中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-16

论文引用格式:Chen Huachao, Chen Yingzhou, Wang Bin, Wang Feng, Zhao Jia. XXXX. Co-Salient Object Detection with Automatic Semantic Discovery. Journal of Image and Graphics, XX(XX):0001-0016(陈华超, 陈颖州, 王斌, 王峰, 赵佳. XXXX. 结合自动语义发现的协同显著性目标检测. 中国图象图形学报, XX(XX):0001-0016)[DOI:10.11834/jig.250377]

结合自动语义发现的协同显著性目标检测

陈华超,陈颖州,王斌,王峰,赵佳

阜阳师范大学　计算机与信息工程学院,安徽　阜阳　236037

摘　要: **目的**　协同显著性目标检测Co-SOD(Co-Salient Object Detection)旨在从一组相关图像中识别出既具有视觉显著性又具备语义一致性的目标区域。然而,现有方法在判别性建模方面存在局限,特别是在开放词汇场景下,语义泛化能力较弱,难以准确区分协同显著目标与复杂背景干扰。为此,本文提出一种新型框架以提升Co-SOD的语义理解能力和检测准确性。**方法**　本文设计了一种语义引导的开放词汇协同显著性目标检测框架。首先,利用CLIP(Contrastive Language - Image Pre-training)模型在组内图像与候选类别之间进行跨模态相似度建模,自动生成与组内语义一致的类别名称,从而减少人工指定类别的繁琐工作。随后,采用OWLv2(Open-Vocabulary Object Detector v2)模型将生成的类别名称与图像进行联合建模,输出与语义一致的候选目标框;再通过SAM(Segment Anything Model)对候选区域进行精细分割,提取高质量的显著性掩码。为增强语义表达的覆盖性与匹配稳定性,本文为每类标签构建多种语言描述,并采用多提示并行输入机制,以提升模型对语言歧义的鲁棒性与适应能力。针对复杂背景、遮挡和弱显著目标等挑战,本文还设计了一种动态阈值自适应策略,能够依据候选区域的语义一致性自适应调整检测门限,从而在保证检测准确性的同时有效提升召回率。**结果**　实验在CoCA、CoSal2015和CoSOD3k三个公开数据集上进行,并与当前主流的15种先进协同显著性检测方法进行了比较,在CoCA数据集中,相比于性能第2的模型,F值提高了25.7%,MAE降低了5.2%;在CoSOD3k数据集中,相比于性能第2的模型,F值提高了2.5%;在CoSal2015数据集中,相比于性能第2的模型,F值提高了2.2%,MAE降低了1.2%。此外,在CoSOD3k数据集上进行了消融实验,进一步验证了本文所提方法在提升协同显著性检测效果方面的有效性。**结论**　本文所提出的语义引导的开放词汇协同显著性目标检测框架,通过引入CLIP的组内一致类别生成方法,有效缓解了人工指定类别的局限性;结合OWLv2的开放词汇检测能力与SAM的通用分割优势,显著增强了模型的语义泛化能力和目标识别鲁棒性,为开放场景下的协同显著性检测提供了新思路与有效解决方案。

关键词: 协同显著性目标检测;语义引导;开放词汇;协同显著性目标检测框架;跨模态相似度;动态阈值自适应策略

Co-Salient Object Detection with Automatic Semantic Discovery

Chen Huachao, Chen Yingzhou, Wang Bin, Wang Feng, Zhao Jia

College of Computer and Information Engineering, Fuyang Normal University, Fuyang 236037, China

收稿日期:2025-08-08;**修回日期**:2025-11-19

*通信作者:赵佳　zhaojia11b@mails.ucas.ac.cn;赵佳,通信作者,男,教授,主要研究方向为计算机视觉。E-mail:zhaojia11b@mails.ucas.ac.cn

基金项目:国家自然科学基金(61906044)、安徽省自然科学基金(2408085MF154)、安徽省教育厅自然科学研究重点项目(2023AH050406, 2023AH050418)

Supported by: National Natural Science Foundation of China (61906044), Natural Science Foundation of Anhui Province (2408085MF154), Key Natural Science Research Projects of Anhui Provincial Department of Education (2023AH050406, 2023AH050418)

Abstract: **Objective** Co-SOD aims to identify target regions from a set of related images that are not only visually salient but also semantically consistent across the group. The core challenge of Co-SOD lies in not only detecting visually prominent objects within individual images but also determining which objects exhibit cross-image consistency among multiple images. Traditional Co-SOD methods mainly rely on low-level visual features to construct inter-image relationships. These approaches extract discriminative representations based on handcrafted features such as color, shape, and texture to mine common salient objects. While such methods perform reasonably well under simple backgrounds, they often fail in complex real-world scenarios with strong background clutter, severe occlusions, or multiple coexisting objects. In such cases, non-co-salient but visually salient objects can easily interfere with detection, leading to high false detection rates, poor robustness, and limited

Key words: Co-salient object detection; semantic guidance; open vocabulary; co-salient object detection framework; Cross-modal Similarity; dynamic threshold adaptation strategy

generalization. In recent years, with the rapid progress of deep neural networks, Co-SOD methods based on CNN (Convolutional Neural Network) or ViT (Vision Transformer) have been proposed. By learning high-level semantic features across images, these methods significantly improve detection performance. However, they typically still rely on modeling image groups as the basis of detection, without sufficiently incorporating explicit semantic priors. As a result, they lack the capability for proactive understanding of image categories or user-specific targets, and therefore struggle in open-domain multi-object scenarios or when fine-grained semantic distinctions are required. Moreover, when an image contains multiple visually salient but semantically irrelevant objects, these models often fail to distinguish them correctly, mistakenly including non-co-salient objects in the detection results, which reduces overall performance. For instance, in multi-class image datasets such as CoCA, a single image often contains multiple visually prominent objects, but only a subset of them are semantically related and constitute co-salient targets. Other salient regions, although visually striking, act as semantic distractors. Without explicit semantic filtering mechanisms, traditional methods are often unable to distinguish which objects are the true co-salient targets, thus decreasing the accuracy of Co-SOD. To address these limitations, this paper proposes a semantic-guided open-vocabulary co-salient object detection framework. The proposed method leverages language prompts to guide the model's attention toward target regions that are semantically aligned with image categories, thereby enabling semantic filtering of co-salient objects in complex multi-object scenarios and improving detection accuracy and semantic generalization ability. **Method** The proposed semantic-guided open-vocabulary Co-SOD framework integrates three key components: a CLIP-based group-consistent category generation method, the open-vocabulary object detection model OWLv2, and the universal segmentation model SAM, forming a complete pipeline that spans from semantic prompt generation to object detection and pixel-level segmentation. First, the system introduces a CLIP-based group-consistent category generation method to automatically derive the most representative group-consistent category for a given image set. This eliminates the need for manually specifying class names, reduces human effort, and provides accurate and stable semantic prompts for subsequent detection. Specifically, the method computes cross-modal similarities between the visual features of each image in the group and candidate category text features, and automatically generates the most representative category name that is consistent across the group, serving as reliable input for the detection model. Subsequently, the generated group-consistent category name is used as a natural language prompt and fed into the open-vocabulary object detection model OWLv2. OWLv2, pre-trained on large-scale image-text pairs, possesses strong cross-modal semantic alignment and matching capabilities. By mapping both image region features and text prompt features into a unified semantic space, OWLv2 can

identify target regions most relevant to the natural language prompts. To further improve the robustness of semantic understanding, this work constructs multiple semantically related natural language phrases for each category name and adopts a multi-prompt parallel input mechanism to feed them into OWLv2 simultaneously. This enhances the model's capacity to capture diverse semantic descriptions, improves stability and robustness in semantic matching, and effectively addresses uncertainties caused by language ambiguity. During the candidate region generation stage, similarities between region feature vectors and text semantic vectors are calculated to identify candidate regions most related to the input prompts, with the model outputting corresponding bounding boxes and confidence scores. To further improve detection flexibility and adaptability, a dynamic thresholding strategy is introduced. This mechanism automatically adjusts detection thresholds based on the confidence levels of semantic matches, thereby increasing the recall of low-confidence targets while maintaining precision. As a result, the framework is better able to detect occluded or weakly salient targets and avoid missed detections caused by overly strict thresholds. In the saliency segmentation stage, the bounding box regions filtered during the semantic detection stage are passed to the SAM. SAM then performs pixel-level mask generation for each region of interest. Leveraging its powerful structural perception and boundary modeling capabilities, SAM produces high-quality segmentation results. The final system outputs salient masks that are structurally complete and have clear boundaries. This design effectively decouples the two subtasks of semantic perception and visual detail modeling, allowing each module to fully leverage its strengths, thereby enhancing the overall semantic guidance ability and object delineation quality of the system. **Result** Experiments were conducted on three publicly available datasets: CoCA, CoSal2015, and CoSOD3k. The proposed method was compared with 15 state-of-the-art Co-SOD methods. On the CoCA dataset, our method achieved a 25.7% improvement in F-measure and a 5.2% reduction in MAE compared to the second-best model. On the CoSOD3k dataset, our method improved the F-measure by 2.5%. On the CoSal2015 dataset, it improved the F-measure by 2.2% and reduced the MAE by 1.2%. These results comprehensively validate the strong ability of the proposed framework to recognize semantically relevant co-salient objects in multi-object open scenarios. Furthermore, ablation experiments conducted on the CoSOD3k dataset further confirm the effectiveness of the proposed approach in enhancing co-salient object detection performance. **Conclusion** This paper proposes a semantic-guided open-vocabulary co-salient object detection framework. The framework first leverages a CLIP-based group-consistent category generation method to automatically generate semantic prompts for image groups, and then integrates OWLv2 for open-vocabulary object detection with SAM for universal segmentation. This enables explicit semantic filtering of co-salient targets as well as fine-grained pixel-level segmentation. The proposed framework significantly enhances semantic generalization and target recognition robustness, enabling effective localization of semantically consistent salient objects in multi-object and complex background scenarios. By unifying the three critical components—semantic prompt generation, open-vocabulary detection, and saliency segmentation—this framework provides an innovative and practical solution for advancing co-salient object detection in open environments.

