面向灾后视觉场景理解的开放词汇与监督学习方法对比

Anna Michailidou*、Georgios Angelidis*、Vasileios Argyriou†、Panagiotis Sarigiannidis‡、Georgios Th. Papadopoulos*§
*雅典哈罗科皮奥大学信息学与远程信息处理系，雅典，希腊
†金斯顿大学网络与数字媒体系，伦敦，英国
‡西马其顿大学电气与计算机工程系，科扎尼，希腊§Archimedes，雅典娜研究中心，雅典，希腊 电子邮箱：{amichailidou, gangelidis, g.th.papadopoulos}@hua.gr, Vasileios.Argyriou@kingston.ac.uk, psarigiannidis@uowm.gr

arXiv:2603.01324v1 [cs.CV] 1 Mar 2026

摘要——航空影像对于大规模灾后损害评估至关重要。由于场景杂乱、视觉变异性以及强烈的跨事件域偏移，自动化解译仍然具有挑战性；而监督方法仍依赖成本高昂、覆盖范围有限的任务特定标注，难以覆盖各类灾害类型和地区。近期的开放词汇与基础视觉模型提供了一种有吸引力的替代方案，可降低对固定标签集和大量任务特定标注的依赖。相反，它们利用大规模预训练和视觉-语言表征。这些特性对于灾后领域尤为重要，因为其中的视觉概念往往模糊不清，且数据可得性受限。在本工作中，我们对监督学习与开放词汇视觉模型在灾后场景理解中的表现进行了对比评估，重点关注多个数据集上的语义分割与目标检测，包括FloodNet+、RescueNet、DFire和LADD。我们考察了不同学习范式的性能趋势、失效模式及实际权衡，从而为它们在真实世界灾害应急响应中的适用性提供洞见。在所有评估基准测试中，最显著的结论是：当标签空间固定且标注可用时，监督训练仍然是最可靠的方法，尤其是在杂乱场景中的小物体检测和精细边界勾勒方面。

索引词—灾后分析、航空影像、语义分割、目标检测、开放词汇模型、基础模型。

I. 引言

快速的灾后态势感知日益依赖来自无人机（UAV）和卫星的遥感影像，从而能够对灾害和损毁进行大范围制图 [1]。然而，灾后航空影像具有高度异质性 [2]–[5]：目标呈现极大的尺度变化，场景中存在杂乱背景和遮挡，且目标外观会随灾害类型和环境条件的不同而显著变化。这些因素，加上标注数据有限，持续制约着模型的可靠部署，也促使研究者探索超越狭窄固定词汇监督的稳健学习策略。

本研究获得了欧盟“地平线欧洲”研究与创新计划的资助，资助协议编号为 101168042，项目名称为 TRIFFID（提升应急救援人员效率的自主机器人援助）。本文所表达的观点和意见仅代表作者本人，并不一定反映欧盟或欧盟委员会的立场。

这一进展得益于：（i）针对航空与卫星影像中灾害与损毁理解的基准测试不断改进 [2], [3], [6], [7]；（ii）通用计算机视觉架构 [8], [9] 及预训练 [10],[11]不断进步。借助大规模监督式预训练和架构改进 [12], 的卷积骨干网络，在密集预测和实时部署方面依然表现出色；而视觉变换器 [13] 则提升了复杂场景的全局上下文建模能力。尤其是在目标检测领域，该方向已从快速的单阶段检测器 [14] 发展到基于 Transformer 的端到端集合预测 [15],，其中也包括面向实时运行的设计 [16], [17]。

近期的一项进展是开放词汇（OV）和基础模型（FMs）的兴起。它们通过利用视觉-语言预训练[18]或可提示分割 [19]，降低了对固定标签集的依赖。这些模型能够在检测和分割中实现文本条件预测，因此在灾害应急响应中颇具应用价值，因为目标概念可能发生变化，且标注数据有限。

