

基于最大差异化竞争的通用视觉难样本挖掘^{*}

鄢杰斌^{1,2}, 祝文涛^{1,2}, 刘学林^{1,2}, 陈俊杰^{1,2}, 钱峰^{1,2}, 方玉明^{1,2}

¹(江西财经大学 计算机与人工智能学院, 江西 南昌 330013)

²(多媒体智能处理江西省重点实验室 (江西财经大学), 江西 南昌 330013)

通信作者: 方玉明, E-mail: leo.fangyuming@foxmail.com



摘要:近年来,深度学习发展迅速,在计算机视觉研究中取得了巨大的成功.在发展过程中,模型的测试和改进方向是研究者们关注的核心.然而,视觉模型比较范式是在封闭数据集上训练(验证)和测试,然后通过测试结果和真实标签的偏差来获得难样本,用于反馈模型的问题和改进方向.这种方式存在的问题包括:1)数据集中少量的数据无法真实反映模型的问题;2)模型预训练等一些操作可能导致数据泄露,因此展现的性能可能有偏差.提出基于最大差异化竞争的通用视觉难样本挖掘算法,自动挖掘真实的难样本,用于指出模型的问题.所提算法遵循“通过模型博弈来比较模型”的思想,联合视觉任务内和多视觉任务间预测结果的“不相似性”优化挖掘潜在的难样本,旨在以可控的、高效的方式为计算机视觉领域提供新的测试基准.实验证明,所构建的测试基准 GHS-CV 相较于单视觉任务的难样本挖掘(语义分割难样本集 SS-C,显著目标检测难样本集 SOD-C)更能暴露出模型的缺陷.其中,相较于 DeepLabv3+模型在 SS-C 数据集上的性能,DeepLabv3+在 GHS-CV 数据集上的 mIoU 下降了约 20%;相较于 VST 模型在 SOD-C 数据集上的性能,VST 在 GHS-CV 数据集上的 F_β 下降了约 36%.

关键词:深度学习;计算机视觉;难样本挖掘;最大差异化竞争;模型评价

中图法分类号: TP391

中文引用格式: 鄢杰斌, 祝文涛, 刘学林, 陈俊杰, 钱峰, 方玉明. 基于最大差异化竞争的通用视觉难样本挖掘. 软件学报, 2026, 37(8): 3386–3404. <http://www.jos.org.cn/1000-9825/7551.htm>

英文引用格式: Yan JB, Zhu WT, Liu XL, Chen JJ, Qian F, Fang YM. General Visual Hard Sample Mining Based on Maximum Discrepancy Competition. Ruan Jian Xue Bao/Journal of Software, 2026, 37(8): 3386–3404 (in Chinese). <http://www.jos.org.cn/1000-9825/7551.htm>

General Visual Hard Sample Mining Based on Maximum Discrepancy Competition

YAN Jie-Bin^{1,2}, ZHU Wen-Tao^{1,2}, LIU Xue-Lin^{1,2}, CHEN Jun-Jie^{1,2}, QIAN Feng^{1,2}, FANG Yu-Ming^{1,2}

¹(School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang 330013, China)

²(Jiangxi Provincial Key Laboratory of Multimedia Intelligent Processing (Jiangxi University of Finance and Economics), Nanchang 330013, China)

Abstract: In recent years, deep learning has developed rapidly and achieved significant success in computer vision, with model evaluation and improvement remaining central concerns for researchers. However, the commonly used model comparison paradigm relies on training (or validation) and testing on closed datasets, and then identifies hard samples based on discrepancies between predictions and ground-truth labels, which provide feedback on model weaknesses and directions for improvement. This paradigm suffers from two major limitations: 1) the limited size and coverage of datasets often fail to faithfully reflect the true weaknesses of models; 2) procedures such as pretraining may introduce data leakage, resulting in potential biases in the demonstrated performance. To address these issues, this study proposes a general visual hard sample mining algorithm based on maximum discrepancy competition, which automatically mines real hard samples to

* 基金项目: 国家自然科学基金 (62461028, U24A20220, 62562034, 62402201); 江西省自然科学基金 (20243BCE51139, 20232BAB202001, 20252BAC240197)

收稿时间: 2025-02-24; 修改时间: 2025-07-27; 采用时间: 2025-09-05; jos 在线出版时间: 2026-02-11

CNKI 网络首发时间: 2026-02-12

reveal models' deficiencies. The proposed algorithm follows the principle of "comparing models through competition" and optimizes the discovery of potential hard samples by jointly exploiting the intra-task and cross-task prediction dissimilarities, aiming to provide new test benchmarks for the field of computer vision in a controllable and efficient manner. Experimental results demonstrate that the constructed benchmark named GHS-CV exposes models' weaknesses more effectively than single-task hard sample benchmarks (i.e., the semantic segmentation hard sample set SS-C and the salient object detection hard sample set SOD-C). Specifically, compared to DeepLabv3+ on SS-C, the mIoU drops by about 20% on GHS-CV, while compared to VST on SOD-C, the F_{β} decreases by about 36%.

Key words: deep learning; computer vision; hard sample mining; maximum discrepancy competition; model evaluation

在大量视觉数据和高性能计算设备的支撑下, 计算机视觉研究取得了令人瞩目的成就, 包括图像质量评价^[1]、图像美感评价^[2]、图像分类^[3]、目标检测^[4]和图像语义分割^[5]等。许多科研工作者投入了大量精力用于构建各类任务的数据集和设计模型。然而, 如何准确地评价现有模型已成为一个至关重要的问题。计算机视觉模型的主流评估方法包括主观评价和客观评价等^[6]。其中, 主观评价通过人工的方式评估模型的预测结果, 该类方法虽然准确, 但是人工评估的方式耗费较高且难以大规模地开展; 客观评价指的是使用客观指标定量描述模型的预测结果, 如目标检测使用平均精度均值 (mean average precision, mAP)、图像语义分割使用交并比 (intersection over union, IoU) 等评价模型性能。

传统的模型评估方法通常会遵循以下过程。首先, 从所有可能的图像中预选出一批具有代表性的样本作为测试集, 这些样本应覆盖模型在实际应用中可能遇到的多种场景。测试集中每个样本的真实标签或期望输出通常由标准方法进行标注, 作为模型预测的基准。模型评估的依据通常是模型在测试集上的性能指标, 表现最佳的模型会被视为最优。当前的评估方法对深度学习模型的研究起到了极大的推动作用, 大多数计算机视觉任务的评估范围也趋于一致。但这种评估方法的主要问题在于, 自然图像的高维度和大数据量与可负担的测试规模之间存在显著矛盾, 使得测试集的规模受到限制。

随着计算机视觉研究的深入, 基于深度学习的模型在许多基准数据集上表现优异。然而, 当模型性能趋于饱和时, 研究者逐渐将部分注意力转向数据本身的问题, 包括数据质量、标注准确性、类别分布不均以及难样本等。很多的数据问题都使用人工的方式去解决, 如对数据重新标注以提高标注质量, 或有针对性地增加某些类别数量以解决长尾分布问题等。特别是在模型分析方面, 研究者往往通过对难样本特征的深入研究来发现模型的不足。例如, Yan 等人^[7]发现语义分割模型在处理具有奇特外观、不同视角、小规模以及仅部分结构可见的物体时, 分割效果显著下降。当前, 难样本的识别通常在大规模数据集上测试模型来完成。例如, PASCAL VOC^[8]数据集包含约 1 万张图像, COCO^[9]数据集则包含约 11 万张图像。在测试过程中, 通过分析模型在这些数据集上的表现, 结合人为筛选的方法, 找出那些模型难以正确处理的样本。这种方法虽能有效识别难样本, 但依赖人工筛选的方式效率较低, 难以适应更大规模数据集的需求。

当前, 有研究者提出了多种自动难样本挖掘的方法。Wang 等人^[10]采用最大差异化竞争的方法, 从无标签图像数据中挖掘图像分类任务的难样本; Wang 等人^[11]基于在线难样本挖掘框架, 使图像质量评价模型能够更关注这些困难样本, 从而提升模型的鲁棒性和性能。上述方法均是在单一的计算机视觉任务上挖掘难样本, 对相应的模型而言具有较大的挑战性。然而, 这些在单一任务上挖掘出的难样本对其他任务的模型不一定同样具有挑战性 (具体可见第 4.1、4.2 节)。例如, 在图像分类任务中被标记为难样本的图像, 可能对目标检测、语义分割等其他任务模型的挑战较小。另一类难样本挖掘方法是对抗攻击, 其通过对输入数据施加微小扰动, 使得深度学习模型产生错误预测。这些扰动对于人类视觉系统通常是不可察觉的, 但却能有效地欺骗模型。这种方式生成的难样本属于合成难样本, 因为扰动是人为设计并添加到原始数据中的, 而非数据本身自然存在的特征。

针对上述问题, 本文提出一种联合多任务的难样本自动挖掘方法, 旨在挖掘通用的计算机视觉任务难样本, 为计算机视觉模型的优化提供新思路。该方法通过联合优化从粗到细 3 种粒度 (即图像级、区域级和像素级) 的视觉任务, 共同识别使得多类模型表现较差的样本。这些样本可能因多任务间的共享特性而具有更广泛的代表性, 从而为提升模型在复杂场景中的泛化能力提供数据支持。

本文的贡献总结如下。

1) 提出一种联合多任务的通用视觉难样本自动挖掘方法, 自动、高效地挖掘潜在的计算机视觉任务的难样

本, 形成公平、统一的测试基准, 指出计算机视觉模型优点和不足.

2) 构建一个通用计算机视觉任务难样本数据集 (general hard samples for computer vision, GHS-CV), 包括图像级、区域级和像素级注释. GHS-CV 数据集可用于验证除本研究实验模型外的其他图像分类任务、语义分割任务、显著目标检测任务模型.

3) 使用 GHS-CV 数据集在多个视觉任务中进行测试, 包括图像分类、显著目标检测、语义分割、目标检测、伪装目标检测等任务. 实验结果表明, 本文提出的数据集在其他计算机视觉任务同样具有挑战性.

1 相关工作

本节介绍相关工作. 第 1.1 节概述模型评估的基本框架和方法, 回顾度量空间与多度量空间的相关工作. 第 1.2 节回顾测试计算机视觉模型泛化能力的一些新兴思路和方法.

1.1 模型评估

随着深度学习的迅速发展和广泛应用, 计算机视觉领域也进入了高速发展阶段. 例如, 卷积神经网络 (convolutional neural network, CNN), 如 ResNet^[3]、DenseNet^[12]等模型, 在大规模基准数据集上的性能超过了人类; Faster R-CNN^[13]、YOLO^[14]、SSD^[15]等模型在目标检测应用场景中取得了显著成绩; 生成对抗网络和变分自编码器在图像生成和风格迁移方面获得了较大的突破^[16,17]. 然而, 诸如对抗性攻击^[18–20]、训练数据有限^[10,21,22]、共振攻击^[23]等问题, 这些因素给计算机领域研究带来了新的挑战^[24–26], 并推动了研究者们寻求新的解决方案^[27–30].