0 引 言

显著性目标检测 SOD(Salient Object Detection)旨在从图像中识别视觉上最引人注目的目标。随着研究的不断深入,传统的 SOD 被扩展至多图协同理解场景,形成了协同显著性目标检测(Co-SOD)这一新兴任务。Co-SOD 的核心在于,在一组相关图像中发现既具有视觉显著性,又在语义上跨图像一致的目标区域。该任务在图像检索(Cheng 等,2014)、共同分割(Wang 等,2016)、语义分割(Zeng 等,2019)、视频分析(Jerripothula 等,2018)等多个计算机视觉

领域具有广泛应用。

传统 Co-SOD 方法主要依赖低层视觉特征构建图像间的关联性。通过手工特征学习颜色、形状和纹理等判别性表征，进而提取共享显著目标(Zhang 等，2015)。然而，当面对复杂背景、遮挡、多目标共存等实际场景时，这类方法的性能显著下降。其判别性和鲁棒性有限，容易受到干扰，难以准确识别协同显著目标。

随着深度学习技术的发展，基于卷积神经网络 CNN (Jiang 等，2021) 和视觉 Transformer (Zhou 等，2023) 的 Co-SOD 方法逐渐出现。这些方法通过学习高层语义特征，在一定程度上提升了检测性能。然而，它们仍存在判别性表征能力不足的问题，尤其在图像包含多个视觉显著但语义无关目标时，容易将非协同显著目标纳入检测结果。例如，如图 1 所示，当协同目标为“牛油果”时，传统方法通过拉近组内协同目标的表征距离、拉远与背景目标的表征距离来学习判别性特征。尽管该策略能够区分差异明显的对象，但对视觉相似且语义不同的目标(如“柠檬”“鸡蛋”“橙子”)常常难以做出正确判断。这暴露了现有方法在细粒度语义理解与判别性建模上的局限性，也凸显了引入开放词汇视觉语言模型和语义引导机制的必要性。

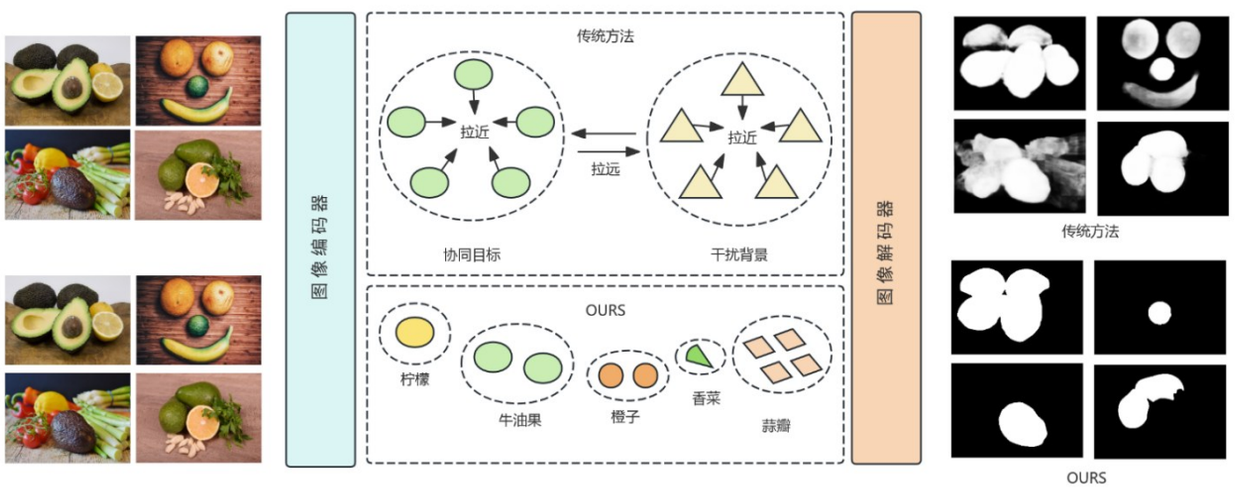


图 1　传统 Co-SOD 方法与本文提出方法的对比示意图

Fig. 1　Comparison diagram between traditional Co-SOD methods and the proposed method

视觉基础模型 SAM(Kirillov 等，2023)作为一种类别无关的分割方法，在多种分割任务中展现出强大的泛化能力。然而，SAM 依赖人工交互式提示(如点、框或掩码)进行分割，无法在无需人工干预的情况下自动识别图像中语义一致且视觉显著的目标区域，这限制了其在 Co-SOD 场景中的直接应用。

为解决上述问题，本文提出了一种融合自然语言理解与开放词汇检测能力的语义引导协同显著性目标检测框架。该框架结合开放词汇检测器 OWLv2(Minderer 等，2023)与自然语言提示，实现对语义相关候选目标区域的检测，并通过 SAM 模型对候选区域进行精细像素级分割，从而在多目标、多样性场景下有效提升语义一致性和分割精度。为保证检测阶段输入提示的准确性与一致性，本文进一步引入基于 CLIP 的组内一致类别生成方法。该方法自动生成最具代表性的类别名称，并作为 OWLv2 的自然语言输入，从而减少人工指定类别的繁琐工作，提升框架的自动化程度和鲁棒性。

本文的主要研究贡献如下：

1) 提出一种语义引导的开放词汇协同显著性目标检测框架，引入自然语言提示以增强模型对语义一致性目标的建模能力，从而提升在开放场景下的显著性目标检测与分割性能；

2) 设计一种基于 CLIP 的组内一致类别生成方法，可在无需人工标注的情况下自动生成语义类别名，显著增强了框架的自动化与鲁棒性；

3) 提出一种动态阈值自适应策略，能够根据候选区域的语义置信度自适应调整检测阈值，从而提

升候选目标筛选的准确性与召回能力；

4）设计一种多提示并行输入机制，支持在检测阶段同时输入多个语义相关的自然语言短语，提升模型对多样化语义表达的理解能力与鲁棒性。

1 相关工作

1.1 OWLv2模型

开放词汇目标检测近年来取得了显著进展，得益于对比训练的图文模型与经典目标检测器的结合（Gu 等，2021；Kamath 等，2021），使模型能够根据自然语言提示识别任意类别的语义目标。其中，Minderer 等人（2022）提出了 OWL-ViT（Vision Transformer for Open-World Localization）模型，通过结合简化的 ViT 架构与大规模图文对的对比预训练策略，实现了开放词汇目标检测。与传统方法相比，OWL-ViT 在架构上仅进行最小改动，便可将预训练的图像编码器与文本编码器直接应用于目标检测任务，避免了复杂的融合层或知识蒸馏技术。该模型通过将文本嵌入作为类别提示，与图像区域特征进行匹配，能够识别超出固定词汇表的新类别目标，适用于零样本（zero-shot）与单样本（one-shot）检测场景。

在实际应用方面，Alhadidi 等人（2024）将 OWL-ViT 应用于交通基础设施检测任务，用于识别交通标志、路面裂缝、井盖和杆状物等关键部件。在交通标志检测方面，其 F1 分数高达 88.61%，表明模型具备良好的语义泛化能力和现实应用潜力。Prajapati 等人（2024）则在交通工程场景中开展了零样本目标检测研究，尝试将 OWL-ViT 与大型多模态语言模型（如 Llava-Next 与 GPT-4o）结合，以提升在开放场景下的目标检测性能。

尽管 OWL-ViT 展现出较强的零样本检测能力，但在高置信度阈值条件下仍存在性能下降问题。为此，Minderer 等人（2023）提出了 OWLv2 框架，并引入自训练方法 OWL-ST（OWL Self-Training），利用现有检测器对大规模图文数据集（如 WebLI）生成伪边框标签，训练新模型，从而降低对人工标注数据的依赖，增强开放类别空间下的泛化能力。