本文通过一项比较评估研究灾后航空感知，在统一的实验方案下，涵盖语义分割和目标检测，对比闭集监督学习方法与开放词汇替代方法。我们的贡献包括：

- 对比闭集与开放词汇方法在灾后航空场景理解中的表现，涵盖语义分割和目标检测。
- 分析卷积神经网络（CNNs）与 Transformer 在杂乱航空影像中的架构趋势。
- 总结常见失效模式（例如小物体、遮挡和视觉模糊区域）以及与实际部署相关的权衡，为实践提供参考。

本文其余部分组织如下。第II节回顾了有监督与开放词汇灾后场景理解的相关工作。第III节介绍了所选分析方法及其分类体系。第IV 节描述了实验方案，并分析

相应结果。最后，第V 节总结了结论，并概述了未来工作的潜在方向。

II. 相关工作

公开基准在规范评估实践、提升灾后感知任务的真实性方面发挥了重要作用，涵盖洪水场景理解[6]，、基于无人机的损伤分割 [2],以及涉及多种灾害的大规模数据集 [20]。相关的建筑物损伤评估卫星基准进一步展现了损伤模式的多样性以及跨事件泛化所面临的挑战 [7]。然而，尽管数据资源日益丰富，大多数研究仍局限于单一灾害数据集或单一监督范式，因此难以清晰了解不同架构和学习策略如何迁移到不同灾害情境中。

在语义分割领域，经典的端到端密集预测方法从全卷积表述 [21] 和带有跳跃连接的编码器-解码器结构 [22] ，逐步发展为多尺度上下文聚合和边界细化；DeepLab [23] 的空洞上下文和 PSPNet [24]的金字塔池化便是其中的代表。近期研究借助 Transformer [25] 和掩码注意力解码器 [26], 增强全局推理能力，这对于灾害影像中常见的大面积同质区域和复杂空间布局可能大有裨益。在目标检测领域，YOLO [14],等实时单阶段检测器仍广泛用于航空部署，而 Transformer检测器 [15] 则将检测重新表述为不依赖锚框或非极大值抑制（NMS）的集合预测，并推动了更稳健的训练方法和实时变体的发展 [16], [27]–[29], ，其中常借助自监督表征 [30]。

近年来，开放词汇方法和基础模型引入了语言条件预测与可提示分割，减少了对特定任务标签的依赖，并支持在推理时灵活定义用户指定的概念 [18], [19]。开放词汇检测和视觉定位支持文本条件定位 [31], [32], ，而开放词汇分割则将视觉-语言语义迁移到像素空间 [33]–[35]。

灾后影像给这些范式带来了不同寻常的挑战，例如严重杂乱、反射、烟雾、极端尺度变化以及跨事件域偏移。现有研究大多在灾害场景之外 [36] 或仅针对范围狭窄的单一数据集 [2], [6],[7]评估此类方法。最常见的情况是，这些研究未纳入开放词汇方法 [37], ，因此难以描述真实灾害条件下的失效模式和鲁棒性。

特别是，据我们所知，目前尚无在同一协议下进行的直接比较评估，能够（i）比较闭集监督流程与开放词汇替代方法，（ii）涵盖语义分割和目标检测两项任务，以及（iii）分析开放词汇方法在何处以及为何失效。本文针对这些空白，在灾后航空影像上对两项任务中的代表性监督方法和开放词汇方法进行基准测试，并报告性能趋势和定性失效模式。

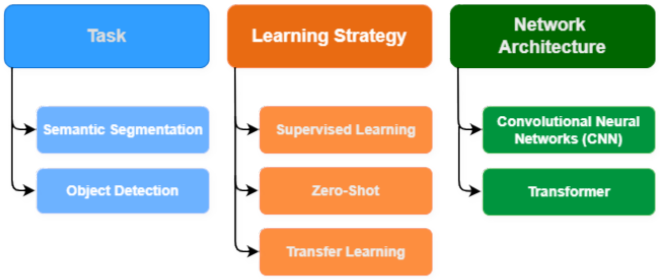

图 1. 灾后视觉场景理解方法的分类体系

III. 实验框架

我们评估多灾害场景（地震、洪水、野火）下灾后场景理解的两项核心任务：（i）语义分割，为每个像素分配标签；以及（ii）目标检测，定位感兴趣的目标。方法的选择旨在（a）体现灾害与遥感文献中的常见基线，并（b）涵盖互补的架构系列（卷积神经网络、Transformer）和学习范式（闭集监督与开放词汇／零样本）。这支持在杂乱场景、尺度变化和域偏移条件下进行受控比较。