在应对这些挑战的过程中, 越来越多的研究者开始从“难样本”视角重视计算机视觉模型的稳定性和局限性. 现有的难样本研究大体可以分为两个方向: 一方面是在模型训练阶段的难样本挖掘与利用. 这类方法通过关注误判或预测不稳定的样本来提升模型表现. 苏亚鹏等人^[31]提出的 SA-Focal Loss 可动态调整参数以聚焦难样本; Smirnov 等人^[32]利用嵌入向量和相似度对比构造硬样本对; Xia 等人^[33]则通过混合特征融合机制提升难样本检测能力. 此外, 还有研究通过更强大的特征提取器、多任务结构或模型修剪差异指导, 增强模型对细粒度难样本的表征能力^[11,34–36]. 另一方面是在模型评估阶段的难样本挖掘与测试方法. 这类方法更关注模型的泛化能力与鲁棒性, 通常基于开放世界的大规模未标注图像集, 利用模型间预测差异自动筛选出最具挑战性的样本. 这一方向的代表性方法是最大差异化竞争.

在这一背景下, Hoiem 等人^[37]通过分析目标检测器中的物体特征及不同类型的误报, 揭示了模型错误的主要类型. Zhang 等人^[38]则专注于单一数据集的错误案例, 深入研究了行人检测器的性能表现. Ronchi 等人^[39]采用了预定义的错误类别, 对多实例姿态估计进行了详细的诊断分析, 从而提出了针对性的改进方案. 在显著性目标检测任务中, Fan 等人^[40]构建了具有鲜明特点的显著性目标检测数据库, 该数据库专门包含无显著目标图像, 指出了现有数据库中存在的严重数据偏差问题. Wang 等人^[41]进一步对显著性目标检测任务进行了全面研究, 通过在一个新数据库上比较不同的显著性目标检测架构, 证明了它们的可转移性和鲁棒性. 在语义分割任务中, 同样面临类似的数据偏差和模型鲁棒性问题. 许多研究致力于提升语义分割模型在复杂场景中的表现, Hendrycks 等人^[42]在应对分布外数据问题上提出了基于不确定性估计的方法, 通过对模型在未知数据上的置信度分布进行评估, 有效检测出分布外数据样本.

虽然用于计算机视觉任务的“大规模”测试集通常包含数以万计的自然图像, 但这些图像的分布相对现实世界的自然图像中仍是极其稀疏的. Recht 等人^[43]的研究表明, 即使是微小的自然分布变化, 也会导致 CIFAR-10^[44]和 ImageNet^[45]上许多图像分类模型的准确率显著下降, 这表明当前的测试集可能远不能代表现实世界中可能遇到的复杂自然图像. 由于测试集是固定的, 计算机视觉模型可能通过大量的超参数调优适应这些特定的测试图像, 进而导致在测试集上表现优异的模型无法有效推广到内容更为丰富和多样的真实世界图像. 近年来, 大语言模型 (large language model, LLM), 如 GPT-4-Turbo^[46]、Gemini-Pro^[47]和 OpenChat-3.5^[48]的开发进入爆发式增长阶段, 关于大模型评估的工作主要还是在大规模的数据集上进行测试, 如 C-Eval^[49]、MMLU^[50]和 BIGbench^[51]等. 目前, 大多数模型评估框架仍采用传统的基于固定数据集的评估方法, 然而, 传统评估方法通常依赖人工标注的大规模数据集,

而数据集的构建往往需要投入大量时间和精力,特别是在处理复杂任务或多模态数据时,成本更为高昂。此外,数据集的覆盖范围有限,难以全面反映现实世界丰富多样的场景,导致模型在测试集上的性能可能无法准确反映其在实际应用中的表现。固定测试集的重复使用也可能导致模型过拟合于测试集,掩盖其在未知场景中的泛化能力。很多数据集仅针对单一任务而设计,缺乏跨任务的通用性,这使得基于这些数据集的模型评估难以全面反映模型在多任务场景中的性能表现。

针对上述问题,我们提出了一种自动挖掘通用视觉难样本的方法。与传统方法不同,该方法挖掘出的难样本不仅适用于单一的视觉任务,还能广泛适用于其他视觉任务。这种通用性难样本挖掘方法能够为计算机视觉任务模型优化提供更具代表性的数据支持,提升模型在复杂场景中的鲁棒性和泛化能力。

1.2 最大差异化竞争

传统的模型评估方法通常依赖于大规模数据集,通过衡量模型在这些数据集上的性能表现(如准确率、召回率等)来评估其优劣。然而,与单纯依赖性能指标进行评估不同,在计算神经科学^[52]、软件测试^[53]、图像处理^[54]和计算机视觉^[7,10]等领域,研究者逐渐采用“攻击”或“对抗性测试”的方法来评估模型的能力。这些方法通过构造对抗性样本或故意设计难度较大的输入来欺骗模型,从而分析其鲁棒性、泛化能力和潜在漏洞。这种方式不仅能够揭示模型的不足,还能为模型优化提供有针对性的改进方向。

Wang 等人^[55]提出了最大差异化竞争(maximum differentiation competition, MAD)方法,该方法通过计算机生成能够区分模型的边界样本,从而减少比较样本数量。具体而言,MAD 通过合成一对极端样本来最大化一个模型的响应,同时保持另一个模型的响应不变。然后进行主观测试,判断哪一对样本更容易区分,从而比较模型性能。该方法显著减少了样本数量需求,减少了人工标注成本,提高了模型比较的效率,特别是在样本空间维度很高的情况下,比如图像质量评价^[28,56]。并且,通过对生成样本的研究,MAD 方法还能够为模型改进提供指导,甚至可以结合多个模型的优点形成更优的模型。尽管 MAD 方法能够显著减少样本数量,但它在复杂模型上的应用存在一定局限性。MAD 方法依赖于梯度信息来解决约束优化问题,而对于一些复杂的模型,梯度计算可能非常困难甚至不可行。此外,MAD 方法生成的样本可能不自然,这限制了其在实际应用中的改进作用。为了解决这些问题,Ma 等人^[54]提出了组最大差异化竞争(group maximum differentiation competition, gMAD)方法,将 MAD 方法推广至可以比较任意的模型,而不受模型是否可微这一限制的影响。gMAD 通过一组攻击者模型来最大化与防御者模型的差异,同时保持防御者的预测不变,以此找到攻击防御者的最佳样本对。该方法引入了攻击性和抵抗性两种衡量标准,分别表示模型在攻击其他模型和抵御攻击时的表现。同时,也验证了 gMAD 在图像质量评价、图像美感评价和流媒体质量评价方面的有效性。

gMAD 在多个视觉任务中验证其在模型比较方面的有效性。在图像分类任务^[10]中,通过在大量未标注图像集中挑选那些能使分类器预测产生最大分歧的图像,从而有效识别分类器之间的差异。这样的自适应测试集能够更加准确地揭示模型在现实场景中的表现差异。通过人类标注这些有争议的图像,可以有效评价模型的泛化能力,进而发现改进方向。在图像质量评价任务中,gMAD 能够在多个图像质量评价模型中自动选择出最具区分性的样本对,从而揭示模型的优缺点^[29]。例如,在多个图像质量评价模型(如对比度、锐度、噪声评估模型)的比较中,gMAD 可以生成极端样本对,使得一组模型的响应差异最大化,从而为改进模型、提升预测准确性提供数据支持^[23]。这种方法能够在大规模、维度高的图像空间中有效提高比较效率,减少了大量的主观测试成本。在 LLM 的评估中,gMAD 方法也展现了重要的应用价值。为了对 LLM 进行自动化且公正的评估,并同时降低人工数据处理的成本,Feng 等人^[56]引入了 gMAD 方法,从极大的指令数据集中自动筛选出小部分信息丰富且具有多样性的指令集,这些指令能够最大化不同模型之间的响应差异,从而用作模型间对比的基准。筛选出来的指令不仅能捕捉模型在理解和生成能力上的差异,还能为模型改进提供针对性反馈。

然而,gMAD 方法在实际使用中也暴露出一些局限性。在缺乏约束条件的情况下,gMAD 方法可能生成大量具有重复特征信息的样本,这不仅无益于模型评估,还可能导致样本类别分布不均以及生成样本的自然性不足。此外,以往的相关研究多集中于单一视觉任务中的模型比较和难样本挖掘,本文通过结合图像级、区域级和像素级

等不同视觉任务, 将这些任务联合起来, 自动挖掘能够使计算机视觉任务模型失效的通用难样本. 这些难样本不仅对联合的视觉任务模型具有挑战性, 对于其他任务模型仍有显著的挑战性 (详细分析见第 4.2 和 4.3.1 节).

2 实验方法

在本节中, 首先介绍问题的表述. 然后, 详细介绍本文提出的联合多视觉任务模型的难样本挖掘框架. 最后, 引入挖掘限制策略, 保证挖掘难样本的多样性.

2.1 问题表述

真实开放的视觉世界包含了各种各样的图像, 如何在真实开放的视觉世界评估不同任务的模型或挖掘出计算机视觉任务潜在的难样本是一个巨大的挑战. 传统方法往往使用性能计算指标或人为的方式从测试集中挑选出难样本, 这种方式带有一定的主观性和样本模式单一性, 难以完整地揭示模型的不足. 针对上述问题, 本文提出联合多视觉任务的难样本自动挖掘框架 (如图 1). 框架从一个无标签的数据集 I (如包括 50 万张图像) 开始, 它尽可能地涵盖开放视觉世界的图像. 本文分别将不同计算机视觉任务的模型在数据集 I 上进行测试, 通过最大化每个任务中两两模型的性能差异, 挖掘出每个任务的难样本集合; 接着, 将两个任务进行关联, 找出能够同时对两个任务模型表现出高难度的样本, 通过新加入任务与前两个任务 (其中之一) 的联系, 进一步引入第 3 个任务, 依此可推广到 m 个任务. 与传统方法相同, 模型评估或者难样本挖掘都是从一个比较大的图像数据集 I 开始, 包含 N 张图像 (如 50 万张), 这些图像涵盖了与现实世界相似的复杂场景. 然而, 随着数据集 I 的不断扩大, 人工标注数据集变得难以承受. 为此, 在本文框架中, 不需要事先对数据集进行标注, 而是选择性地挖掘出一个包含 L 张图像的代表性子集 ($L \ll N$), 子集的每张图片都可能是视觉任务的难样本. 本文目的是通过分析视觉模型 F_i 在数据集 I 上执行不同任务时的预测结果, 自动挖掘出计算机视觉任务的难样本.