为提高大规模训练效率，OWLv2 还引入了 token dropping（分词丢弃）与 mosaic 图像增强等机制，使模型能够在超过 20 亿图文对样本上进行有效自训练。实验表明，该策略显著提升了 OWLv2 在常见类别与未见类别目标上的泛化能力。在多个开放词汇检测基准（如 LVIS 和 ODinW）上，OWLv2 均取得优异成绩，进一步巩固了其在开放世界视觉理解中的基础地位。

在图像标注与细粒度目标识别任务中，Matuzevičius 等人（2024）评估了 OWL-ViT 与 OWLv2 在 CelebAMask-HQ 和 Face Synthetics 等数据集上的性能表现。结果显示，OWL-ViT 在高置信度阈值下精度与召回率急剧下降，而 OWLv2 受益于改进的训练策略和架构优化，依然保持良好的检测性能，尤其在检测小目标方面表现出色。例如，在 Face Synthetics 数据集中，OWLv2 实现了最高的 AP/AR 分数，展现出其在多样化视觉任务中的适应能力。

1.2 SAM 模型

SAM 是由 Meta AI 提出的通用分割基础模型，通过在 1100 万张图像和 10 亿个掩码上进行预训练，展现出强大的泛化能力和高质量的分割性能（Kirillov 等，2023）。SAM 支持点、框、掩码等多种提示输入，能够在无需类别先验的情况下完成高质量像素级分割，为通用分割任务提供了强有力的基础工具。

尽管 SAM 在一般图像分割任务中表现出色，但其在协同显著性目标检测（Co-SOD）中的直接应用仍存在挑战。首先，SAM 对用户交互提示（如前景点或边界框）依赖较高，而在自动化 Co-SOD 场景中，这类精确提示往往难以获取。其次，SAM 原生的掩码解码器结构较为简单，难以充分建模多图之间的语义一致性及多尺度显著特征，这在处理复杂背景或多目标图像时可能导致分割精度下降。

为解决这些问题，已有研究尝试将 SAM 应用于显著性检测任务。例如，Cui 等人（2023）提出 SSOM 模型，将 SAM 转化为单图显著性检测器，通过冻结图像编码器并引入 AdaLoRA 低秩微调策略，仅需训练 4M 参数即可实现高质量显著性分割；Xu 等人（2025）则提出对 SAM 进行多维探索，以适应弱监督的 SOD 任务；Ke 等人（2023）则在 SAM 的掩码解码器中引入可学习的高质量输出 token，以预测更精细的显著性掩码。

综上所述，SAM 在通用图像分割中具有强大的结构感知和边界建模能力，能够生成高质量的像素级掩码。然而，其在 Co-SOD 场景中的直接应用仍

受限于提示获取困难以及多图语义一致性建模不足。这一局限性为后续引入语义引导机制提供了明确的研究动机，从而提升协同显著性目标检测的语义一致性与分割精度。

1.3 语义引导的检测方法

近年来，语义引导的检测方法逐渐成为开放词汇目标检测与显著性分割领域的重要发展方向。相比传统依赖固定类别词汇表的检测器，语义引导方法通过引入跨模态表征和语言提示，将类别语义信息直接融入检测与分割过程，从而显著提升模型在开放场景下的泛化能力与鲁棒性。这类方法的核心在于利用视觉-语言模型（如 CLIP）或开放词汇检测器（如 OWLv2）提供的语义嵌入，将视觉特征与文本特征对齐，使检测器能够识别超出训练词汇表的目标类别。

已有研究表明，语义引导能够在复杂场景下显著提升目标检测与分割性能。例如，Zhou 等人（2024）提出的 SOGDet 框架通过引入 3D 语义占据预测分支，显式建模场景的物理上下文信息（如道路、植被等），并利用模态融合机制将语义信息嵌入目标检测特征中，实现端到端优化，有效增强模型对结构化环境的感知能力。Wang 等人（2024）提出的语义引导少样本检测方法通过三个核心操作提升稀缺样本条件下的泛化能力：首先，利用 CLIP 生成的语义嵌入构建相似性分类器（SSC），以余弦相似度替代传统线性分类决策；其次，采用多模态特征融合模块（MFF）实现视觉与语言特征的深层交互；最后，引入语义感知间隔损失，依据类别间语义相似性自适应施加边界约束，从而缓解少样本场景下的类间混淆问题。

此外，一些研究尝试将语义引导机制与通用分割模型结合，以提升语义一致性和分割精度。Ding 等人（2024）利用 CLIP 提供跨模态语义提示，引导分割模型在复杂背景中定位语义相关区域；Liu 等人（2024）基于 Grounding DINO 生成语义相关检测框，并结合 SAM 完成精细分割，实现语义一致的显著性区域提取。这类方案启发了本论文的设计思路：首先通过基于 CLIP 的组内一致类别生成方法，为 OWLv2 提供准确的语义提示；随后，OWLv2 利用这些语义提示生成语义相关的候选目标区域，并结合 SAM 提取高质量的显著性掩码，从而有效提升检测结果的语义一致性与分割精度。

2 方　法

2.1 总体框架

在开放域多目标图像中，协同显著性目标检测（Co-SOD）面临诸多挑战，尤其是在 CoCA 等多类别图像数据集中，一张图像通常包含多个视觉上显著的物体，而其中仅有部分与图像的真实类别语义密切相关。其余显著区域尽管在视觉上突出，但在语义层面上却可能构成干扰，从而降低协同显著性检测的准确性，并影响后续分割、识别等任务的鲁棒性与有效性。

为解决上述问题，本文提出一种语义引导的开放词汇协同显著性目标检测框架，旨在通过语言提示引导模型关注与图像所属类别语义一致的区域，实现复杂图像中协同显著目标的语义筛选。该框架首先通过 CLIP 模型对图像组内的视觉特征进行语义聚合，生成组内一致的类别候选，为每组图像确定最可能的语义类别。随后，将该类别作为提示输入，结合多模态目标检测模型（OWLv2）生成语义相关的候选显著性区域。最终，引入 SAM 对候选区域进行像素级分割，生成高质量显著性掩码，实现目标的精确定位与语义一致性。

整体流程如图 2 所示。

综上，该框架实现了在不依赖组间协同关系的前提下，对语义相关显著性目标的自动筛选与精确定位，有效抑制了语义无关区域的干扰，为后续的分割和识别任务提供了更可靠的输入基础。

2.2 基于 CLIP 的组内一致类别生成方法

在协同显著性数据集中，一组图像往往存在语义相关的公共显著性目标。为在图像组内自动确定与真实类别语义一致的主要目标类别，本文引入基于 CLIP 模型的组内一致类别生成方法。该方法通过对图像组中的每张图像提取视觉特征，并与候选类别文本特征进行相似度匹配，从而选取组内最可能的类别。

具体而言，设一组图像为 $\{\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_N\}$，候选类别集合为 $\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_M\}$。首先，使用 CLIP 的图像编码器对每张图像进行特征提取，得到图像特征向量 $u_i\in\mathbf{R}^d$，并对其进行归一化处理，可表示为如公式（1）所示：