A. 闭集监督语义分割

对于监督分割，我们纳入了成熟的卷积神经网络基线和近期的 Transformer 设计，以利用互补的归纳偏置——卷积神经网络中的多尺度上下文聚合和边界细化，以及 Transformer 中的全局上下文建模。

DeepLabV3+ [23] 是一个性能强劲的卷积神经网络基线，它结合空洞卷积与空洞空间金字塔池化（ASPP）来捕获多尺度上下文，同时在高分辨率场景中保持良好的边界质量。

PSPNet [24]通过金字塔池化聚合全局上下文，可作为航空影像的稳健参考；在这类影像中，局部纹理可能模糊不清（例如阴影、反射、瓦砾）。

CCNet [38]引入交叉注意力，以低于完整自注意力的计算成本高效捕获长程依赖。我们将其纳入研究，以检验结构化的全局上下文注意力相较于纯卷积上下文聚合，是否能提升杂乱灾害场景中的性能。

SegFormer [25] 是一种高效的 Transformer 分割设计，采用分层混合 Transformer（MiT）编码器和轻量级解码器，能够进行长程建模，同时适用于大规模遥感输入。

Mask2Former [26]被纳入作为一种现代的掩码注意力分割框架。它通常使用强大的主干网络进行实例化（例如 ResNet [39]或层次化 Transformer）。由于我们的数据集提供语义标注，我们严格在语义设定下对其进行评估。

B. 开放词汇语义分割

为了将闭集监督与减少对任务特定标签依赖的方法加以对比，开放词汇分割方法在使用视觉-语言表示的零样本文本提示下进行评估，此外还在模型适配到目标域的迁移学习设定下进行评估。

MaskCLIP [33]采用 CLIP 风格的视觉—语言预训练方法 [18]进行密集预测，通常使用 CLIP 视觉变换器图像编码器（视觉变换器 [13]），将语言对齐语义迁移为像素标签，而无需在目标数据集类别上进行训练。

FC-CLIP [34]使用基于 CLIP 的特征和轻量级密集预测头进行开放词汇分割，其中 CLIP 特征保持冻结，从而有力地评估语言对齐特征在多样性极高的灾后类别上的迁移能力。

SegEarth-OV-3 [40] 旨在通过调整 SAM3 [41]以适用于密集航空场景，在无需训练的设置下实现遥感开放词汇语义分割（OVSS）。它通过简单的掩码融合策略结合 SAM3 的语义和实例输出，并利用模型的存在性评分抑制不存在的类别，从而减少基于图块的遥感处理过程中由大规模词汇表导致的假阳性。我们将其视为适用于航空灾害影像的领域感知开放词汇基线。

C. 闭集监督目标检测

在目标检测方面，纳入了实时卷积神经网络检测器和高效的 DETR 风格 Transformer，以比较卷积神经网络与 Transformer 在小目标及杂乱场景下的表现。

YOLO 系列 [14] 涵盖了本文采用的一阶段实时检测器，作为具有代表性的卷积神经网络基线。YOLOv5 提供了广泛采用的基于锚框的参考方案，其骨干网络为 CSPDarknet，并采用 SPPF/PAN 特征聚合；YOLOv8 和 YOLO11 则通过无锚框预测及改进的检测头优化了流程，从而实现更优的精度与延迟权衡。YOLO26 [42] 则通过移除 DFL 并支持端到端无NMS推理，进一步面向边缘部署，以减少后处理开销及实时系统中的延迟抖动。

RT-DETRv2 [17] 是一种改进的实时 DETR 风格检测器，在保留端到端集合预测的同时，优化了多尺度特征处理，以提升实用性和鲁棒性。本文将其纳入作为现代实时 Transformer 基线；它通过选择性多尺度采样和便于部署的算子选择，力求在不牺牲速度的前提下提升杂乱场景及尺度变化下的精度。