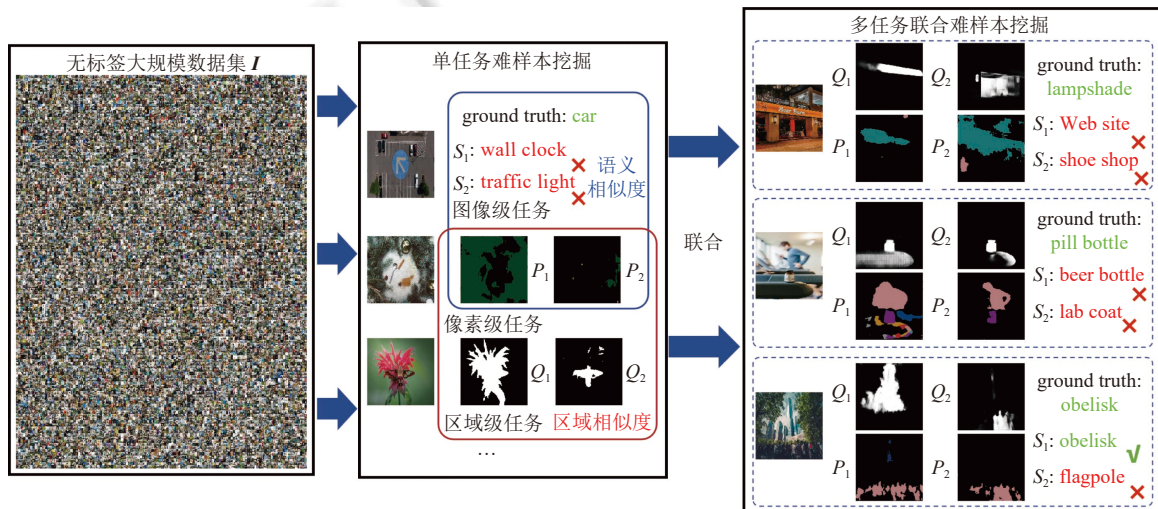


图 1 通用视觉任务难样本挖掘方法

2.2 联合多任务模型挖掘难样本

给定原始图像集 I 、计算机视觉任务 $F=\{F^1, F^2, \dots, F^m\}$, 通过比较这些任务以及任务模型之间的预测结果来挖掘潜在的难样本集 O . 具体而言, 本文利用图像级任务 F^1 、区域级任务 F^2 、像素级任务 F^3 之间的联系, 以及它们的模型预测结果, 最大化模型之间的差异. 通过这种方式, 从图像集 I 中挑选出那些在多任务上表现出显著差异的图像, 这些图像很可能是模型的难样本, 从而构成潜在的难样本集 O . 需要注意的是, 本文提出的框架具有易扩展性, 不仅限于 3 个视觉任务 (图像分类、显著目标检测和语义分割), 本文选取了图像级、区域级和像素级的任务做了具体的研究, 在 O 中的图像 I_n 可由公式 (1) 获得:

$$I_n^* = \underset{I_n \in I}{\operatorname{argmax}} (w_1 D_1(F_i^1(I_n), F_j^1(I_n)) + w_2 D_2(F_i^2(I_n), F_j^2(I_n)) + \dots + w_m D_m(F_i^m(I_n), F_j^m(I_n))) \quad (1)$$

其中, D_m 为不同任务的测度函数, w_m 是不同任务的权重系数. 以图像分类模型、语义分割模型以及显著目标检测模型为例, 对于图像分类模型性能的度量, 与传统的二元判断 (对与错) 相比, 本文使用了 WordTree 层次结构上的加权距离方法^[10], 这个方法在语义上更符合人类的认知; 对于语义分割和显著目标检测模型的度量, 则采用常见的度量方法 mIoU 和 F_β . 然而, 公式 (1) 中考虑的仅是单一任务的联系, 没有考虑到任务间的联系. 本文提出的框架不仅考虑相同任务模型的差异, 还要联系不同任务模型的差异. 计算方式如下:

$$I_n^* = \underset{I_n \in I}{\operatorname{argmax}} (w_1 D_1(F_i^1(I_n), F_j^1(I_n)) + w_{1,2} D_{1,2}(F^1, F^2) + w_2 D_2(F_i^2(I_n), F_j^2(I_n)) + \dots + w_{n,m} D_{n,m}(F^n, F^m) + w_m D_m(F_i^m(I_n), F_j^m(I_n))) \quad (2)$$

其中, $D_{n,m}$ 为不同任务之间的测度函数, $w_{n,m}$ 为不同任务之间的权重系数. 考虑到视觉任务模型不同粒度的预测情况, 我们选择从粗到细的 3 种粒度的视觉模型去实例化公式 (2) 中的 $D(\cdot)$, 即全局预测的图像分类任务、区域级预测的显著目标检测任务以及像素级预测的语义分割任务. 需要注意, 此 3 个模型可以替换成其他视觉任务模型, 其中, 图像分类任务和语义分割任务是基于 Wang 等人^[10]的方法设计出的一个多对多的语义距离度量 (WordTree), 用于衡量图像分类和语义分割任务之间的语义相似性; 而语义分割任务和显著目标检测任务则采用了 greedy-IoU (gIoU) 方法^[57], 该方法忽略了类别信息的影响, 专注于目标形状的匹配, 用以评估语义分割任务与显著性目标检测任务之间的相似度. 由于 F_β 、mIoU 和 gIoU 指标均处于 $[0, 1]$ 区间, 而 WordTree' 的数值范围大于 1. 因此, 我们在各任务内部对差异分数进行 Min-Max 标准化. 需要指出的是, WordTree' 在归一化后会出现大量取值接近 1 的情况, 即较大的 WordTree' 值被压缩到归一化上界. 考虑到区域级任务 (显著目标检测)、像素级任务 (语义分割) 对图像特征信息更敏感, 根据公式 (2), 我们对不同任务设置了权重系数, 对于图像分类以及其与语义分割之间的差异度量权重系数设置为 0.3, 显著目标检测、语义分割以及它们之间的差异度量权重系数都设置为 1.

对于多任务模型的预测结果 (如图 2 所示, 以联合语义分割和显著性目标检测为例), 可能会出现以下几种情况.

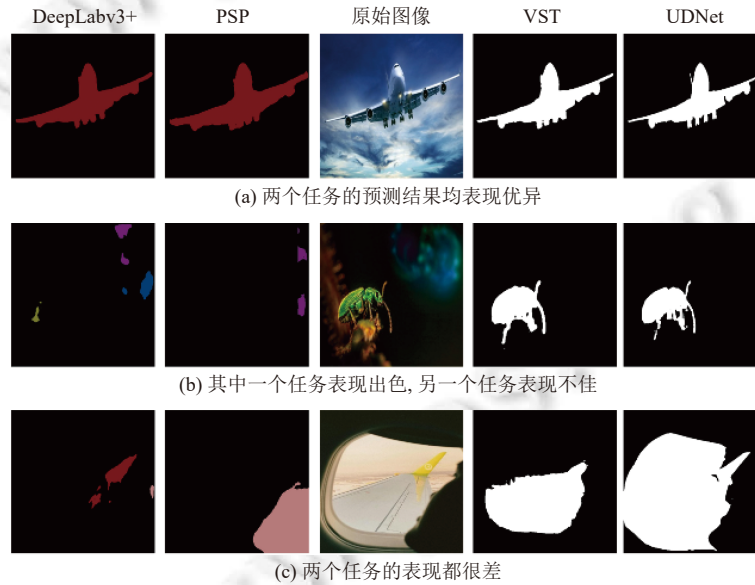


图 2 多任务联合预测差异产生的 3 种情况说明

(1) 如图 2(a) 所示, 两个任务的预测结果都非常出色. 这通常发生在数据集较小时, 或两个任务模型对数据分布的表示能力较强的情况下. 然而, 随着图像集 I 规模的扩大, 这种情况可能会得到缓解.

(2) 如图 2(b) 所示, 其中一个任务的预测表现良好, 而另一个任务的表现则相对较差. 在这种情况下, 图像样本 I_n 在区分这两个任务的性能方面起到了关键作用.

(3) 如图 2(c) 所示, 两个任务的预测结果都较差. 虽然这种情况下图像样本对区分任务性能的作用不大, 但它们揭示了这两个任务模型的共同弱点.

本文的工作旨在避免第 1 种情况的发生, 并尽可能实现第 3 种情况. 在实践中, 为了获得计算机视觉任务通用的难样本, 通过公式 (2) 计算出 n 个最大的距离, 选择其中 Top- k 张图像, 形成难样本图像集 O .

2.3 多样化难样本挖掘

本节将进一步详细说明如何挖掘多任务难样本. 通过优化公式 (2), 直接在原始图像集 I 上对测试模型应用最大差异化竞争. 然而, 这种方法往往会暴露出一些信息量少、相同属性多的图像, 会出现许多无显著目标、语义信息少、无法分辨的图像, 如图 3(a). 这些图像样本并不是本文期望的难样本, 有两个原因: (1) 大多数语义分割数据集上至少包含一种语义信息 (如 PASCAL VOC^[8]、MS COCO^[9]), 大多数显著目标检测数据集至少包含一个显著性物体 (如 DUTS^[23]), 以及 ImageNet^[45] 中图像都是可以被人识别的, 因此, 不管是图像分类任务、显著目标检测任务, 还是语义分割任务模型在这类图像上表现不佳都是不能接受的. (2) 挖掘出的样本的独特性与开放视觉世界的多样性特征不匹配, 这一点在很多工作^[16,54,55]中被忽视. 此外, 由于预训练数据集 (本文采用 PASCAL VOC^[8]) 的影响, 语义分割模型可能无法识别某些类别, 例如建筑和风景, 这类图像样本也不是本文需要的难样本. 因此, 需要进一步优化样本选择策略, 以确保所挖掘的难样本能够准确反映模型在复杂场景中的表现.

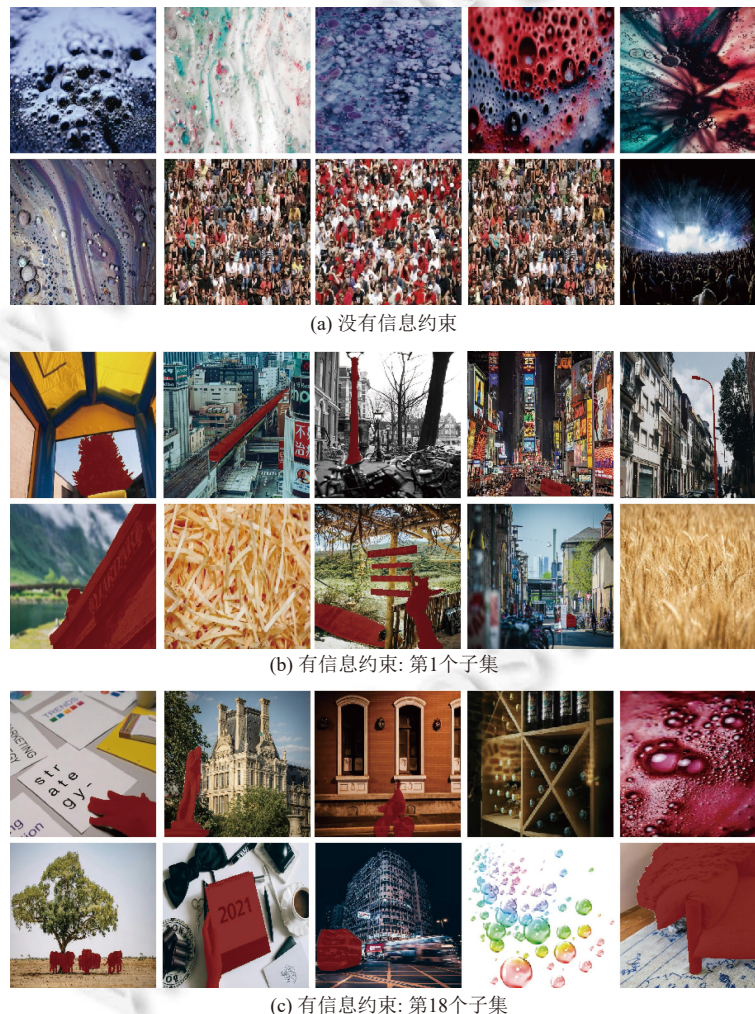


图3 没有和有信息约束的情况下选择视觉图像样本

基于上述考虑, 本文设置了一个图像信息约束, 增加区域级视觉任务模型测试图像样本形状多样性和像素级视觉任务模型测试图像样本类别多样性, 从多个角度对模型进行验证. 首先, 将一对区域级任务模型分别设置为防御模型和攻击模型. 然后, 根据防御模型预测的目标大小和纵横比经验性地将对象大小和纵横比设置为 s 个尺度 (默认设置为 5), 这样有助于增加所选图像样本的多样性. 本文还在每个尺度上增加一个像素级任务模型预测的类别约束, 在不同尺度中, 出现不同类别信息的图像样本, 这样有助于增加每个尺度所选图像样本的类别多样性. 如图 3(b) 和图 3(c) 所示, 本文通过图像信息约束为每组测试模型生成了 $s \times s \times k$ 张图像. 所有测试模型选择的图像样本构成最终的难样本集 $\mathcal{O}$.