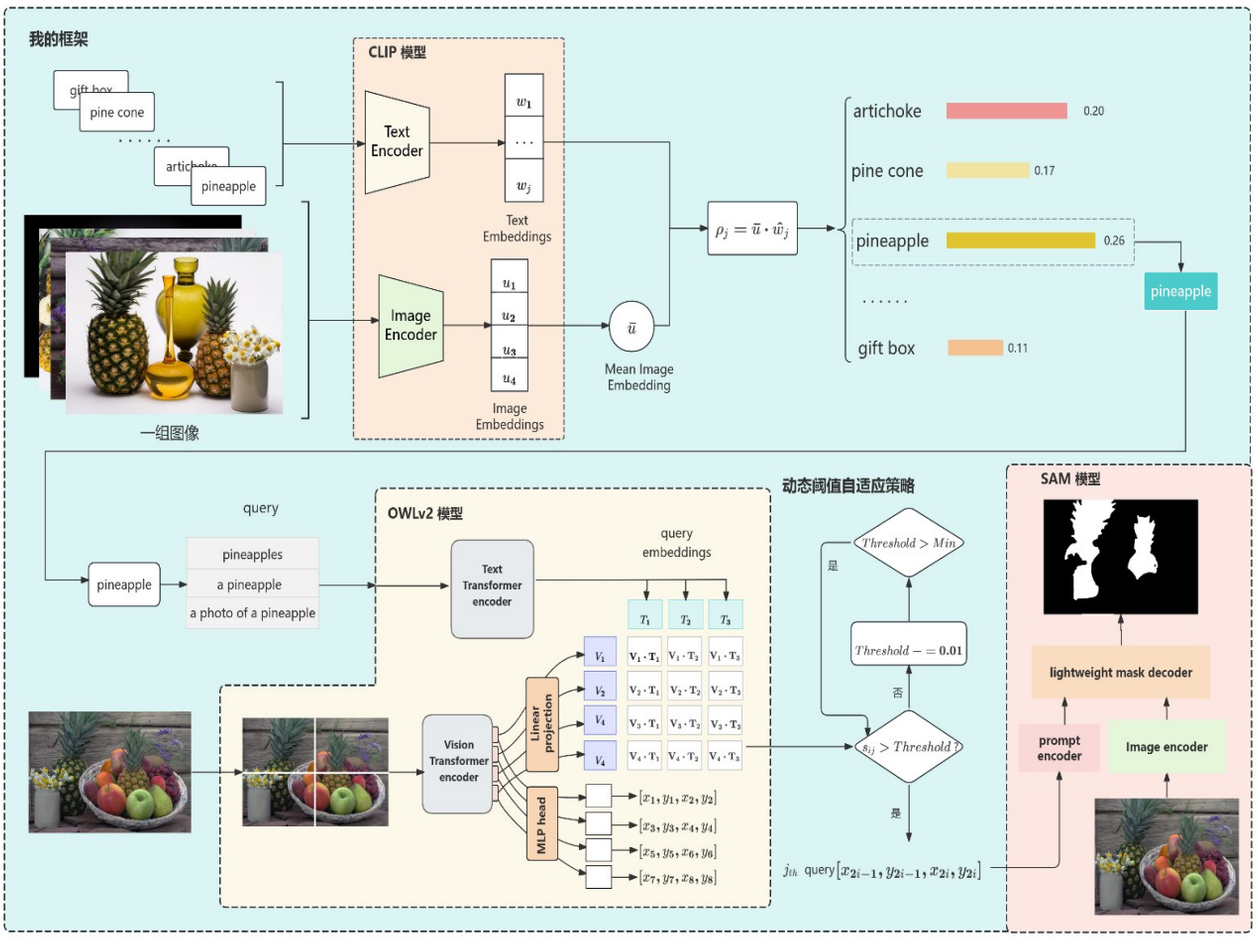


图2　整体流程图

Fig. 2　Overall Framework Diagram

$$\hat{u}_i = \frac{u_i}{\|u_i\|_2}, i = 1,2,\cdots,N \tag{1}$$

同时，将候选类别文本 C_j 输入 CLIP 文本编码器，得到文本特征向量 $w_j \in \mathbf{R}^d$，并归一化，可表示为如公式(2)所示：

$$\hat{w}_j = \frac{w_j}{\|w_j\|_2}, j = 1,2,\cdots,M \tag{2}$$

随后，计算图像组内平均图像特征向量 $\bar{u}$，可表示为如公式(3)所示：

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{N}\hat{u}_i \tag{3}$$

通过计算 $\bar{u}$ 与每个候选类别文本向量 w_j 的余弦相似度，评估类别与图像组的匹配程度，可表示为如公式(4)所示：

$$\rho_j = \bar{u} \cdot \hat{w}_j, j = 1,2,\cdots,M \tag{4}$$

式中，ρ_j 为图像组特征向量与候选类别 C_j 的相似度得分。组内最一致的类别 $\hat{C}$ 由最大相似度得分决定，可表示为如公式(5)所示：

$$\hat{C} = \arg\max_j \rho_j \tag{5}$$

为进一步分析潜在候选类别，本文还保留相似度最高的 Top-K 类别，用于后续语义引导或辅助模型优化。该方法的核心优势在于利用 CLIP 的跨模态对齐能力，无需提前标注单张图像的类别，即可自动推断图像组的主要语义类别，且与数据集中真实类别具有较高的一致性。

通过对图像组平均特征向量的方式，可有效降低个体图像噪声的影响，从而提高组内类别预测的稳定性和鲁棒性。该方法可作为语义引导模块的前置步骤，为后续的目标检测、像素级分割和显著性分析提供可靠的类别提示。

2.3　OWLv2：基于语言提示的开放词汇目标检测

OWL-ViT 是一种支持开放词汇目标检测的视觉语言模型(Minderer 等，2022)。OWL-ViT 基于

Vision Transformer 架构，结合了类 CLIP 模型的跨模态语义对齐机制，能够根据任意自然语言提示检测图像中与之语义相关的目标区域。

与传统目标检测方法依赖封闭类别集合的方式不同，OWL-ViT 通过计算图像区域与文本提示之间的语义相似度来完成目标检测，具备更强的开放性与泛化能力，特别适用于多目标混杂和语义干扰显著的复杂图像场景。

OWL-ViT的核心思想是通过跨模态对齐，实现图像区域与文本语义之间的直接匹配。为提升协同显著性目标检测中的语义泛化能力，本研究引入其增强版本 OWLv2作为语义候选区域检测模块，借助其开放词汇检测能力，实现对图像中与类别语义一致目标区域的精确筛选。具体而言，输入图像与类别文本提示 p 共同输入 OWLv2模型。模型首先将输入图像 I 划分为若干视觉区域，并将每个区域编码为特征向量 $V_i \in \mathbf{R}^d$，同时将每条类别提示文本编码为语义向量 $T_j \in \mathbf{R}^d$。随后，模型通过计算区域与文本之间的余弦相似度 s_{ij} 来度量其语义匹配程度，可表示为如公式(6)所示：

$$s_{ij} = \frac{V_i \cdot T_j}{\|V_i\| \cdot \|T_j\|} \tag{6}$$

为确定每个图像区域的预测类别，OWLv2 在所有文本提示中选取与其相似度最高者，可表示为如公式(7)所示：

$$\hat{j} = \arg\max_j s_{ij},\ Score_i = \max_j s_{ij} \tag{7}$$

式中 $\hat{j}$ 表示与区域 v_i 最匹配的文本提示索引，$Score_i$ 表示其对应的最大相似度得分。模型将得分超过设定阈值的区域视为检测目标，并生成对应的边界框与预测标签。该机制赋予 OWLv2 强大的开放词汇检测能力，使其能够根据任意语言提示灵活识别图像中的相关目标区域。

其中，OWLv2 的 Linear Projection 层用于将视觉 patch 特征映射到与文本对齐的对比空间，MLP Head 则对每个匹配到的 region token 独立回归检测框及其匹配概率，形成先特征对齐、后区域定位的完整流程，从而保证了区域的语义一致性与检测精度。

例如，给定文本提示 "pineapple"，模型并不会先识别每个框是什么“名字”，而是直接判断哪个框的语义向量最接近这个短语，所以它不需要事先知道“这张图里有无菠萝”，也能找到匹配目标。

当候选区域的相似度得分超过预设阈值，即视为与类别语义相关的目标区域，并输出对应的边界框(bounding box)及其匹配置信度。对于输入的一条文本提示，若图像中仅存在一个语义一致的目标，OWLv2 通过 one-to-one 匹配机制将预测并输出唯一的候选检测框，有效避免同一目标的重复检测。为增强模型对弱目标或低对比区域的适应能力，本文引入动态阈值自适应策略，提升检测的鲁棒性与覆盖率。

此外，为增强语义提示对检测结果的引导能力，本文在每类标签基础上扩展构造多个语义相近的文本表达(如 "a photo of a pineapple"、" pineapples"、"a partially occluded pineapple"、等)，联合输入模型，提升对多样化语言描述的适应性。

最终，所有匹配置信度超过自适应阈值的候选区域将作为目标框输出，其边界框坐标采用标准形式 $[x_1, y_1, x_2, y_2]$，并传递至后续的 SAM 模块用于像素级显著性目标分割。

2.4 SAM：目标分割与掩码生成

为了实现对候选目标区域的精细化分割，本文引入 SAM 作为掩码生成模块，对 OWLv2 输出的候选区域进行像素级处理，生成边界准确、结构清晰的显著性掩码图。SAM 是一类基于提示驱动(prompt-driven)的通用分割模型，具备强大的零样本分割能力，能够在多类图像任务中展现出良好的泛化性能，尤其适用于开放域、多目标及复杂背景场景下的目标提取。

在本方法中，SAM 接收原始图像及其对应的候选边界框(bounding boxes)作为空间提示输入，首先提取图像的全局视觉特征，并结合每个候选框，引导模型在该区域内进行像素级掩码生成。对于一张输入图像，SAM 会针对每个候选框返回一个或多个独立的分割掩码，每个掩码均附带一个预测 IoU(Predicted Intersection over Union)得分，用以衡量该掩码与潜在真实边界之间的匹配程度。输出结果包括一组候选掩码及其对应的预测 IoU 分数，为后续掩码筛选与整合提供依据。

为确保融合掩码的可靠性，本文引入基于预测 IoU 的掩码筛选机制：具体而言，仅保留预测 IoU 值高于设定阈值且具有明确结构的掩码图用于最终合并，构建整张图像的联合掩码(combined mask)。对于预测 IoU 低于阈值或结构不完整的掩码区域，则