D. 开放词汇目标检测

开放词汇检测采用零样本文本提示进行评估，从而能够在无需针对固定标签集重新训练的情况下，对用户指定的概念进行定位。

OWL-ViT [31]使用与 CLIP 对齐的视觉变换器图像编码器（ViT [13], CLIP [18]）进行文本条件检测，是基于提示的定位的简洁基线。

表 I 本工作所用数据集。

数据集	应用场景	Task	图像数量	类别数量	年份
FloodNet+ [6]	洪水	语义分割	2289	9	2024
RescueNet [2]	地震	语义分割	4494	10	2023
D-Fire [5]	野火	检测	21527	2	2021
LADD [4]	SAR	检测	1365	1	2023

Grounding DINO [32]结合了 DETR 风格的检测与基于图像–文本数据的视觉定位预训练，与早期开放集检测器相比，提升了文本到边界框的视觉定位质量。

YOLOE [43] 是一种实时开放词汇 YOLO 变体，将单阶段检测/分割扩展到文本提示、视觉提示和无提示推理，并整合于单一模型中。它实现了用于区域–文本和视觉提示条件化的轻量级对齐模块（RepRTA、SAVPE）以及一种无提示方案（LRPC）。Ultralytics 的 [44] 无提示检查点附带固定的 4,585 类词汇表（RAM++ 标签集 [45]），从而能够在部署时像标准闭集 YOLO 一样进行“无提示”操作。

IV. 实验评估

本节介绍所选语义分割和目标检测方法在灾害遥感数据集上的实验配置和定量结果，这些数据集涵盖地震、洪水、野火和空中人员检测。所有实验均遵循一致的方案：在适用情况下，数据集划分遵循各自数据集的惯例；模型采用基于参考实现的标准设置进行训练或微调；结果则使用密集预测和检测领域广泛采用的指标进行报告。

A. 数据集

我们在 RescueNet [2] 上评估用于震后损伤评估的语义分割，在 FloodNet+ [3]上评估用于洪水场景理解的语义分割。在与搜索和救援（SAR）任务相关的 LADD [4], 空中行人检测以及 D-Fire [5] 火灾与烟雾检测上评估目标检测。这些数据集呈现了互补的挑战，包括严重的杂乱背景、尺度变化、遮挡以及显著的类内差异。

B. 评估指标

对于语义分割，我们按照语义分割基准测试（例如 PASCAL VOC [48]）中采用的标准定义，报告所有 C 类别上的平均交并比（mIoU）。为揭示严重像素不平衡下的类别特定表现，我们还报告逐类交并比。

对于目标检测，我们报告交并比阈值为 0.50 时的平均精度均值 (mAP_{50})，这与检测基准测试（例如 PASCAL VOC）中的常见做法一致 [48]。在评估协议支持的情况下，我们还报告 COCO 风格的 $mAP^{[50:95]}$，以反映模型在更严格交并比阈值下的定位质量 [49]。

表 II RescueNet 与 FloodNet+ 基准测试上的总体分割性能。

模型	骨干网络	学习范式	神经网络类型	RescueNet mIoU	FloodNet+ mIoU
监督学习					
PSPNet [24]	ResNet-101 [39]	监督学习	CNN	78.90%	76.87%
CCNet [38]	ResNet-101 [39]			**79.95**%	75.15%
DeepLabV3+ [23]	ResNet-101 [39]			78.49%	76.27%
SegFormer [25]	分层混合 Transformer-B4 [25]		Transformer	76.10%	75.61%
Mask2Former [26]	SwinT [46]			77.55%	**79.20**%
开放词汇					
MaskCLIP [33]	视觉变换器 [13]	零样本	Transformer	11.21%	10.19%
SegEarth-开放词汇-3 [40]	SAM3 [41]			**26.16**%	**46.08**%
FC-CLIP [34]	ConvNeXt-L [47]		CNN	21.66%	45.43%
MaskCLIP [33]	视觉变换器 [13]	迁移学习	Transformer	28.94%	24.38%
SegEarth-开放词汇-3 [40]	SAM3 [41]			**52.63**%	**53.92**%
FC-CLIP [34]	ConvNeXt-L [47]		CNN	51.85%	52.34%