3 实验设置与主观标注

3.1 实验设置

3.1.1 构建数据集

为了使原始图像集 $\mathcal{I}$ 更接近现实世界中的复杂场景, 受 Wang 等人^[10]的启发, 本文收集现实世界中常见的 200 个类别, 以避免稀有和抽象的类别以及自 2012 年以来变化较大的类别. 此外, 为了适应语义分割任务的需求, 本文将 PASCAL VOC 数据集^[8]的 20 种类别作为关键词来爬取图像, 从 Pexels (<https://www.pexels.com/zh-cn/>)、百度 (<https://image.baidu.com/>)、微软 (<https://cn.bing.com/images/>) 和 Flickr (<https://www.flickr.com/>) 等平台爬取了大量图像, 每个类别包含大量样本, 最终收集到超过 50 万张图像, 并将它们统一裁剪为 512×512 的尺寸. 与以往的研究^[28,56]不同, 本文没有对原始图像集 $\mathcal{I}$ 进行任何过滤, 因为希望 $\mathcal{I}$ 能够覆盖开放视觉世界中广泛的复杂场景, 而不是预先定义好的内容^[56].

3.1.2 模型选择与度量函数

本次实验分别为基于全局预测的图像分类任务、基于区域级预测的显著目标检测任务以及基于像素级预测的语义分割任务选择了 4 个不同的模型. 图像分类任务所采用的模型包括了 ResNet101^[3]、VGG16^[58]、NASNet-A-Large^[59]以及 PNASNet-5-Large^[60], 这些模型在精度和测试速度上各具特色, 适合不同的应用场景. 在显著目标检测任务中, 我们选择了 EGN^[61]、VST^[62]、PoolNet^[63]和 DCN^[64], 这些模型在检测显著目标时表现优异, 且能够在复杂背景下保持较高的鲁棒性. 对于语义分割任务, 我们则选用了 DeepLabv3+^[65]、PSPNet^[66]、CorrMatch^[67]和 U2PL^[68], 这些模型在像素级标注的精度和边界保留方面具有较高的性能表现.

对于度量函数, 图像分类任务采用了基于 WordTree 的加权距离来衡量模型性能; 语义分割任务使用了常见的 mIoU 度量标准; 而显著目标检测任务则使用了标准的 F_β 度量函数. 此外, 在图像分类任务中, Top-5 准确率是常用的性能评估指标之一. 因此, 为了度量图像分类与语义分割任务之间的联系, 本文基于图像分类模型的 Top-5 预测结果与语义分割的结果进行语义相似性度量. 由于这本质上是一个多对多的关系, 通过合理归类语义分割的类别 (如后文表 1 所示), 把占比最多的像素类别作为语义分割任务分类的结果, 将其简化为一对多的关系. 我们统计了图像的语义分割结果中占比最大的类别, 将其作为语义分割任务的关键类别, 并与图像分类任务进行语义相似性计算, 从中选择最相似的结果作为两个任务之间的度量. 虽然语义分割和显著目标检测都是分类任务, 不同的是语义分割任务是多分类, 显著目标检测则是二分类. 假如将背景视为假, 那么含有目标物体信息即为真, 这样语义分割任务和显著目标检测任务两者可以简单联合在一起. 然而, 由于语义分割任务是对目标对象都进行语义的分割, 显著目标检测任务则是更关注突出的物体, 这会导致前面的方法出现更高的误差. 我们采用了 gIoU^[57]的方法, 使用语义分割结果中一个类别标签与显著图做交并比计算, 该方法忽略了语义信息, 仅选择与显著目标具有最大交并比的类别作为关键信息, 并将该值作为语义分割任务与显著性目标检测任务之间的度量标准, 更合理地反映出语义分割任务和显著目标检测任务之间的联系.

3.1.3 实验约束

本节详细介绍实验约束的相关细节. 首先, 我们定义了 3 个关键参数, 即显著目标的大小、长宽比, 还有语义分割的类别属性, 分别为 *size*、*aspect_ratio*、*semantic_info*. 其中, *size* 表示显著目标在图像中所占的比例, *aspect_ratio* 则表示显著目标的宽度与高度的比值. *semantic_info* 代表语义分割图像中像素最多的类别. 需要特别

注意的是, 在应用这些约束时, 我们无法直接获取测试图像中物体的具体信息, 如数量 (如图 4(a) 和 (b) 所示) 及其标签 (如图 4(c)). 然而, 我们可以通过模型的预测结果来定位显著区域的位置和像素类别信息. 因此, 本文提出了简单且公平的方法来计算由显著目标检测模型预测的显著目标的 *size* 和 *aspect_ratio*, 还有语义分割模型预测的 *semantic_info*, 以便在实验中对这些目标进行有效的约束和分析.

表 1 PASCAL VOC 基准中的对象类别映射关系

Category	PASCAL VOC Category
Person	person
Craft	aero plane, boat
Wheeled vehicle	bicycle, car, motorbike
Placental mammal	sheep, horse, cow, dog, cat
Public transport	bus, train
Vessel	bottle
Plant	potted plant
Furnishing	chair, dining table, sofa
Bird	bird
Electronic device	TV/monitor

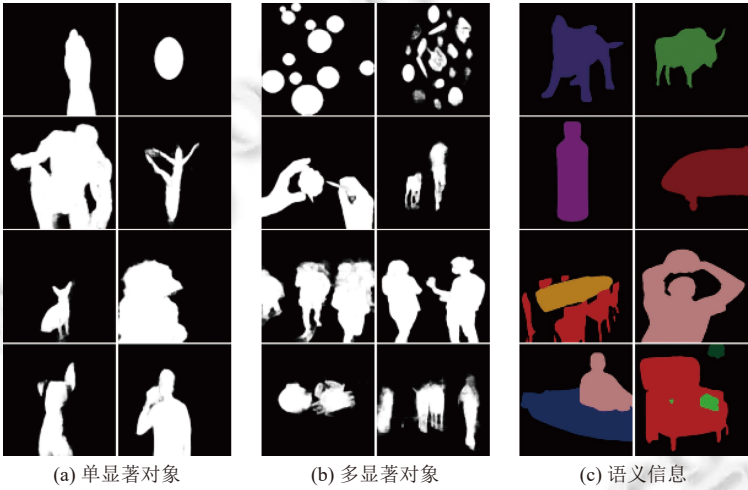


图 4 单显著对象、多显著对象和语义信息示例

具体计算方式如下:

$$size = \frac{1}{H \times W} \sum_{x=1}^H \sum_{y=1}^W G_{xy} \tag{3}$$

$$aspect_ratio = \frac{x_{right}(G) - x_{left}(G) + 1}{y_{bottom}(G) - y_{top}(G) + 1} \tag{4}$$

$$semantic_info = \max_{c \in C} \frac{\sum_{(x,y) \in \Omega} 1(L_{xy} = c)}{\sum_{(x,y) \in \Omega} 1} \tag{5}$$

其中, G 表示防御者模型预测得到的二值显著图. 在生成这个二值显著图 G 时, 我们采用了一种自适应阈值方法, 该阈值设置为原始显著性图平均值的两倍. H 和 W 分别表示图像 G 的高度和宽度. x_{left} 和 x_{right} 分别表示图像 G 中最左侧和最右侧显著像素所在的列, 而 y_{bottom} 和 y_{top} 分别表示图像 G 中最上面和最下面的显著像素所在的行. L_{xy} 代表在某个位置上的语义类别信息, Ω 表示图像区域的像素集合. 在计算显著目标的大小和长宽比之

前, 我们对显著性图像进行一次腐蚀操作. 该操作使用 3×3 滤波核, 目的是去除图像中的孤立显著像素, 从而确保计算出的显著目标大小和长宽比更加准确和稳健. 我们将显著目标大小和纵横比分别设置了 5 个尺度, 共有 5×5 个子集. 在每个子集中, 我们根据语义信息对筛选出的图像进行类别约束, 确保每个尺度筛选出的图像样本更丰富. 计算方式如下:

$$\text{s.t. } \sum_{i=1}^N x_{ic} \leq k, \quad x_{ic} \in \{0, 1\}, \quad \forall c = 1, 2, \dots, 10 \quad (6)$$

其中, x_{ic} 表示图像 i 是否被选中用于类别 c , 被选中则 x_{ic} 为 1, 否则为 0; c 表示类别的种类数量, 具体种类名称及映射关系见表 1. 每个类别选中的图像不超过 k . 如图 5(a) 所示, 可以利用显著目标检测模型 (如 VST) 的预测结果 (显著图), 自动计算每个样本中显著目标的包围框尺寸和纵横比. 基于这两个维度, 我们构建了二维分组空间, 也可以对这两个维度进行细粒度的划分, 如 6×6 、 7×7 等, 如图 5(b) 展示了更细粒度的划分 (7×7). 相比于更细粒度的划分 (如 6×6 或 7×7), 5 个区间能够获得同样类型的样本, 当测试图像总数量较少时, 设置细粒度的划分可能由于样本分布问题导致某些图像被误判为难样本; 当测试图像总数量较多时 (如本文使用的测试图像), 相邻的细粒度划分均可达到本文的目标. 需要注意的是, 本文难样本的选取规则是从每个分组中选取预测差异最大的 Top- k 样本, 可以通过调整 k 的值来覆盖更多的样本. 为了验证基于此分组策略挖掘的合理性, 我们进行了人工标注分析, 如图 5(c) 所示. 这些自动挖掘出的难样本, 其人工标注揭示的挑战特性 (如小目标、低对比度、复杂背景等) 与预期的一致.



图 5 约束可视化

3.2 主观标注

本文提出的 GHS-CV 与其他单任务数据集 (如语义分割难样本 SS-C^[7]、显著目标检测难样本 SOD-C^[69]) 不同, GHS-CV 不仅包含图像级标签信息、区域级标签信息, 还有像素级标签信息.