视为该区域分割结果不可信，将其排除在最终掩码融合过程之外，以避免低质量掩码干扰整体检测效果。

此外，为进一步提升掩码的整洁度与边界质量，本文对保留的掩码图引入后处理策略，主要包括形态学闭运算与开运算，以填补内部空洞、平滑边缘结构；同时采用连通域分析滤除小面积孤立区域及冗余碎片，从而保证分割结果的连通性与边缘平滑性。

综上所述，SAM 模块在本框架中完成了从粗粒度目标框到高质量显著性掩码的转换，并通过 IoU 筛选与后处理机制，进一步提升了分割结果的准确性与语义一致性，为复杂图像中的目标分割与分析提供了可靠支撑。

2.5 动态阈值自适应策略

在开放词汇目标检测任务中，固定置信度阈值往往难以在检测精度与召回率之间取得理想平衡。尤其是在处理弱显著性目标、边缘区域目标或被部分遮挡的物体时，过高的置信度阈值可能导致语义相关目标漏检，进而影响整体检测性能。为缓解此类问题，本文设计并引入了一种动态阈值自适应策略，通过逐步调整置信度阈值，实现对潜在目标区域的更全面覆盖。

具体而言，系统从预设的初始置信度阈值（如 $T_{init}=0.4$）开始执行语义匹配检测。为便于表述，置信度阈值 *Threshold* 在后文中简化记为 T。

若当前轮次未检测到任何候选目标框（$N_k=0$），则自动将阈值按设定步长（如 $\Delta T=0.01$）递减，并重新执行检测过程。该迭代过程持续进行，直到成功检出候选框或阈值下降至设定下限（如 $T_{min}=0.001$）为止。

为便于形式化表示，阈值更新过程可表示为如公式(8)所示：

$$T^{(k+1)}=\begin{cases}T^{(k)}-\Delta T, if\ N^{(k)}=0\ and\ T^{(k)}>T_{min}\\ T^{(k)}, T^{(0)}=T_{init}, otherwise\end{cases} \tag{8}$$

式中，$T^{(k)}$ 表示第 k 轮检测的置信度阈值，N_k 表示在第 k 轮检测中成功检测到的候选目标框数量。

该策略的核心优势体现在以下两个方面：

• 高阈值阶段：优先保留语义匹配度高的候选区域，输出结果具有更高的准确性与语义一致性，适用于显著性强或目标清晰的场景；

• 低阈值阶段：逐步放宽匹配条件，引入更多潜在但置信度较低的候选区域，显著提升整体的目标召回率，特别适用于复杂背景或弱语义目标的补全。

最终，所有通过动态阈值筛选的候选框将作为提示输入传递至SAM，用于获取各区域的高质量像素级分割掩码。该机制确保了在复杂多目标场景中，尤其在存在显著性干扰或类别间语义模糊的情况下，依然能够实现语义引导下的目标全面提取，显著提升了系统的稳健性与适应性。

3 实　验

3.1 实验设计

本研究提出的显著性目标检测方法基于 OWLv2-large-patch14-ensemble 模型，在 NVIDIA RTX A6000 GPU 上完成实验评估。该模型具备开放词汇目标检测能力，能够根据自然语言短语提示在图像中检测与语义相关的目标区域。

在模型输入方面，本文结合 CLIP 的跨模态相似度计算机制，从组内多张图像中自动生成语义一致的类别名称，并将其转化为符合 OWLv2 提示格式的多样化自然语言短语，作为模型文本输入以实现语义对齐与目标检测。

在实验评估中，本文选用 CoSal2015（Zhang 等，2016）、CoCA（Zhang 等，2020）和 CoSOD3k（Fan 等，2020）三个具备语义一致性特征的协同显著性目标检测数据集作为测试集，以更好地支持基于提示词的开放词汇检测评估。推理阶段，模型输入为一张图像及其对应的语义提示词，通过跨模态对齐机制匹配提示词与图像区域，输出为图像中语义相关目标的检测框。后续可进一步结合分割模块（如 SAM）对检测框区域进行像素级掩码提取，实现显著性目标的高质量分割。

为保证评估的公平性，本文在统一评测代码框架下，与多种主流协同显著性检测方法进行对比，包括 DeepACG（Co-Saliency Detection via Semantic-Aware Contrast Gromov-Wasserstein Distance）（Zhang 等，2021）、DMT（Discriminative Co-Saliency and Background Mining Transformer）（Li 等，2023）、DCFM（Democratic Comprehensive Feature Mining）（Yu 等，2022）、GCoNet+（Group Collaborative Co-Salient Object Detector +）（Zheng 等，2023）、HrSSMN（Hierarchical high-order Spatial-Semantic Modulation Network）（Zhang 等，2022）、GCoNet（Group Collaborative

Network)(Fan 等,2021)、GICD(Gradient-Induced Co-Saliency Detection)(Zhang 等,2020)、CoGEM(Group Exchange-Masking)(Wu 等,2023)、GCAGC(Graph Convolutional network with Attention Graph Clustering)(Zhang 等,2020)、CADC(Consensus-Aware Dynamic Convolution)(Zhang 等,2021)、UFO(Unified Framework for Co-Object Segmentation)(Su 等,2023)、ICSM(Information Calibration and Sliding Mining)(Wei 等,2024)、UniTR(A Unified TRansformer-Based Framework for Co-Object and Multi-Modal Saliency Detection)(Guo 等,2024)以及 ASCoD(Appearance and Shape Co-Representations for Co-Saliency Detection)(Guo 等,2025)等。评估指标包括平均绝对误差(*MAE*)(Wang 等,2019)、最大 F 值(F_β^{max})(Achanta 等,2009)、最大 E 值(E_ϕ^{max})(Fan 等,2018)和 S_α(Fan 等,2017),其中 *MAE* 越低表示误差越小,其余指标数值越高代表检测性能越优。

3.2 语义一致性验证

在本研究的方法框架中,CLIP 被用于在图像组内自动生成语义一致的类别提示,从而为后续的开放词汇检测与分割提供有效的语义引导。为验证该模块的合理性与有效性,本文在 CoCA、CoSal2015 和 CoSOD3k 三个数据集上,对 CLIP 自动生成的类别名称与数据集真实类别进行了对比分析,并统计了语义一致性准确率。结果如表 1 所示。

表 1 语义一致性验证实验

Table 1 Semantic Consistency Verification Experiment

数据集	组数	准确率
CoCA	80	100.0%
Cosal2015	50	100.0%
CoSOD3k	160	97.5%

从表中可以看出,CLIP 自动生成的类别名称在绝大多数情况下能够与数据集真实类别保持高度一致。即便出现少量不一致的情况,也多为语义相近的替换(如 sheep 与 antelope),不会影响语义对齐和后续检测任务的效果。这一结果表明,基于 CLIP 的组内一致类别生成方法能够在无需额外标注的情况下,有效发现图像组的公共语义目标,为后续的开放词汇检测和分割提供了可靠的类别提示。

综上所述,该实验进一步验证了本文所提方法在自动语义生成方面的可行性和有效性,并为整体框架在开放场景下的语义引导奠定了基础。

3.3 实验结果

表 2 展示了本文所提出方法与多种最新主流方法的定量对比结果。得益于所设计的语义引导检测与分割模块及动态阈值自适应策略,本文方法能够更充分地利用自然语言提示信息,实现对图像中非显著性区域的有效抑制,并进一步提升对显著性目标区域的分割精度。当阈值初设为 0.4 时,所提方法在多个真实场景数据集上均取得了领先的检测性能。

具体来看,在 CoCA 数据集上,本文方法分别在 E_ϕ^{max}、S_α、F_β^{max}、*MAE* 指标上取得了 0.948、0.907、0.897 和 0.025 的优异结果;在 CoSOD3k 数据集上,取得了 0.915、0.880、0.884 和 0.046 的优异结果;在 CoSal2015 数据集上,同样取得了 0.952、0.919、0.938 和 0.033 的优异结果。

相较于性能第二的模型,本文方法在 CoCA 数据集上,E_ϕ^{max}、S_α、F_β^{max} 指标分别提升了 13.1%、16.9% 和 25.7%,*MAE* 指标相比现有方法下降了 5.2%,充分验证了方法在开放词汇场景下的显著性检测能力。在 CoSal2015 数据集上,方法在 E_ϕ^{max}、S_α、F_β^{max} 指标上分别有 1%、1.3%、2.2% 的提升,*MAE* 降低 1.2%,进一步验证了其在协同显著性检测场景下的有效性。在 CoSOD3k 数据集上,方法同样取得了较好的性能表现,进一步证明了所提出各模块在大规模复杂场景下的适用性与鲁棒性。