C. 实验设置

所有实验均采用统一训练协议，以确保公平比较。除非另有说明，模型均使用预训练权重初始化，并采用 AdamW 优化器进行优化，学习率为 0.01，权重衰减为 10^{-4}，训练最多 100 个轮次，并使用早停（耐心值为 10）；批大小设为 8，输入分辨率为 768 像素。每项实验均应用标准数据增强，以减轻过拟合。所有数据集划分均遵循各基准所采用的标准协议。

对于迁移学习，采用部分微调，仅在训练期间更新较深层中的一部分层和任务特定头。在此设置下，使用 AdamW，学习率为 10^{-4}，权重衰减为 0。05，训练 10 个训练轮次，批大小设为 2，图像分辨率为 768 像素。

D. 评估结果

为突出本研究的核心重点，结果按学习范式组织，分别讨论闭集监督方法与开放词汇方法，并随后针对每个数据集展开讨论。

1) 语义分割：

a) 闭集监督分割： 如表II所示，在两个分割数据集上，监督模型的表现都显著优于开放词汇方法。这与监督模型直接针对目标标签空间和域内外观统计进行优化相一致。相比之下，开放词汇流水线必须将大规模预训练中获得的语义迁移到具有显著域偏移（损伤纹理、瓦砾、反射）和严重类别不平衡的灾后航空影像中。

在监督学习方法中，各方法之间的差异相对较小，整体上与架构的归纳偏置相符。在 RescueNet 上，CCNet 的表现略优于 PSP-Net 和 DeepLabV3+，这可能是因为交叉注意力能够传播场景级上下文，适用于损伤证据在空间上分散的情况。在 FloodNet+ 上，Mask2Former排名靠前，这与其掩码级解码相符，后者有助于区域推理以及大面积“背景”区域（如水域）与细长结构之间的边界细化。

b) 开放词汇分割： 零样本开放词汇分割的表现仍远低于监督分割，这反映出视觉-语言预训练与灾害特定航空场景之间的对齐有限，同时也说明将数据集分类体系映射到简短提示词存在歧义。经过迁移学习后，平均交并比显著提升，表明即使是适度的微调，也有助于模型将其表征调整至目标域。开放词汇方法之间的性能差距依然存在，这取决于复杂背景下的掩码质量，以及微调使预测结果与数据集分类体系对齐的效果。</output>

2) 目标检测：

a) 闭集监督检测： 表 III 显示，监督检测器在两个基准数据集上均具有明显优势，尤其是在航空小目标占主导的 LADD 上。得益于密集多尺度特征和定位先验，单阶段卷积神经网络检测器在小目标场景中依然表现强劲；而实时 DETR 风格模型则受益于全局推理和端到端匹配，这有助于提升杂波环境下的一致性。

b) 开放词汇检测：零样本与迁移学习： 零样本开放词汇检测的表现高度依赖模型和数据集（表III），这反映出文本—区域视觉定位对域偏移和小目标可见性的敏感性。基于对齐的设计通常比纯文本条件基线具有更强的零样本定位能力，但航空小目标和背景杂波仍会导致漏检和提示敏感性。迁移学习缩小了 LADD 上的差距，表明域内适配主要提升候选区域质量和定位能力，进而实现更有效的文本对齐。

3) 数据集特定观察：

a) RescueNet：在闭集监督下，视觉上易于区分的类别（树木、水体、道路畅通）在各架构上均持续取得较高的交并比，表明域内训练已基本使这些类别趋于饱和。其余误差