本节详细介绍数据集收集显著真值图的流程 (如图 6 所示). 为了确保 GHS-CV 数据集标注的高质量与一致性, 主观标注过程分为两个阶段: 粗标注和逐像素精标注^[40]. 在标注任务正式开始之前, 向参与人员详细说明了任务要求和标注信息格式, 确保他们充分理解相关任务. 同时, 还预先清理了数据集, 主动剔除了诸如卡通、游戏截图、拼图以及佛像等非正常图像. 在粗标注阶段, 我们邀请了 5 名具备计算机视觉领域经验的研究生 (2 男 3 女, 平均年龄 24 岁) 参与. 参与者通过投票的方式确定数据库中每张图像的显著目标. 如果某一目标获得半数以上的票数, 即被认为是显著目标, 并进一步确定其具体对象.

接下来, 参与者使用 LabelMe (<https://github.com/wkentaro/labelme>) 标注工具, 采用多边形粗略地标记出每张图像中的显著物体 (见图 6(b)). 第 2 阶段, 邀请了 15 个大学生, 对每个显著目标进行逐像素的精确边界标注 (见图 6(c)), 并同时进行了语义分割标注. 此外, 我们还收集了数据库中图像的属性信息^[40,70,71] (见表 2), 以便分析测试模型在开放视觉世界中应对不同场景的鲁棒性. 在标注完成后, 进行交叉验证, 以进一步提高标注的平滑性和一致性. 参与者轮流检查其他参与者的标注, 并在第 2 个标注阶段允许彼此沟通, 以确保标注的准确性和一致性. 最终,

我们为数据集中的图像生成了人工标注结果 (二值图和语义分割图) 以及相关属性信息. 图 7 展示了 GHS-CV 数据集中各相关属性的占比, 内圈表示语义信息属性占比, 外圈表示显著对象属性信息占比. 从图 7 中可以看出, 显著目标检测模型在识别大目标时表现出可靠的性能, 但在小目标检测方面仍然存在显著的缺陷. 然而, 小目标的检测仍然是许多显著目标检测模型的一个缺陷. 从语义类别的角度来看, 语义分割模型在处理只有单一语义信息的图像时表现不佳, 无语义信息 (NS) 包括了风景图和 PASCAL VOC 类别之外的其他类别.



图 6 GHS-CV 区域级和像素级信息标注过程

表 2 GHS-CV 区域级与像素级标签属性描述	
属性	属性描述
大目标 (big object, BO)	目标对象与背景之比大于0.5
多目标 (multiple object, MO)	图像存在两个及以上目标对象
小目标 (small object, SO)	目标对象与背景之比小于0.2
无对象 (null object, NO)	图像没有显著目标对象
无语义信息 (null semanticity, NS)	图像没有语义信息
单语义信息 (unique semanticity, US)	图像只包含一种语义信息
多语义信息 (multiple semanticity, MS)	图像包含两种以上的语义信息

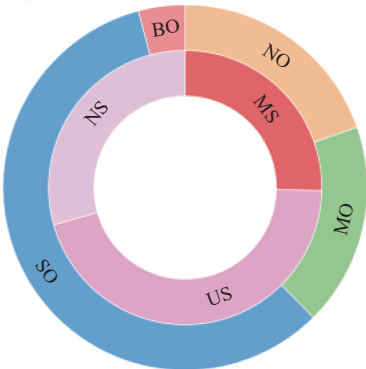


图 7 GHS-CV 数据集的属性分布

在图像分类任务中, 对图像进行标注存在类别定义模糊、复杂背景干扰、目标遮挡和数据不完整等问题, 使数据集进行图像级标注产生了极大的困扰. 我们使用一种更加便捷的方法消除类别定位模糊的问题, 即对图像进行多标签分类, 本文对图像规定最多 3 个图像级标签. 为了方便参与者更好地标注, 专门设计了一个工具 (<https://github.com/wentaozhu-jxufecai/GHS-CV>), 这个工具提供了其他模型的预测结果作为参考, 还包括查询 wnid (WordTree 中类别编号). 为了减少标注问题对实验结果的影响, 结合 WordTree 的结构, 本文提出了一种新型的图像分类任务精确度计算方法 (详细介绍见第 3.1.2 及 3.1.3 节).

4 实验评估

4.1 多任务驱动的通用难样本挖掘评估

本节探讨了通用视觉任务难样本的构建及其在多任务场景下的表现. 与单一视觉任务难样本挖掘不同, 通用视觉任务难样本不只是简单的对单一任务的度量指标相加 (如公式 (1)), 这样会淡化不同视觉任务之间的联系. 考虑到这个问题, 我们通过联合优化从粗到细的 3 种粒度的视觉任务, 共同识别在多个模型中表现缺陷的样本. 本文采用图像分类任务、显著目标检测任务和语义分割任务模型, 通过不同任务之间的关联, 自动挖掘出通用视觉任务难样本.

如表 3 所示, 本文的难样本数据集在几个联合视觉任务中的模型测试, 这些模型包括进行图像分类任务的 VGG16^[58]、ResNet101^[3]和 ViT^[72]模型; 进行显著目标检测任务的 PoolNet^[63]、VST^[62]、UDNet^[73]和 ADMNet^[74]模

型; 进行语义分割任务的 U2PL^[68]、CorrMatch^[67]和 DeepLabv3+^[65]模型. 除此之外, 我们还测试了性能优异的大模型 SAM^[75]和 SAM 2^[76]. 实验结果表明, 在 GHS-CV₁ (仅含有显著目标图像) 和 GHS-CV₂ (同时包含有显著目标和无显著目标图像) 的测试中, 所有模型性能均出现了不同程度的下降; 语义分割任务模型在 GHS-CV 上的性能下降尤为显著, 表明该任务对通用难样本更为敏感. SAM^[75]的表现远超其他模型, 这得益于大模型的学习能力远超其他任务模型. 从整体平均结果来看, GHS-CV 数据集相较于语义分割难样本集 SS-C^[7]和显著目标检测难样本集 SOD-C^[69], 在跨任务评估中展现出更强的挑战性, 更能揭示模型的潜在不足.

表 3 不同任务模型在 GHS-CV 数据集上测试表现

模型	数据集	有显著目标	无显著目标	Acc/ F_{β} /mIoU/gIoU
VGG16 ^[58]	GHS-CV ₁	√	—	0.559
	GHS-CV ₂	√	√	0.703
ResNet101 ^[3]	GHS-CV ₁	√	—	0.560
	GHS-CV ₂	√	√	0.706
ViT ^[72]	GHS-CV ₁	√	—	0.570
	GHS-CV ₂	√	√	0.727
PoolNet ^[63]	GHS-CV ₁	√	—	0.282/0.190
	GHS-CV ₂	√	√	0.223/0.180
VST ^[62]	GHS-CV ₁	√	—	0.335/0.264
	GHS-CV ₂	√	√	0.263/0.259
UDNet ^[73]	GHS-CV ₁	√	—	0.367/0.265
	GHS-CV ₂	√	√	0.291/0.258
ADMNet ^[74]	GHS-CV ₁	√	—	0.223/0.178
	GHS-CV ₂	√	√	0.175/0.176
U2PL ^[68]	GHS-CV ₁	√	—	0.110/0.310
	GHS-CV ₂	√	√	0.104/0.307
CorrMatch ^[67]	GHS-CV ₁	√	—	0.246/0.315
	GHS-CV ₂	√	√	0.227/0.302
DeepLabv3+ ^[65]	GHS-CV ₁	√	—	0.193/0.272
	GHS-CV ₂	√	√	0.183/0.265
SAM ^[75]	GHS-CV ₁	√	—	0.826
	GHS-CV ₂	√	√	0.648
SAM 2 ^[76]	GHS-CV ₁	√	—	0.486
	GHS-CV ₂	√	√	0.382

注: GHS-CV₁ 只包含显著目标的图像, GHS-CV₂ 包含无显著目标图像和显著目标图像; 绿色代表准确率 Acc (↑), 蓝色代表gIoU (↑), 黑色代表 F_{β} (↑), 红色代表mIoU (↑); ↑表示值越大越好

我们使用多模态大模型 X-SAM^[77]在本文提出的难样本集 GHS-CV (只含显著目标) 上进行了测试. 具体而言, 在 Referring Segmentation 模式下输入自然语言提示 (如“Can you segment the prominent targets in the picture?”), 并要求模型根据提示完成显著目标的分割任务. 与 MS COCO testB 数据集测试比较, X-SAM 在 GHS-CV 上的表现出现显著下降, 在 MS COCO testB 上的条件交并比 (cIoU) 指标达到 0.839, 而在 GHS-CV 上的 cIoU 仅为 0.359. 该结果证明了将本文提出的方法迁移至多模态任务的通用难样本挖掘的可行性.

4.2 单一视觉任务难样本

在本节中, 我们考虑单一视觉任务的难样本是否对其他任务具有挑战性, 通过实验验证单一视觉任务难样本在不同视觉任务中的表现. 例如, 在显著目标检测任务中表现较差的样本, 是否同样会对图像分类或语义分割任务带来挑战? 有研究已经提出了在语义分割任务和显著目标检测任务的难样本数据集^[7,69]. 此外, 本文将这两类单任务难样本与所提出的通用视觉任务难样本集 (GHS-CV) 进行对比分析.

在语义分割任务中, Yan 等人^[7]提出了数据集 SS-C, 这个数据集是针对语义分割任务模型识别失败情况构建的. 本文利用 SS-C 分别在显著目标检测任务模型和图像分类任务模型上进行测试. 其中, 对于显著目标检测任务模型在 SS-C 数据集的度量指标, 我们选用了 gIoU 度量方法. 在图 8(a) 中, 可以直观地看出显著目标检测模型的表现, 而对于图像分类任务模型, 使用不同颜色代表图像分类模型预测的准确性 (绿色代表预测正确, 红色代表预测错误, 黄棕色代表不确定). 在显著目标检测任务中, 张海艳^[69]的研究提出的显著目标检测难样本数据集 SOD-C, 并分别在语义分割任务模型和图像分类任务模型上进行了测试. 对于语义分割任务模型在 SOD-C 的度量, 我们也使用了与上面相同度量方式.

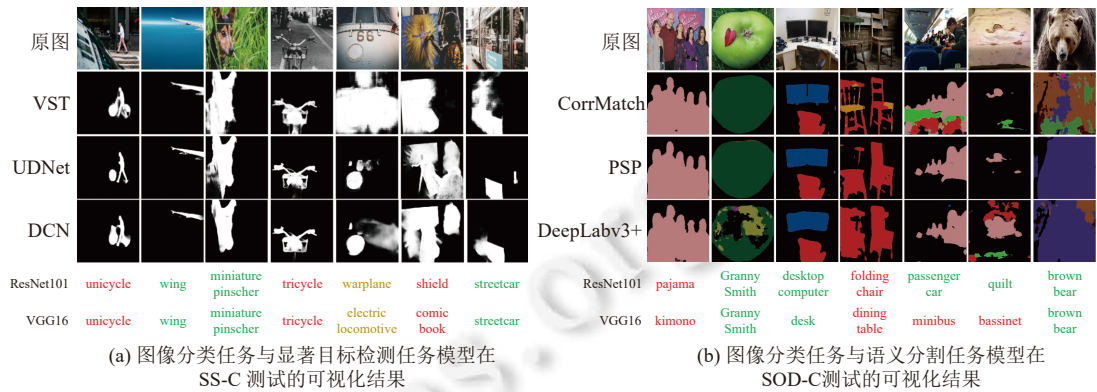


图 8 单一视觉任务难样本测试的可视化结果

如图 8(b) 所示, 在第 1、3、4 列中, 语义分割模型对图像进行了完美的分割; 而在第 2、3、7 列中, 图像分类模型对图像进行了精确的预测. 这些对比展示了不同任务在处理显著性目标检测任务中的优势和局限性, 特别是在面对复杂场景和难样本时的表现差异. 表 4 进一步展示了不同任务模型在 SS-C 和 SOD-C 数据集上的性能比较. 从显著目标检测任务模型测试 SS-C 和 SOD-C 单一难样本数据集测试结果, 可以发现显著目标检测任务模型在 SS-C 数据集比 SOD-C 数据集上的性能数值低, 需要注意的是, 显著目标检测任务模型与 SOD-C 使用的度量指标不同, 所以并不能说明 SS-C 数据集对显著目标检测任务模型也具有更大的挑战. 我们对单任务难样本与本文提出的通用视觉任务难样本进行了对比实验, 如表 4 所示. 实验中, 选择的模型覆盖了语义分割和显著性检测的主流方法, 确保具有代表性. 结果显示, 通用视觉任务难样本 (GHS-CV) 在跨任务模型测试中, 比单任务难样本 (SS-C 和 SOD-C) 更能暴露模型的弱点. 这也说明单任务难样本在跨任务模型中可能失效, 即某些在单一任务中困难的样本, 在其他任务中可能很容易预测.