此外,如图 3 所示,本文进一步给出了与多种最新方法的定性可视化对比结果。从可视化结果可以看出,本文方法在复杂背景下能够更好地抑制无关区域的干扰,显著提升了显著性目标的边缘细节刻画效果。同时,对于多尺度目标,所提方法也能更好地捕获语义一致性与区域完整性,明显优于现有主流方法的分割表现。

综上所述,本文方法在 CoCA、CoSOD3k 和 CoSal2015 等多个标准测试集上的检测结果均优于对比方法,证明了所提语义引导显著性检测框架在开放场景下具备较强的泛化能力与分割精度。

3.4 消融实验

为验证所提出各核心模块对整体检测与分割性能的贡献,本文在规模较大的 CoSOD3k 测试集上进行了系统的消融实验。实验主要从模块组合与阈值

表2　与其他先进方法在三个基准数据集上的对比

Table 2　Comparison with other state-of-the-art methods on three benchmark datasets

方法	CoCA				CoSOD3k				CoSal2015			
	$E_\phi^{max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{max}\uparrow$	$MAE\downarrow$	$E_\phi^{max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{max}\uparrow$	$MAE\downarrow$	$E_\phi^{max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{max}\uparrow$	$MAE\downarrow$
GCAGC	0.754	0.669	0.523	0.111	0.816	0.785	0.740	0.100	0.866	0.817	0.813	0.085
GICD	0.701	0.658	0.504	0.125	0.831	0.794	0.743	0.089	0.884	0.842	0.834	0.072
DeepACG	0.771	0.688	0.552	0.102	0.838	0.792	0.756	0.089	0.892	0.854	0.842	0.064
CADC	0.744	0.681	0.548	0.132	0.840	0.801	0.859	0.096	0.906	0.866	0.862	0.064
GCoNet	0.760	0.673	0.544	0.105	0.860	0.802	0.777	0.071	0.887	0.845	0.847	0.068
HrSSMN	0.739	0.671	0.532	0.106	0.842	0.788	0.753	0.087	0.895	0.845	0.841	0.062
DCFM	0.783	0.710	0.598	0.085	0.874	0.810	0.805	0.067	0.892	0.838	0.856	0.067
UFO	0.782	0.697	0.571	0.095	0.874	0.819	0.797	0.073	0.906	0.860	0.865	0.064
GCoNet+	0.814	0.738	0.637	0.081	0.901	0.843	0.834	0.062	0.924	0.881	0.891	0.056
DMT	0.800	0.725	0.619	0.108	0.895	0.851	0.835	0.063	0.936	0.897	0.905	0.045
CoGEM	0.808	0.726	0.599	0.095	0.911	0.853	0.829	0.061	0.933	0.885	0.882	0.053
CLIP-COD	0.817	0.735	0.635	0.077	**0.916**	0.858	0.838	0.053	0.939	0.906	0.916	0.049
ICSM	0.806	0.725	0.626	0.099	0.890	0.858	0.843	0.061	0.942	0.899	0.896	0.052
UniTR	0.782	0.702	0.567	0.096	0.879	0.827	0.802	0.072	0.906	0.866	0.862	0.064
ASCoD	0.816	0.738	0.640	0.083	0.901	0.830	0.819	0.063	0.907	0.861	0.867	0.061
OURS	**0.948**	**0.907**	**0.897**	**0.025**	0.915	**0.880**	**0.884**	**0.046**	**0.952**	**0.919**	**0.938**	**0.033**

注：加粗字体表示各列最优结果

策略对比两个方面展开，具体结果如表3和表4所示。

首先，在模块组合实验表3中，我们分别比较了仅使用SAM分割、在此基础上引入CLIP自动类别生成与OWLv2语义候选区域检测、以及进一步结合动态阈值自适应策略的完整方案。结果表明，单独使用SAM时性能较低，难以保证分割区域的语义一致性。当引入CLIP与OWLv2后，模型能够有效定位语义相关的候选区域，并在像素级分割中显著提升边界完整性和语义对齐效果。在此基础上，进一步结合动态阈值自适应策略后，模型在CoSOD3k上取得了最优性能，E_ϕ^{max}、S_α、F_β^{max}分别达到0.915、0.880、0.884，*MAE*降至0.046，充分验证了所提出各模块的有效性与互补性。

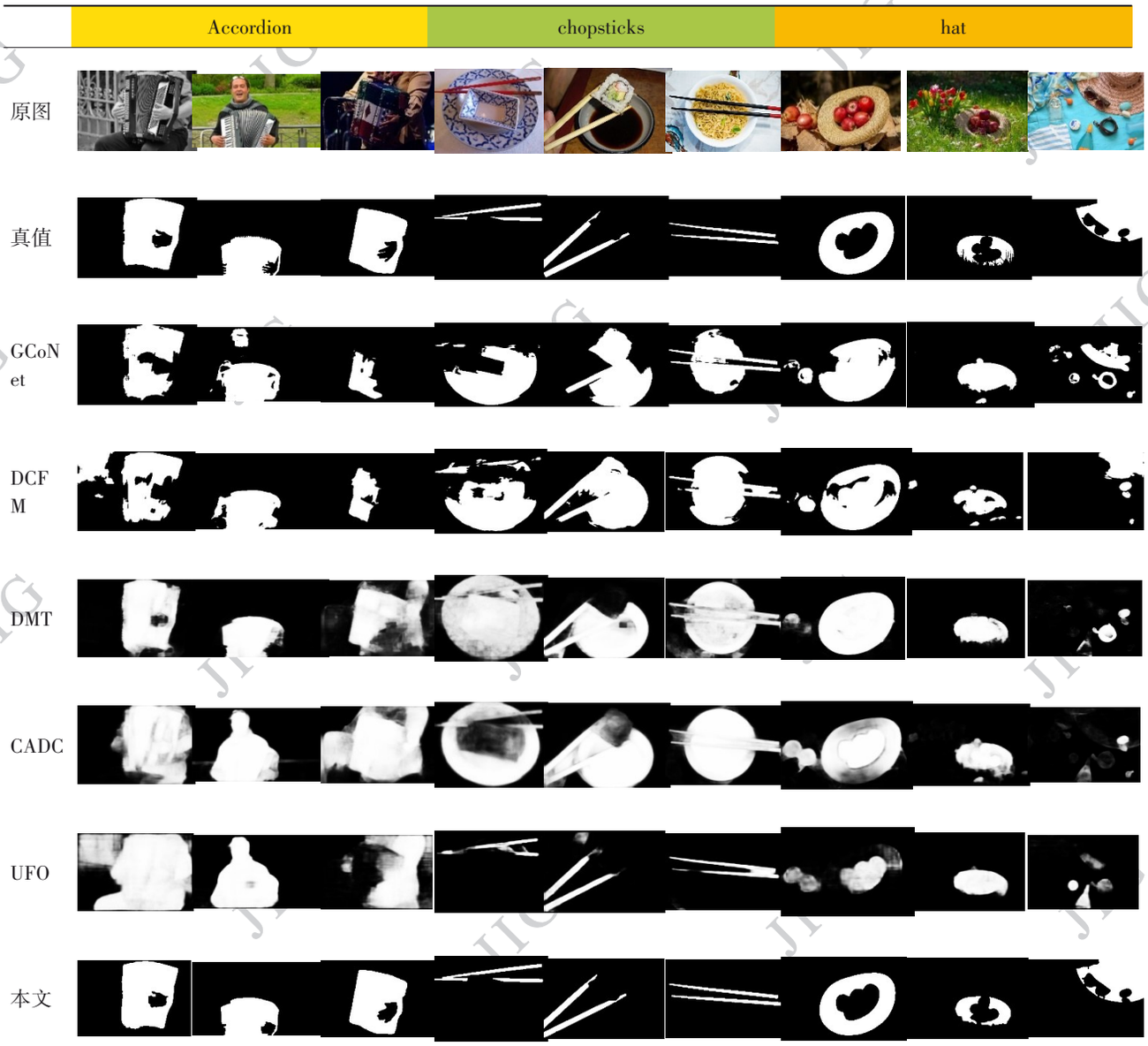
其次，在阈值策略对比实验表4中，我们系统评估了固定阈值（*T*=0.1－0.8）与动态阈值自适应策略在CoSOD3k数据集上的性能差异。结果显示，当采用固定阈值时，模型的整体性能对阈值的选择高度敏感：当阈值较低时，虽然能够覆盖更多潜在目标区域，但会引入大量噪声，导致分割边界模糊、背景区域误检率上升，从而造成性能明显下降；而当阈值较高时，虽然可以减少误检，但弱显著性目标和遮挡区域中的真实目标往往被漏检，整体性能同样下降。这一现象充分说明，固定阈值策略难以在不同场景下同时兼顾目标召回率与精确性。

相比之下，动态阈值自适应策略能够根据候选区域的置信度分布自适应调整筛选标准，有效提升弱显著性目标和复杂遮挡场景下的检测与分割性能。最终结果表明，动态阈值在所有指标上均显著优于固定阈值方案，进一步验证了其在开放词汇协同显著性检测中的优势与鲁棒性。