表 III LADD 和 D-Fire 基准测试上的总体检测性能。

模型	骨干网络	学习范式	神经网络类型	LADD		D-Fire	
				平均精度均值$_{50}$	平均精度均值$_{50:95}$	平均精度均值$_{50}$	平均精度均值$_{50:95}$
监督学习							
YOLOv11l [14]	基于CSP [44]	监督学习	CNN	92.2%	58.2%	**80.6%**	48.9%
YOLO26L [42]	基于CSP [42]			93.6%	61.8%	80.3%	**49.0%**
RT-DETRv2-L [17]	ResNet [39]			**94.7%**	**62.68%**	77.43%	43.97%
开放词汇							
YOLOE26 [43]	基于YOLO26 [42]	零样本	CNN	24.7%	11.3%	14.9%	7.1%
Grounding DINO [32]	SwinT [46]		Transformer	**61.0%**	28.2%	27.5%	13.4%
OWL-ViT [31]	视觉变换器 [13]			6.2%	1.9%	**36.4%**	**18.0%**
YOLOE26 [43]	基于YOLO26 [42]	迁移学习	CNN	37.5%	19.0%	24.2%	10.7%
Grounding DINO [32]	SwinT [46]		Transformer	**92.2%**	**52.1%**	**65.6%**	**33.1%**
OWL-ViT [31]	视觉变换器 [13]			68.6%	38.4%	48.9%	19.4%

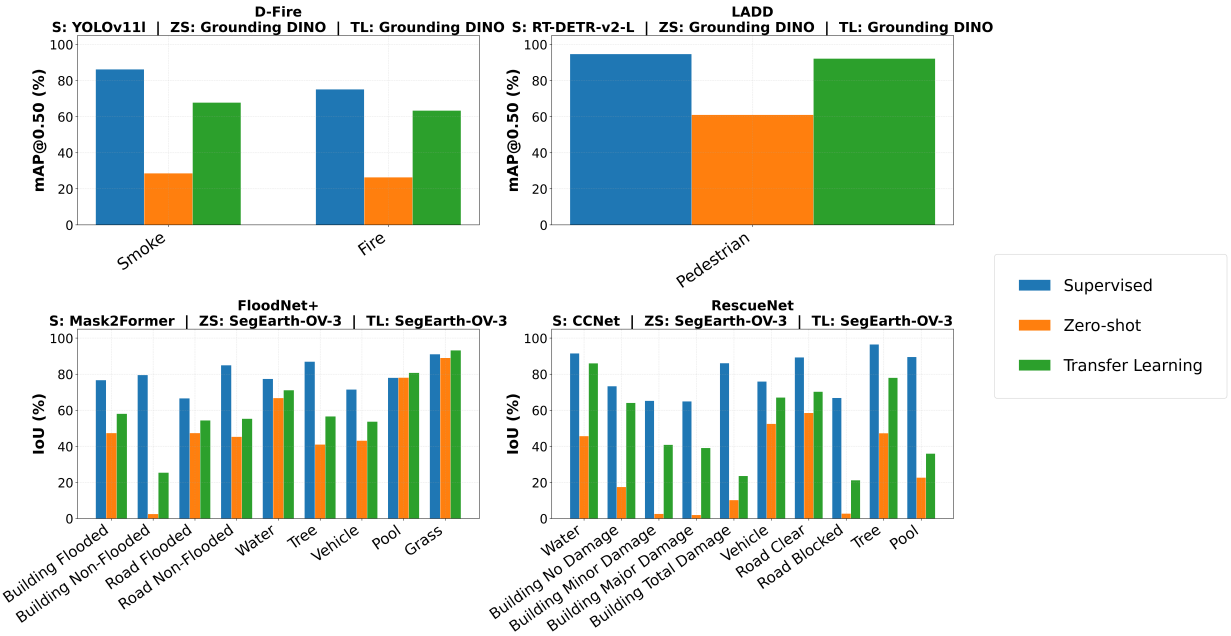

图2. 各数据集在不同训练设置下顶尖模型的类别级图示

集中于需要上下文解释的类别，尤其是中等损毁等级（轻微/严重）和道路受阻；在这些类别中，瓦砾、阴影和混合材料降低了可分性。CCNet取得了最佳的总体平均交并比，并且在这些更难的类别上相对更强，这与在局部证据不完整时传播场景级上下文的优势相一致。在开放词汇设置中，零样本预测大多只能恢复粗粒度语义，而细粒度损毁等级和道路受阻仍然较弱；迁移学习显著提升了大多数类别的表现，但同样这些高度依赖上下文的类别仍在限制性能。