表 4 不同任务模型在 SS-C、SOD-C 和 GHS-CV 数据集测试验证

任务	模型	SS-C	SOD-C	GHS-CV	任务	模型	SS-C	SOD-C	GHS-CV
ISOD	PoolNet ^[63]	0.386	0.546	0.282	ISS	U2PL ^[68]	0.216	0.479	0.110
	VST ^[62]	0.558	0.699	0.335		CorrMatch ^[67]	0.440	0.486	0.246
	UDNet ^[73]	0.481	0.691	0.367		DeepLabv3+ ^[65]	0.390	0.409	0.193

注: 蓝色代表gIoU (↑), 黑色代表 F_{β} (↑), 红色代表mIoU (↑)

综上所述, 单一视觉任务的挑战性并不一定能够直接推广到其他任务. 为此, 本文研究通用计算机视觉任务的挑战, 筛选能够对通用计算机视觉任务构成挑战的样本.

4.3 其他计算机视觉任务对比实验与分析

4.3.1 伪装目标检测任务

为了验证本文的难样本数据集对其他视觉任务也具有挑战性, 我们使用了伪装目标检测任务模型 FSPNet^[78]、POPNet^[79]、ZoomNet^[80]、SINet^[81]进行测试. 测试结果如后文表 5 所示, 清晰地表明本文的难样本数据集对这些伪

装目标检测模型提出了较高的挑战. 这进一步验证了数据集在不同模型和任务中的普适性和针对性. 具体而言, 表 5 中的测试结果显示, 相较于常规数据集, 这些难样本在伪装目标检测任务中能够有效暴露模型的潜在不足.

4.3.2 目标检测任务

如表 6 所示, 我们使用 YOLO 系列目标检测模型, 包括 YOLOv9^[82]、YOLOv10^[83]和 YOLO11 (<https://github.com/ultralytics/ultralytics>), 在本文构建的难样本数据集上进行测试. 目标检测结果包含检测框的位置、目标类别以及对应的置信度分数. 为了直观比较不同模型的检测性能, 我们采用 greedy-IoU (gIoU) 作为度量指标, 对模型的检测能力进行评估. gIoU 指标能够更全面地衡量检测框与真实目标之间的重叠程度, 尤其在难样本数据集集中的复杂目标场景下, 更能合理地反映模型性能差异. 测试结果显示, YOLO 系列模型在本文数据集上的表现还是偏低的. 我们对 YOLOv9 进行了可视化结果分析 (如图 9 所示), 可以观察到, 目标检测模型识别结果存在 4 种情况, 分别是目标检测位置和类别都识别正确 (图 9(a)); 目标检测类别识别错误 (图 9(b)); 目标检测位置识别错误 (图 9(c)); 目标检测位置和类别都识别错误 (图 9(d)).

表 5 伪装目标检测任务模型在 GHS-CV 数据集上的表现

模型	数据集	有显著目标	无显著目标	F_{β} (↑)
FSPNet ^[78]	GHS-CV ₁	√	—	0.134
	GHS-CV ₂	√	√	0.099
POPNet ^[79]	GHS-CV ₁	√	—	0.218
	GHS-CV ₂	√	√	0.171
ZoomNet ^[80]	GHS-CV ₁	√	—	0.226
	GHS-CV ₂	√	√	0.260
SINet ^[81]	GHS-CV ₁	√	—	0.304
	GHS-CV ₂	√	√	0.238

表 6 目标检测任务模型在 GHS-CV 数据集上的表现

模型	数据集	有显著目标	无显著目标	gIoU (↑)
YOLOv9 ^[82]	GHS-CV ₁	√	—	0.345
	GHS-CV ₂	√	√	0.357
YOLOv10 ^[83]	GHS-CV ₁	√	—	0.384
	GHS-CV ₂	√	√	0.407
YOLO11	GHS-CV ₁	√	—	0.360
	GHS-CV ₂	√	√	0.373



图 9 YOLOv9 在 GHS-CV 数据集上测试的可视化结果

4.4 可视化分析

我们对本文构建的难样本数据集在多种视觉任务模型上的表现进行了可视化对比,直观地展示了这些模型的性能及其局限性.如图 10 所示(最后两列分别为显著目标和语义分割的真值图),从图中可以观察到,大多数视觉任务模型在检测大目标时表现较好,但对于小目标的检测则受到背景复杂程度的显著干扰.这表明,现有视觉模型在应对复杂背景中的小目标检测方面仍有较大的改进空间.值得注意的是,伪装目标检测模型在该数据集上的表现极差,这可能是由于该类模型在任务设计上更关注于隐藏目标,从而对显著目标的处理较为薄弱.此外,目标检测模型在不完整目标的识别能力显著低于完整目标,这进一步反映了模型在处理目标完整性变化时的适应能力不足.这些结果揭示了现有视觉任务模型在复杂场景下的性能瓶颈,突出了在未来研究中优化小目标检测、背景干扰抑制以及不完整目标识别的挑战性.

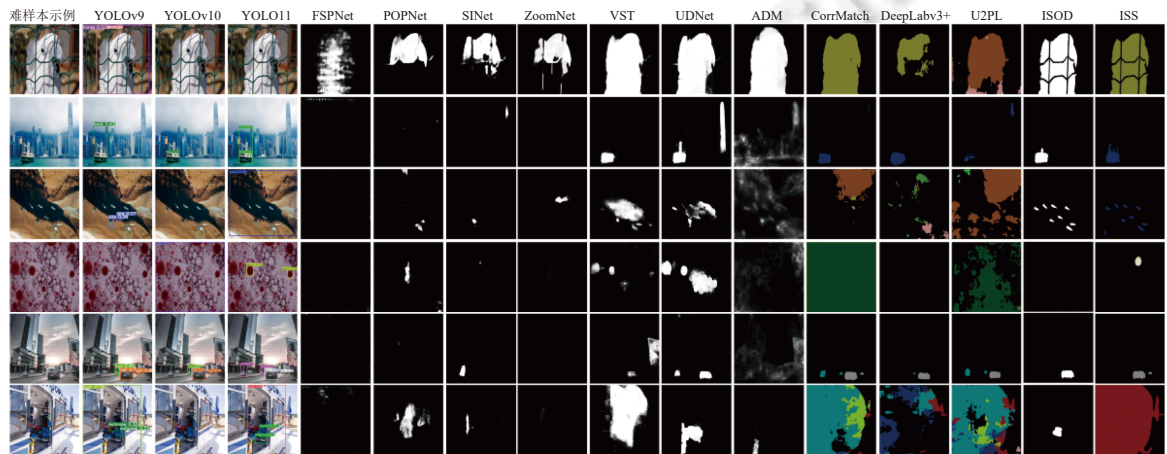


图 10 不同视觉任务模型在 GHS-CV 数据集上测试的可视化比较

5 总 结

在本文中,我们提出了一种基于最大差异化竞争的通用视觉难样本挖掘方法,针对现有计算机视觉模型在封闭测试集下泛化能力不足和评估不公平的问题,利用图像级、区域级和像素级信息等多任务联合分析,自动挖掘具有挑战性的通用视觉任务难样本.基于该方法,构建了一个通用视觉难样本数据集(GHS-CV),该数据集包含详细的像素级、区域级和图像级注释,能够广泛应用于计算机视觉任务,为模型的公平评估和缺陷发现提供了强有力的支持.实验结果表明,GHS-CV 数据集显著削弱了伪装目标检测模型等在复杂场景下的表现,验证了其在不同任务中的迁移性和泛化性.

本文的研究为计算机视觉模型的评估与优化开辟了新思路,通过难样本挖掘,深入揭示了模型在复杂场景中的薄弱环节,为进一步改进设计提供了重要依据.未来的研究可以将该方法扩展到更多的视觉任务和实时应用中,以充分挖掘其潜在价值.

References

[1] Yan JB, Rao JL, Chen JJ, Tan ZW, Liu WD, Fang YM. Multitask auxiliary network for perceptual quality assessment of non-uniformly distorted omnidirectional images. IEEE Trans. on Circuits and Systems for Video Technology, 2025, 35(3): 2782–2793. [doi: 10.1109/TCSVT.2024.3497994]

[2] Shen Z, Cui CR, Dong GX, Yu J, Huang J, Yin YL. Unified image aesthetic and emotional prediction based on deep multi-task learning. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2494–2506 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/6487.htm> [doi: 10.13328/j.cnki.jos.006487]