综上所述，消融实验结果表明，自动语义生成、语义引导检测、动态阈值调节与提示驱动分割在开放词汇显著性检测任务中相互协同，可在保证检测准确率的同时，有效提升弱目标和复杂场景下的目标召回率与掩码分割质量。

3.5　协同对象的可选择性

Fig. 3　Visual Comparison of Our Method and State-of-the-Art Methods on Test Images



注:图3我们的方法和其他最先进的测试图像方法的视觉比较

表3　CoSOD3k的模块组合消融实验

Table 3　Ablation Study of Module Combinations on CoSOD3k

组合方式				CoSOD3k			
CLIP	SAM	OWLv2	阈值	E_ϕ^{max} ↑	S_α ↑	F_β^{max} ↑	*MAE* ↓
	√			0.612	0.564	0.498	0.281
√	√	√	固定	0.870	0.819	0.809	0.093
√	√	√	动态	0.915	0.880	0.884	0.046

如图4所示,图像组中有多个共同目标(如棒球,棒球棒,人等),在输入图像中,我们可以通过修改文本提示来选择我们想要的分割对象。以前的方法都不能选择期望的分割目标,但这在实际应用中是非常重要的。

表4　CoSOD3k的阈值策略对比实验

Table 4　Threshold Strategy Comparison on CoSOD3k

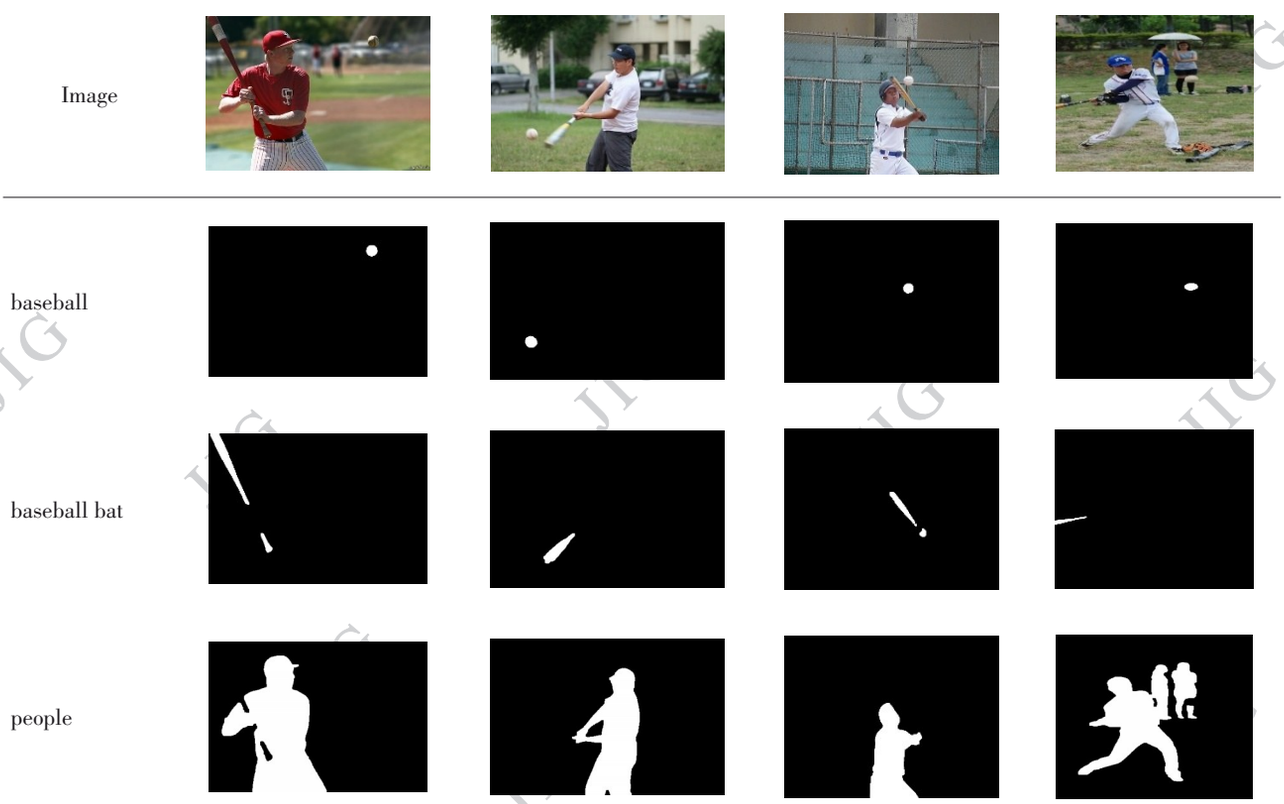
阈值	CoSOD3k			
	$E_\phi^{max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{max}\uparrow$	$MAE\downarrow$
0.1	0.778	0.749	0.727	0.152
0.2	0.870	0.830	0.818	0.086
0.3	0.884	0.839	0.832	0.081
0.4	0.870	0.819	0.809	0.093
0.5	0.831	0.775	0.753	0.118
0.6	0.756	0.685	0.634	0.173
0.7	0.654	0.565	0.476	0.250
0.8	0.546	0.434	0.302	0.332
动态	0.915	0.880	0.884	0.046

4　结　论

针对Co-SOD中判别性表征能力受限及在开放词汇场景下语义泛化能力不足的问题，本文提出了一种语义引导的开放词汇协同显著性目标检测框架。该方法首先引入基于CLIP的组内一致类别生成机制，通过跨模态相似度计算自动生成与组内语义一致的类别名称。在此基础上，充分利用OWLv2的开放词汇目标检测能力，生成与类别语义一致的多目标检测框，并结合SAM强大的分割能力提取高质量掩码，从而在图像中有效保留语义一致的显著性目标，显著提升了协同目标的可分性与检测精度。同时，设计了一种基于语义置信度的动态阈值调节策略，能够根据候选区域的语义一致性自适应调整检测阈值，有效提升候选目标筛选的准确性与召回能力。

实验结果表明，本文方法在多个Co-SOD基准数据集上均优于GCoNet、DMT、CoGEM等现有方法，能够更好地区分协同显著目标与干扰背景。未来工作将进一步探索跨模态表征在协同显著性目标理解中的潜力，优化语义提示生成策略，提升推理效率等挑战性的任务中。

Fig. 4　Segmentation Results Guided by Textual Prompts



注：图4根据提示分割相应的目标图示

参考文献

Achanta R, Hemami S, Estrada F and Susstrunk S. 2009. Frequency-Tuned Salient Region Detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE: 1597-1604[DOI: 10.1109/CVPR.2009.5206596]

Alhadidi T, Jaber A, Jaradat S, Ashqar H and Elhenawy M. 2024. Object Detection Using Oriented Window Learning Vision Transformer: Roadway Assets Recognition//International Conference on Intelligent Systems, Blockchain, and Communication Technologies. Cham: Springer Nature Switzerland: 506-522[DOI: 10.1007/978-3-031-82377-0_41]

Cheng M M, Mitra N J, Huang X and Hu S M. 2014. Salientshape: group saliency in image collections. The Visual Computer, 30(4): 443 - 453 [DOI: 10.1007/s00371-013-0867-4]

Cui R, He S and Qiu S. 2023. Adaptive low rank adaptation of segment anything to salient object detection[EB/OL]. [2025-08-08]. https://arxiv.org/pdf/2308.05426.pdf

Ding C, Wu Y, Song H, Zhang K, Zhang X and Guo Z. 2024. Language-Guided Semantic Alignment for Co-Saliency Detection//Proceedings of the IEEE International Conference on Multimedia and Expo. Singapore: IEEE: 1-6[DOI:10.1109/ICME57554.2024.10687964]

Fan D P, Cheng M M, Liu Y, Li T and Borji A. 2017. Structure-Measure: A New Way to Evaluate Foreground Maps//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4548-4557[DOI:10.48550/arXiv.1708.00786]

Fan D P, Gong C, Cao Y, Ren B, Cheng M M and Borji A. 2018. Enhanced-alignment measure for binary foreground map evaluation [EB/OL]. [2025-08-08]. https://arxiv.org/pdf/1805.10421.pdf

Fan DP, Lin Z, Ji GP, Zhang D, Fu H and Cheng MM. 2020. Taking a Deeper Look at Co-Salient Object Detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 2919-2929 [DOI: 10.1109/CVPR42600.2020.00299]

Fan Q, Fan DP, Fu H, Tang CK, Shao L and Tai YW. 2021. Group Collaborative Learning for Co-Salient Object Detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 12288-12298[DOI: 10.1109/CVPR46437.2021.01211]

Gu X, Lin T Y, Kuo W and Cui Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation[EB/OL]. [2025-08-8]. https://arxiv.org/pdf/2104.13921.pdf