b)FloodNet+：对于监督分割，所有方法都能稳健地分割较大的“背景”区域（草地、树木、未被洪水淹没的道路），而被洪水淹没的道路和较小类别（车辆、水池）仍是主要的误差来源。Mask2Former 的平均交并比最高，这主要是因为它改善了这些难分类别，表明在反光水面和杂乱场景中，它具有更好的区域级细化和边界处理能力。在开放词汇分割中，零样本表现对广义自然类别的识别相对较强，但对洪水特定区分（被淹与未被淹）的识别较弱。迁移学习在洪水特定类别和小目标类别上带来了最大幅度的提升，这表明有限的适应过程主要是将决策边界校准至目标分类体系和图像统计特性。

c) LADD：对于单一目标类别（行人），性能取决于小目标召回率和精确定位。监督检测器的结果表明，域内优化足以在杂乱背景下对航空影像中的小目标实例实现较高的检出灵敏度；而开放词汇检测器受限于较弱的小目标定位能力，且迁移学习只能使其性能得到部分恢复。

d) D-Fire：在各类监督检测器中，烟雾始终比火灾更容易检测，这与烟雾占据的区域更大相符；相比之下，火灾目标更小、更零散，外观也更加多变。开放词汇检测在零样本模式下也呈现出相同趋势，而迁移学习通过使候选区域质量和定位适应航空影像采集条件，提升了这两类目标的检测效果。

V. 结论

本研究对封闭标签集监督方法和开放词汇方法在灾后航空场景理解中的表现进行了比较评估，涵盖洪水、地震、野火和合成孔径雷达场景下的语义分割与目标检测。在所有基准测试中，监督训练仍是最可靠的方法：当标签空间固定且有标注可用时，监督模型始终能取得最高准确率，且表现最稳定，尤其是在杂乱场景中的小目标检测和精细边界勾勒方面。

当密集标签稀缺时，开放词汇流水线提供了一种实用替代方案，但在灾害特定领域偏移和细粒度分类体系下，零样本性能明显较弱。迁移学习大幅缩小了这一差距，表明开放词汇模型作为预训练起点最为有效，而非可直接替代高分辨率航空影像中监督系统的方案。总体而言，结果表明，不同范式面临的主要挑战是相同的，包括小物体、遮挡和细微的类内变化；但对于稳健定位和边界质量，监督仍起决定性作用。

未来工作将把评估扩展到更多灾害类型和采集条件，提升开放词汇方法在航空场景中的空间精度，并利用基础模型进行半自动标注，以扩大灾害特定训练数据规模并增强跨事件泛化能力。

此外，将使用功能需求（FRs）评估该系统，以验证其在模拟运行场景中的有效性。进一步而言，该系统将连接实时人工智能事件检测器 [50]–[52], ，以支持涉及人机交互的场景 [53]–[55]以及范围更广的安全响应事件 [56], ，同时支持必要的可解释人工智能（XAI）流程 [9], [57]。

参考文献

[1] J. Cani、P. Koletsis、K. Foteinos、I. Kefaloukos、L. Argyriou、M. Fale-lakis、I. Del Pino、A. Santamaria-Navarro、M. ˇCech、O. Severa 等，“Triffid：用于提升应急救援人员效率的自主机器人辅助系统”，载于*2025*年第六届国际电子工程与信息技术会议（*EEITE*）。IEEE，2025，第1–9页。[2] M. Rahnemoonfar、T. Chowdhury和R. Murphy，“RescueNet：用于自然灾害损害评估的高分辨率无人机语义分割数据集”，*Scientific Data*，第10卷，第1期，第913页，2023。[3] D. Zhao、J. Lu和B. Yuan，“观察、感知并回答：用于遥感图像灾后评估的统一高分辨率基准”，*IEEE*地球科学与遥感汇刊，第62卷，第1–14页，2024。[4] M. Shuranov 等，“LADD：Lacmus无人机数据集”，2023，数据集。[5] P. V. A. de Venancio、A. C. Lisboa和A. V. Barbosa，“一种基于深度卷积神经网络、适用于低功耗资源受限设备的自动火灾检测系统”，*Neural Computing and Applications*，第34卷，第18期，第15 349–15368页，2022。