[3] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision

- and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90)]
- [4] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2818–2826. [doi: [10.1109/CVPR.2016.308](https://doi.org/10.1109/CVPR.2016.308)]
- [5] Yan Y, Deng C, Li L, Zhu LK, Ye B. Survey of image semantic segmentation methods in the deep learning era. Journal of Image and Graphics, 2023, 28(11): 3342–3362 (in Chinese with English abstract). [doi: [10.11834/jig.220292](https://doi.org/10.11834/jig.220292)]
- [6] Yan JB, Fang YM, Liu XL, Yao YR, Sui XJ. A survey on recent advances in video quality assessment. Chinese Journal of Computers, 2023, 46(10): 2196–2224 (in Chinese with English abstract). [doi: [10.11897/SP.J.1016.2023.02196](https://doi.org/10.11897/SP.J.1016.2023.02196)]
- [7] Yan JB, Zhong Y, Fang YM, Wang ZY, Ma KD. Exposing semantic segmentation failures via maximum discrepancy competition. Int'l Journal of Computer Vision, 2021, 129(5): 1768–1786. [doi: [10.1007/s11263-021-01450-2](https://doi.org/10.1007/s11263-021-01450-2)]
- [8] Everingham M, Eslami SMA, van Gool L, Williams CKI, Winn J, Zisserman A. The PASCAL visual object classes challenge: A retrospective. Int'l Journal of Computer Vision, 2015, 111(1): 98–136. [doi: [10.1007/s11263-014-0733-5](https://doi.org/10.1007/s11263-014-0733-5)]
- [9] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: [10.1007/978-3-319-10602-1_48](https://doi.org/10.1007/978-3-319-10602-1_48)]
- [10] Wang HT, Chen TL, Wang ZY, Ma KD. I am going MAD: Maximum discrepancy competition for comparing classifiers adaptively. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020. 1–13.
- [11] Wang ZH, Jiang QP, Zhao SS, Feng WS, Lin WS. Deep blind image quality assessment powered by online hard example mining. IEEE Trans. on Multimedia, 2023, 25: 4774–4784. [doi: [10.1109/TMM.2023.3257564](https://doi.org/10.1109/TMM.2023.3257564)]
- [12] Huang G, Liu Z, van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269. [doi: [10.1109/CVPR.2017.243](https://doi.org/10.1109/CVPR.2017.243)]
- [13] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: [10.1109/TPAMI.2016.2577031](https://doi.org/10.1109/TPAMI.2016.2577031)]
- [14] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788. [doi: [10.1109/CVPR.2016.91](https://doi.org/10.1109/CVPR.2016.91)]
- [15] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. SSD: Single shot MultiBox detector. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 21–37. [doi: [10.1007/978-3-319-46448-0_2](https://doi.org/10.1007/978-3-319-46448-0_2)]
- [16] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139–144. [doi: [10.1145/3422622](https://doi.org/10.1145/3422622)]
- [17] Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv:1312.6114, 2013.
- [18] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015. 1–11.
- [19] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R. Intriguing properties of neural networks. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff: OpenReview.net, 2014.
- [20] Wang ZB, Wang X, Ma JJ, Qin Z, Ren J, Ren K. Survey on adversarial example attack for computer vision systems. Chinese Journal of Computers, 2023, 46(2): 436–468 (in Chinese with English abstract). [doi: [10.11897/SP.J.1016.2023.00436](https://doi.org/10.11897/SP.J.1016.2023.00436)]
- [21] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma SA, Huang ZH, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet large scale visual recognition challenge. Int'l Journal of Computer Vision, 2015, 115(3): 211–252. [doi: [10.1007/s11263-015-0816-y](https://doi.org/10.1007/s11263-015-0816-y)]
- [22] Wang LJ, Lu HC, Wang YF, Feng MY, Wang D, Yin BC, Ruan X. Learning to detect salient objects with image-level supervision. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3796–3805. [doi: [10.1109/CVPR.2017.404](https://doi.org/10.1109/CVPR.2017.404)]
- [23] Chen TY, Zhou HY, He MR, Zhang SH, Yan K, Zhou MM, Li JX. Resonant attack: Reveal cross-modal vulnerability of CLIP. Chinese Journal of Computers, 2023, 46(12): 2597–2611 (in Chinese with English abstract). [doi: [10.11897/SP.J.1016.2023.02597](https://doi.org/10.11897/SP.J.1016.2023.02597)]
- [24] Kamann C, Rother C. Benchmarking the robustness of semantic segmentation models. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8825–8835. [doi: [10.1109/CVPR42600.2020.00885](https://doi.org/10.1109/CVPR42600.2020.00885)]
- [25] Fan H, Yang F, Chu P, Lin YW, Yuan L, Ling HB. TracKlinic: Diagnosis of challenge factors in visual tracking. In: Proc. of the 2021 IEEE Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2021. 969–978. [doi: [10.1109/WACV48630.2021.00101](https://doi.org/10.1109/WACV48630.2021.00101)]
- [26] Gu JJ, Cai HM, Chen HY, Ye XX, Ren JS, Dong C. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 633–651. [doi: [10.1007/978-3-030-58621-8_37](https://doi.org/10.1007/978-3-030-58621-8_37)]
- [27] Ma KD, Duanmu ZF, Wu QB, Wang Z, Yong HW, Li HL, Zhang L. Waterloo exploration database: New challenges for image quality

- assessment models. *IEEE Trans. on Image Processing*, 2017, 26(2): 1004–1016. [doi: [10.1109/TIP.2016.2631888](https://doi.org/10.1109/TIP.2016.2631888)]
- [28] Mađry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: *Proc. of the 6th Int'l Conf. on Learning Representations*. Vancouver: OpenReview.net, 2018.
- [29] Wang ZH, Ma KD. Active fine-tuning from gMAD examples improves blind image quality assessment. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2022, 44(9): 4577–4590. [doi: [10.1109/TPAMI.2021.3071759](https://doi.org/10.1109/TPAMI.2021.3071759)]
- [30] Zhang WX, Ma KD, Zhai GT, Yang XK. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Trans. on Image Processing*, 2021, 30: 3474–3486. [doi: [10.1109/TIP.2021.3061932](https://doi.org/10.1109/TIP.2021.3061932)]
- [31] Su YP, Chen GS, Zhao T. SA-YOLO: Self-adaptive loss object detection method under imbalance samples. *Journal of University of Chinese Academy of Sciences*, 2024, 41(3): 411–426 (in Chinese with English abstract). [doi: [10.7523/j.ucas.2023.013](https://doi.org/10.7523/j.ucas.2023.013)]
- [32] Smirnov E, Ivanova E, Melnikov A, Kalinovskiy I, Oleinik A, Luckyanets E. Hard example mining with auxiliary embeddings. In: *Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*. Salt Lake City: IEEE, 2018. 37–46. [doi: [10.1109/CVPRW.2018.00013](https://doi.org/10.1109/CVPRW.2018.00013)]
- [33] Xia QM, Ye W, Wu H, Zhao SJ, Xing LY, Huang X, Deng JH, Li X, Wen CL, Wang C. HINTED: Hard instance enhanced detector with mixed-density feature fusion for sparsely-supervised 3D object detection. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2024. 15321–15330. [doi: [10.1109/CVPR52733.2024.01451](https://doi.org/10.1109/CVPR52733.2024.01451)]
- [34] Hooker S, Courville A, Clark G, Dauphin Y, Frome A. What do compressed deep neural networks forget? *arXiv:1911.05248*, 2019.
- [35] Cao PB, Li DQ, Ma KD. Image quality assessment: Integrating model-centric and data-centric approaches. In: *Proc. of the 2024 Conf. on Parsimony and Learning*. Hong Kong: PMLR, 2024. 529–541.
- [36] Guo Y, Kong JY, Wang Y, Duan YP. Take the bull by the horns: Learning to segment hard samples. In: *Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2025. 15642–15652. [doi: [10.1109/CVPR52734.2025.01458](https://doi.org/10.1109/CVPR52734.2025.01458)]
- [37] Hoiem D, Chodpathumwan Y, Dai QY. Diagnosing error in object detectors. In: *Proc. of the 12th European Conf. on Computer Vision*. Florence: Springer, 2012. 340–353. [doi: [10.1007/978-3-642-33712-3_25](https://doi.org/10.1007/978-3-642-33712-3_25)]
- [38] Zhang SS, Benenson R, Omran M, Hosang J, Schiele B. How far are we from solving pedestrian detection? In: *Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016. 1259–1267. [doi: [10.1109/CVPR.2016.141](https://doi.org/10.1109/CVPR.2016.141)]
- [39] Ronchi MR, Perona P. Benchmarking and error diagnosis in multi-instance pose estimation. In: *Proc. of the 2017 IEEE Int'l Conf. on Computer Vision*. Venice: IEEE, 2017. 369–378. [doi: [10.1109/ICCV.2017.48](https://doi.org/10.1109/ICCV.2017.48)]
- [40] Fan DP, Cheng MM, Liu JJ, Gao SH, Hou QB, Borji A. Salient objects in clutter: Bringing salient object detection to the foreground. In: *Proc. of the 15th European Conf. on Computer Vision*. Munich: Springer, 2018. 196–212. [doi: [10.1007/978-3-030-01267-0_12](https://doi.org/10.1007/978-3-030-01267-0_12)]
- [41] Wang WG, Lai QX, Fu HZ, Shen JB, Ling HB, Yang RG. Salient object detection in the deep learning era: An in-depth survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2022, 44(6): 3239–3259. [doi: [10.1109/TPAMI.2021.3051099](https://doi.org/10.1109/TPAMI.2021.3051099)]
- [42] Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: *Proc. of the 5th Int'l Conf. on Learning Representations*. Toulon: OpenReview.net, 2017. 1–12.
- [43] Recht B, Roelofs R, Schmidt L, Shankar V. Do ImageNet classifiers generalize to ImageNet? In: *Proc. of the 36th Int'l Conf. on Machine Learning*. Long Beach: PMLR, 2019. 5389–5400.
- [44] Krizhevsky A. Learning Multiple Layers of Features from Tiny Images. Toronto: University of Toronto, 2009.
- [45] Deng J, Dong W, Socher R, Li LJ, Kai L, Li FF. ImageNet: A large-scale hierarchical image database. In: *Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition*. Miami: IEEE, 2009. 248–255. [doi: [10.1109/CVPR.2009.5206848](https://doi.org/10.1109/CVPR.2009.5206848)]
- [46] OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2023.
- [47] Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv:2312.11805*, 2023.
- [48] Wang G, Cheng SJ, Zhan XY, Li XG, Song S, Liu Y. OpenChat: Advancing open-source language models with mixed-quality data. In: *Proc. of the 12th Int'l Conf. on Learning Representations*. Vienna: OpenReview.net, 2024. 1–20.
- [49] Huang YZ, Bai YZ, Zhu ZH, Zhang JL, Zhang JH, Su TJ, Liu JT, Lv CC, Lei JY, Fu Y, Sun MS, He JX. C-EVAL: A multi-level multi-discipline Chinese evaluation suite for foundation models. In: *Proc. of the 37th Int'l Conf. on Neural Information Processing Systems*. New Orleans: Curran Associates Inc., 2023. 2749.
- [50] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. Measuring massive multitask language understanding. In: *Proc. of the 9th Int'l Conf. on Learning Representations*. OpenReview.net, 2021. 1–27.
- [51] Srivastava A, Rastogi A, Rao A, *et al.* Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv:2206.04615*, 2022.
- [52] Golan T, Raju PC, Kriegeskorte N. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proc. of the National Academy of Sciences of the United States of America*, 2020, 117(47): 29330–29337. [doi: [10.1073/pnas.1912334117](https://doi.org/10.1073/pnas.1912334117)]