Guo G T, Hu S L, Song H H and Zhang K H. 2025. Learning Joint Appearance and Shape Co-Representations for Co-Saliency Detection// ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: IEEE: 1-5[DOI:10.1109/ICASSP49660.2025.10887621]

Guo R H, Ying X H, Qi Y Y, and Qu L. 2024. UniTR: A Unified TRansformer-Based Framework for Co-Object and Multi-Modal Saliency Detection. IEEE Transactions on Multimedia, 26: 7622 - 7635 [DOI: 10.1109/TMM.2024.3369922]

Jerripothula K R, Cai J and Yuan J. 2018. Efficient video object co-localization with co-saliency activated tracklets. IEEE Transactions on Circuits and Systems for Video Technology, 29(3): 744 - 755 [DOI: 10.1109/TCSVT.2018.2805811]

Kamath A, Singh M, LeCun Y, Synnaeve G, Misra I and Carion N. 2021. MDETR-Modulated Detection for End-to-End Multi-Modal Understanding//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual Conference: IEEE: 1780-1790 [DOI:10.48550/arXiv.2104.12763]

Ke L, Ye M, Danelljan M, Liu Y, Tai Y W, Tang C K and Yu F. 2023. Segment anything in high quality. Advances in Neural Information Processing Systems, 36: 29914 - 29934 [DOI: 10.48550/arXiv.2306.01567]

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P and Girshick R. 2023. Segment Anything//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3992-4003[DOI: 10.1109/ICCV51070.2023.00371]

Li L, Han J, Zhang N, Liu N, Khan S, Cholakkal H, Anwer RM and Khan FS. 2023. Discriminative Co-Saliency and Background Mining Transformer for Co-Salient Object Detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 7247-7256 [DOI: 10.48550/arXiv.2305.00514]

Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, Jiang Q, Li C, Yang J, Su H, Zhu J and Zhang L. 2024. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection// European Conference on Computer Vision. Cham: Springer Nature Switzerland: 38-55[DOI:10.1007/978-3-031-72970-6_3]

Matuzevičius D. 2024. A Retrospective Analysis of Automated Image Labeling for Eyewear Detection Using Zero-Shot Object Detectors. Electronics, 13(23): 4763 [DOI:10.3390/electronics13234763]

Minderer M, Gritsenko A and Houlsby N. 2023. Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems, 36: 72983 - 73007 [DOI: 10.48550/arXiv.2306.09683]

Minderer M, Gritsenko A, Stone A, Neumann M, Weissenborn D, Dosovitskiy A, Mahendran A, Arnab A, Dehghani M, Shen Z, Wang X, Zhai X, Kipf T and Houlsby N. 2022. Simple Open-Vocabulary Object Detection//European Conference on Computer Vision. Cham: Springer Nature Switzerland: 728-755 [DOI: 10.1007/978-3-031-20080-9_42]

Prajapati S, Singh T, Hegde C and Chakraborty P. 2024. Evaluation and comparison of visual language models for transportation engineering problems[EB/OL]. [2025-08-08]. https://arxiv.org/pdf/2409.02278.pdf

Su Y K, Deng J L, Sun R Z, Lin G S, Su H J, and Wu Q Y. 2023. A Unified Transformer Framework for Group-Based Segmentation: Co-Segmentation, Co-Saliency Detection and Video Salient Object Detection. IEEE Transactions on Multimedia, 26: 313-325 [DOI: 10.1109/TMM.2023.3264883]

Jiang T T, Liu Y, Ma X, Sun J L. 2021. Multi-path collaborative salient object detection based on RGB-T images[J]. Journal of image and graphics, 26(10): 2388-2399. (蒋亭亭, 刘昱, 马欣, 孙景林. 2021. 多支路协同的RGB-T图像显著性目标检测. 中国图象图形学报, 26(10):2388-2399)[DOI:10.11834/jig.200317]

Wang C, Zha ZJ, Liu D and Xie H. 2019. Robust Deep Co-Saliency Detection with Group Semantic//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, Hawaii, USA: AAAI Press: 8917-8924[DOI:10.1609/aaai.v33i01.33018917]

Wang W and Shen J. 2016. Higher-order image co-segmentation. IEEE Transactions on Multimedia, 18(6): 1011 - 1021 [DOI: 10.1109/TMM.2016.2545409]

Wang Z, Gao Y J, Liu Q J and Wang Y H. 2024. Semantic Enhanced Few-Shot Object Detection// 2024 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates: IEEE: 575-581[DOI:10.1109/ICIP51287.2024.10647133]

Wei L S and Huang J. 2024. An Information Calibration and Sliding Mining Network for Co-Saliency Object Detection. IEEE Transactions on Instrumentation and Measurement, 73: 1-11 [DOI: 10.1109/TIM.2024.3470967]

Wu Y, Song H, Liu B, Zhang K and Liu D. 2023. Co-Salient Object Detection with Uncertainty-Aware Group Exchange-Masking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 19639-19648[DOI: 10.1109/CVPR52729.2023.01881]

Xu B, Jiang Q, Zhao X, Lu C, Liang H and Liang R. 2025. Multidimensional Exploration of Segment Anything Model for Weakly Supervised Video Salient Object Detection. IEEE Transactions on Circuits and Systems for Video Technology, 35(4): 2987 - 2998 [DOI: 10.1109/TCSVT.2024.3368053]

Yu S, Xiao J, Zhang B and Lim EG. 2022. Democracy Does Matter: Comprehensive Feature Mining for Co-Salient Object Detection// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE: 979-988 [DOI:10.48550/arXiv.2203.05787]

Zeng Y, Zhuge Y, Lu H and Zhang L. 2019. Joint learning of saliency detection and weakly supervised semantic segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE: 7223-7233 [DOI: 10.48550/arXiv.1909.04161]

Zhang D, Han J, Li C, Wang J and Li X. 2016. Detection of co-salient objects by looking deep and wide. International Journal of Computer Vision, 120(2): 215 - 232 [DOI: 10.1007/s11263-016-0907-4]

Zhang D W, Han J W, Li C and Wang J D. 2015. Co-saliency detection via looking deep and wide// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA: IEEE: 2994-3002[DOI:10.1109/CVPR.2015.7298918]

Zhang K, Dong M, Liu B, Yuan XT and Liu Q. 2021. DeepACG: Co-Saliency Detection via Semantic-Aware Contrast Gromov-Wasserstein Distance//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 13703-13712[DOI: 10.1109/CVPR46437.2021.01349]

Zhang K, Li T, Shen S, Liu B, Chen J and Liu Q. 2020. Adaptive Graph Convolutional Network with Attention Graph Clustering for Co-Saliency Detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 9050-9059[DOI:10.48550/arXiv.2003.06167]

Zhang K, Wu Y, Dong M, Liu B, Liu D and Liu Q. 2022. Deep object co-segmentation and co-saliency detection via high-order spatial-semantic network modulation. IEEE Transactions on Multimedia, 25: 5733 - 5746 [DOI: 10.1109/TMM.2022.3198848]

Zhang N, Han J, Liu N and Shao L. 2021. Summarize and Search: Learning Consensus-Aware Dynamic Convolution for Co-Saliency Detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 4167-4176[DOI: 10.1109/ICCV48922.2021.00413]

Zhang Z, Jin W, Xu J and Cheng MM. 2020. Gradient-Induced Co-Saliency Detection//European Conference on Computer Vision. Cham: Springer International Publishing: 455-472[DOI: 10.1007/978-3-030-58610-2_27]

Zheng P, Fu H, Fan D P, Fan Q, Qin J, Tai Y W, Tang C K and Van Gool L. 2023. Gconet+: A stronger group collaborative co-salient object detector. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9): 10929 - 10946 [DOI: 10.1109/TPAMI.2023.3264571]

Zhou L J, Mao J N. 2023. Vision Transformer-based recognition tasks: a critical review. Journal of Image and Graphics, 28(10):2969-3003 (周丽娟, 毛嘉宁. 2023. 视觉Transformer识别任务研究综述. 中国图象图形学报, 28(10):2969-3003)[DOI:10.11834/jig.220895]

Zhou Q, Cao J, Leng H, Yin Y, Kun Y and Zimmermann R. 2024. SOGDet: Semantic-Occupancy Guided Multi-view 3D Object Detection[EB/OL]. [2025-09-11]. https://arxiv.org/pdf/2308.13794.pdf

作者简介

陈颖州，男，硕士研究生，主要研究方向为大模型检索增强生成。 E-mail:2024221407@stu.fynu.edu.cn

王斌,男,硕士研究生,主要研究方向为协同显著性目标检测。 E-mail:3143454479@qq.com
王峰,男,教授,主要研究方向为计算机视觉。 E-mail:wf111625@163.com
赵佳,男,教授,主要研究方向为计算机视觉。 E-mail:zhaojia11b@mails.ucas.ac.cn