long-term spatial-temporal information[J]. IEEE Transactions on Image Processing, 2019, 29: 1090-1100.
[13] SONG H, LIU Z, DU H, et al. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning[J]. IEEE Transactions on Image Processing, 2017, 26(9): 4204-4216.
[14] GUO J, REN T, BEI J. Salient object detection for RGB-D image via saliency evolution[C]//Proceedings of 2016 IEEE International Conference on Multimedia and Expo (ICME). Washington D. C., USA: IEEE Press, 2016: 1-6.
[15] XU Y, YU X, ZHANG J, et al. Weakly supervised RGB-D salient object detection with prediction consistency training and active scribble boosting[J]. IEEE Transactions on Image Processing, 2022, 31: 2148-2161.
[16] 孙福明, 胡锡航, 武景宇, 等. 跨模态交互融合与全局感知的 RGB-D 显著性目标检测[J]. 软件学报, 2024, 35(4): 1899-1913.
SUN F M, HU X H, WU J Y, et al. RGB-D salient object detection based on cross-modal interactive fusion and global awareness[J]. Journal of Software, 2024, 35(4): 1899-1913. (in Chinese)
[17] DONG S, HONG X, TAO X, et al. Few-shot class-incremental learning via relation knowledge distillation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2021, 35(2): 1255-1263.
[18] LIN S, JI R, CHEN C, et al. Holistic CNN compression via low-rank decomposition with knowledge transfer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(12): 2889-2905.
[19] LIU B, RAO Y, LU J, et al. MetaDistiller: network self-boosting via meta-learned top-down distillation[EB/OL]. (2019-01-22)[2024-02-25]. https://arxiv.org/abs/1807.03748.
[20] VAN DEN OORD A, LI Y, VINYALS O. Representation learning with contrastive predictive coding[EB/OL]. (2020-08-27)[2024-02-25]. https://arxiv.org/abs/2008.12094.
[21] 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. Washington D. C., USA: IEEE Press, 2020: 9729-9738.
[22] XU G, LIU Z, LI X, et al. Knowledge distillation meets self-supervision[EB/OL]. (2020-06-12)[2024-02-25]. https://arxiv.org/abs/2006.07114.
[23] TIAN Y, KRISHNAN D, ISOLA P. Contrastive representation distillation[EB/OL]. (2022-01-24)[2024-02-25]. https://arxiv.org/abs/1910.10699.
[24] WU H, QU Y, LIN S, et al. Contrastive learning for compact single image dehazing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2021: 10551-10560.
[25] HUANG R, ZHAO Q, XING Y, et al. A saliency enhanced feature fusion based multiscale RGB-D salient object detection network[C]// Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2024: 9356-9360.
[26] JU R, GE L, GENG W, et al. Depth saliency based on anisotropic center-surround difference[C]// Proceedings of 2014 IEEE International Cconference on Image Processing (ICIP). Washington D. C., USA: IEEE Press, 2014: 1115-1119.
[27] PENG H, LI B, XIONG W, et al. RGBD salient object detection: a benchmark and algorithms[C]//Proceedings of Computer Vision-ECCV 2014, Berlin, Germany: Springer, 2014: 92-109.
[28] LI N, YE J, JI Y, et al. Saliency detection on light field[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 39(8): 1605-1616.
[29] ZHANG J, FAN D P, DAI Y, et al. RGB-D saliency detection via cascaded mutual information minimization[C]// Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2021: 4338-4347.
[30] LIU N, ZHANG N, SHAO L, et al. Learning selective mutual attention and contrast for RGB-D saliency detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(12): 9026-9042.
[31] FAN D P, CHENG M M, LIU Y, et al. Structure-measure: a new way to evaluate foreground maps[EB/OL]. (2017-08-02)[2024-02-25]. https://arxiv.org/abs/1708.00786.
[32] MARGOLIN R, ZELNIK-MANOR L, TAL A. How to evaluate foreground maps?[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 248-255.
[33] FAN D P, LIN Z, ZHANG Z, et al. Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(5): 2075-2089.
[34] CONG R, LIN Q, ZHANG C, et al. CIR-Net: cross-modality interaction and refinement for RGB-D salient object detection[J]. IEEE Transactions on Image Processing, 2022, 31: 6800-6815.
[35] WU Z, ALLIBERT G, MERIAUDEAU F, et al. HiDAnet: RGB-D salient object detection via hierarchical depth awareness[EB/OL]. (2023-01-18)[2024-02-25]. https://arxiv.org/abs/2301.07405.
[36] WU Z, PAUDEL D P, FAN D P, et al. Source-free depth for object pop-out[EB/OL]. (2023-09-25)[2024-02-25]. https://arxiv.org/abs/2212.05370.
[37] CHEN J, KAO S, HE H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[EB/OL]. (2023-05-21)[2024-02-25]. https://arxiv.org/abs/2303.03667.

编辑 陆燕菲