- [53] McKeeman WM. Differential testing for software. *Digital Technical Journal*, 1998, 10(1): 100–107.
- [54] Ma KD, Duanmu ZF, Wang Z, Wu QB, Liu WT, Yong HW, Li HL, Zhang L. Group maximum differentiation competition: Model comparison with few samples. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2020, 42(4): 851–864. [doi: [10.1109/TPAMI.2018.2889948](https://doi.org/10.1109/TPAMI.2018.2889948)]
- [55] Wang Z, Simoncelli EP. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 2008, 8(12): 8. [doi: [10.1167/8.12.8](https://doi.org/10.1167/8.12.8)]
- [56] Feng KH, Ding KY, Tan HZ, Ma KD, Wang ZH, Guo SQ, Chen YZ, Sun G, Zheng GZ, Zhang Q, Chen HJ. Sample-efficient human evaluation of large language models via maximum discrepancy competition. In: *Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics*. Vienna: ACL, 2025. 10913–10947. [doi: [10.18653/v1/2025.acl-long.535](https://doi.org/10.18653/v1/2025.acl-long.535)]
- [57] Zhang CN, Cho J, Puspitasari FD, Zheng S, Li CH, Qiao Y, Kang T, Shan XR, Zhang CS, Qin CY, Rameau F, Lee LH, Bae SH, Hong CS. A survey on segment anything model (SAM): Vision foundation model meets prompt engineering. *arXiv:2306.06211*, 2023.
- [58] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: *Proc. of the 3rd Int'l Conf. on Learning Representations*. San Diego: OpenReview.net, 2015. 1–14.
- [59] Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. In: *Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018. 8697–8710. [doi: [10.1109/CVPR.2018.00907](https://doi.org/10.1109/CVPR.2018.00907)]
- [60] Liu CX, Zoph B, Neumann M, Shlens J, Hua W, Li LJ, Fei-Fei L, Yuille A, Huang J, Murphy K. Progressive neural architecture search. In: *Proc. of the 15th European Conf. on Computer Vision*. Munich: Springer, 2018. 19–35. [doi: [10.1007/978-3-030-01246-5_2](https://doi.org/10.1007/978-3-030-01246-5_2)]
- [61] Zhao JX, Liu JJ, Fan DP, Cao Y, Yang JF, Cheng MM. EGNNet: Edge guidance network for salient object detection. In: *Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision*. Seoul: IEEE, 2019. 8778–8787. [doi: [10.1109/ICCV.2019.00887](https://doi.org/10.1109/ICCV.2019.00887)]
- [62] Liu N, Zhang N, Wan KY, Shao L, Han JW. Visual saliency Transformer. In: *Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision*. Montreal: IEEE, 2021. 4702–4712. [doi: [10.1109/ICCV48922.2021.00468](https://doi.org/10.1109/ICCV48922.2021.00468)]
- [63] Liu JJ, Hou QB, Cheng MM, Feng JS, Jiang JM. A simple pooling-based design for real-time salient object detection. In: *Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019. 3912–3921. [doi: [10.1109/CVPR.2019.00404](https://doi.org/10.1109/CVPR.2019.00404)]
- [64] Wu Z, Su L, Huang QM. Decomposition and completion network for salient object detection. *IEEE Trans. on Image Processing*, 2021, 30: 6226–6239. [doi: [10.1109/TIP.2021.3093380](https://doi.org/10.1109/TIP.2021.3093380)]
- [65] Chen LC, Zhu YK, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proc. of the 15th European Conf. on Computer Vision*. Munich: Springer, 2018. 833–851. [doi: [10.1007/978-3-030-01234-2_49](https://doi.org/10.1007/978-3-030-01234-2_49)]
- [66] Zhao HS, Shi JP, Qi XJ, Wang XG, Jia JY. Pyramid scene parsing network. In: *Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017. 6230–6239. [doi: [10.1109/CVPR.2017.660](https://doi.org/10.1109/CVPR.2017.660)]
- [67] Sun BY, Yang YQ, Zhang L, Cheng MM, Hou QB. CorrMatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In: *Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2024. 3097–3107. [doi: [10.1109/CVPR52733.2024.00299](https://doi.org/10.1109/CVPR52733.2024.00299)]
- [68] Wang YC, Wang HC, Shen YJ, Fei JJ, Li W, Jin GQ, Wu LW, Zhao R, Le XY. Semi-supervised semantic segmentation using unreliable pseudo-labels. In: *Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. New Orleans: IEEE, 2022. 4238–4247. [doi: [10.1109/CVPR52688.2022.00421](https://doi.org/10.1109/CVPR52688.2022.00421)]
- [69] Zhang HY. Research on extensibility of salient object detection models by deep learning [MS. Thesis]. Nanchang: Jiangxi University of Finance and Economics, 2022 (in Chinese with English abstract). [doi: [10.27175/d.cnki.gjxcu.2022.001183](https://doi.org/10.27175/d.cnki.gjxcu.2022.001183)]
- [70] Perazzi F, Pont-Tuset J, McWilliams B, van Gool L, Gross M, Sorkine-Hornung A. A benchmark dataset and evaluation methodology for video object segmentation. In: *Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016. 724–732. [doi: [10.1109/CVPR.2016.85](https://doi.org/10.1109/CVPR.2016.85)]
- [71] Wu Y, Lim J, Yang MH. Online object tracking: A benchmark. In: *Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition*. Portland: IEEE, 2013. 2411–2418. [doi: [10.1109/CVPR.2013.312](https://doi.org/10.1109/CVPR.2013.312)]
- [72] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houtsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: *Proc. of the 9th Int'l Conf. on Learning Representations*. OpenReview.net, 2021. 1–22.
- [73] Fang YM, Zhang HY, Yan JB, Jiang WH, Liu Y. UDNet: Uncertainty-aware deep network for salient object detection. *Pattern Recognition*, 2023, 134: 109099. [doi: [10.1016/j.patcog.2022.109099](https://doi.org/10.1016/j.patcog.2022.109099)]

- [74] Zhou XF, Shen KY, Liu Z. ADMNet: Attention-guided densely multi-scale network for lightweight salient object detection. *IEEE Trans. on Multimedia*, 2024, 26: 10828–10841. [doi: [10.1109/TMM.2024.3413529](https://doi.org/10.1109/TMM.2024.3413529)]
- [75] Kirillov A, Mintun E, Ravi N, Mao HZ, Rolland C, Gustafson L, Xiao TT, Whitehead S, Berg AC, Lo WY, Dollár P, Girshick R. Segment anything. In: *Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision*. Paris: IEEE, 2023. 3992–4003. [doi: [10.1109/ICCV51070.2023.00371](https://doi.org/10.1109/ICCV51070.2023.00371)]
- [76] Ravi N, Gabeur V, Hu YT, Hu RH, Ryali C, Ma TY, Khedr H, Rädle R, Rolland C, Gustafson L, Mintun E, Pan JT, Alwala KV, Carion N, Wu CY, Girshick R, Dollar P, Feichtenhofer C. SAM 2: Segment anything in images and videos. In: *Proc. of the 13th Int'l Conf. on Learning Representations*. Singapore: OpenReview.net, 2025.
- [77] Wang H, Qiao LM, Jie ZQ, Huang ZJ, Feng CJ, Zheng QF, Ma L, Lan XY, Liang XD. X-SAM: From segment anything to any segmentation. *arXiv:2508.04655*, 2025.
- [78] Huang Z, Dai H, Xiang TZ, Wang S, Chen HX, Qin J, Xiong H. Feature shrinkage pyramid for camouflaged object detection with Transformers. In: *Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Vancouver: IEEE, 2023. 5557–5566. [doi: [10.1109/CVPR52729.2023.00538](https://doi.org/10.1109/CVPR52729.2023.00538)]
- [79] Wu ZW, Paudel DP, Fan DP, Wang JJ, Wang S, Demoncaux C, Timofte R, van Gool L. Source-free depth for object pop-out. In: *Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision*. Paris: IEEE, 2023. 1032–1042. [doi: [10.1109/ICCV51070.2023.00101](https://doi.org/10.1109/ICCV51070.2023.00101)]
- [80] Pang YW, Zhao XQ, Xiang TZ, Zhang LH, Lu HC. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: *Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. New Orleans: IEEE, 2022. 2150–2160. [doi: [10.1109/CVPR52688.2022.00220](https://doi.org/10.1109/CVPR52688.2022.00220)]
- [81] Fan DP, Ji GP, Cheng MM, Shao L. Concealed object detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2022, 44(10): 6024–6042. [doi: [10.1109/TPAMI.2021.3085766](https://doi.org/10.1109/TPAMI.2021.3085766)]
- [82] Wang CY, Yeh IH, Liao HYM. YOLOv9: Learning what you want to learn using programmable gradient information. In: *Proc. of the 18th European Conf. on Computer Vision*. Milan: Springer, 2025. 1–21. [doi: [10.1007/978-3-031-72751-1_1](https://doi.org/10.1007/978-3-031-72751-1_1)]
- [83] Wang A, Chen H, Liu LH, Chen K, Lin ZJ, Han JG, Ding GG. YOLOv10: Real-time end-to-end object detection. In: *Proc. of the 38th Int'l Conf. on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 2025. 3429.

附中文参考文献

- [2] 申朕, 崔超然, 董桂鑫, 余俊, 黄瑾, 尹义龙. 基于深度多任务学习的图像美感与情感联合预测研究. *软件学报*, 2023, 34(5): 2494–2506. <http://www.jos.org.cn/1000-9825/6487.htm> [doi: [10.13328/j.cnki.jos.006487](https://doi.org/10.13328/j.cnki.jos.006487)]
- [5] 严毅, 邓超, 李琳, 朱凌坤, 叶彪. 深度学习背景下的图像语义分割方法综述. *中国图象图形学报*, 2023, 28(11): 3342–3362. [doi: [10.11834/jig.220292](https://doi.org/10.11834/jig.220292)]
- [6] 鄢杰斌, 方玉明, 刘学林, 姚怡茹, 睦相杰. 视频质量评价研究综述. *计算机学报*, 2023, 46(10): 2196–2224. [doi: [10.11897/SP.J.1016.2023.02196](https://doi.org/10.11897/SP.J.1016.2023.02196)]
- [20] 王志波, 王雪, 马菁菁, 秦湛, 任炬, 任奎. 面向计算机视觉系统的对抗样本攻击综述. *计算机学报*, 2023, 46(2): 436–468. [doi: [10.11897/SP.J.1016.2023.00436](https://doi.org/10.11897/SP.J.1016.2023.00436)]
- [23] 陈天宇, 周号益, 何铭睿, 仇尚航, 闫坤, 周萌萌, 李建欣. 共振攻击: 揭示跨模态模型 CLIP 的脆弱性. *计算机学报*, 2023, 46(12): 2597–2611. [doi: [10.11897/SP.J.1016.2023.02597](https://doi.org/10.11897/SP.J.1016.2023.02597)]
- [31] 苏亚鹏, 陈高曙, 赵彤. 不平衡样本下的 SA-YOLO 自适应损失目标检测算法. *中国科学院大学学报 (中英文)*, 2024, 41(3): 411–426. [doi: [10.7523/j.ucas.2023.013](https://doi.org/10.7523/j.ucas.2023.013)]
- [69] 张海艳. 基于深度学习的显著目标检测模型可扩展性研究 [硕士学位论文]. 南昌: 江西财经大学, 2022. [doi: [10.27175/d.cnki.gjxcu.2022.001183](https://doi.org/10.27175/d.cnki.gjxcu.2022.001183)]

作者简介

鄢杰斌, 博士, 副教授, CCF 专业会员, 主要研究领域为人工智能, 深度学习, 多媒体技术.

祝文涛, 硕士生, CCF 学生会员, 主要研究领域为深度学习, 模型评估.

刘学林, 博士, CCF 专业会员, 主要研究领域为视觉质量评估, 虚拟现实.

陈俊杰, 博士, 讲师, CCF 专业会员, 主要研究领域为计算机视觉, 姿态估计, 弱监督学习, 迁移学习.

钱峰, 博士, 讲师, CCF 专业会员, 主要研究领域为自动驾驶, 视觉感知, 路径规划.

方玉明, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为视觉大数据, 计算机视觉, 多媒体技术处理, 智能信息处理.