[6] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy，《Floodnet：用于洪后场景理解的高分辨率航空影像数据集》，IEEE Access，第9卷，89 654，2021，89 644页。[7] R. Gupta, R. Hosfelt, S. Sajeev, N. Patel, B. Goodman, J. Doshi, E. Heim, H. Choset, and M. Gaston，《xbd：用于根据卫星影像评估建筑物损害的数据集》，arXiv预印本 arXiv:1911 09296，2019。[8] J. Cani, C. Diou, S. Evangelatos, V. Argyriou, P. Radoglou-Grammatikis, P. Sarigiannidis, I. Varlamis, and G. T. Papadopoulos，《使用深度学习技术检测X射线成像中的违禁物品：比较评估》，IEEE Access，2026。[9] N. Rodis, C. Sardianos, P. Radoglou-Grammatikis, P. Sarigiannidis, I. Varlamis, and G. T. Papadopoulos，《多模态可解释人工智能：方法学进展与未来研究方向综述》，IEEE Access，第12卷，159 820，2024，159 794页。[10] S. Konstantakos, J. Cani, I. Mademlis, D. I. Chalkiadaki, Y. M. Asano, E. Gavves, and G. T. Papadopoulos，《低数据场景中的自监督视觉学习：比较评估》，Neurocomputing，第620卷，第129199页，2025。[11] P. Alimisis, I. Mademlis, P. Radoglou-Grammatikis, P. Sarigiannidis, and G. T. Papadopoulos，《图像数据增强中的扩散模型进展：方法、模型、评估指标与未来研究方向综述》，Artificial Intelligence Review，第58卷，第4期，第112页，2025。[12] O. Elharrouss, Y. Akbari, N. Almadeed, and S. Al-Maadeed，《骨干网络综述：用于计算机视觉中深度学习和深度强化学习方法的特征提取网络》，Computer Science Review，第53卷，第100645页，2024。[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly 等，《一张图像胜过16×16个词：用于大规模图像识别的Transformer》，arXiv预印本 arXiv:2010.11929，2020。[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi，《YOLO：统一的实时目标检测》，载于《IEEE计算机视觉与模式识别会议论文集》（CVPR），2016年，第779–788页。[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko，《使用Transformer进行端到端目标检测》，载于《欧洲计算机视觉会议》。Springer，2020年，第213–229页。[16] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, a nd J. Chen，《DETR在实时目标检测方面胜过YOLO》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2024年，第16 965–16 974页。[17] W. Lv, Y. Zhao, Q. Chang, K. Huang, G. Wang, and Y. Liu，《Rt-detrv2：采用免费技巧集改进实时检测Transformer基线》，2024 [ ] 2407 17140 在[18]线。可访问：https://arxiv.org/abs/。A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark 等，《通过自然语言监督学习可迁移视觉模型》，载于《国际机器学习会议》。PmLR，2021年，第8748–8763页。[19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo 等，《Segment Anything》，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第4015–4026页。[20] E. Weber, D. P. Papadopoulos, A. Lapedriza, F. Ofli, M. Imran, and A. Torralba，《Incidents1M：包含自然灾害、损害和事故图像的大规模数据集》，IEEE模式分析与机器智能汇刊，第45卷，第4期，第4768–4781页，2022。[21] J. Long, E. Shelhamer, and T. Darrell，《用于语义分割的全卷积网络》，载于《IEEE计算机视觉与模式识别会议论文集》，2015年，第3431–3440页。[22] O. Ronneberger, P. Fischer, and T. Brox，《U-net：用于生物医学图像分割的卷积网络》，载于《医学图像计算与计算机辅助干预国际会议》。Springer，2015年，第234–241页。[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam，《用于语义图像分割的编码器-解码器与空洞可分离卷积》，载于《欧洲计算机视觉会议论文集》（ECCV），2018年，第801–818页。

[24] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, 《金字塔场景解析网络》，载于《IEEE计算机视觉与模式识别会议论文集》，2017年，第2881–2890页。[25] E. Xie 等，《Segformer：使用Transformer进行语义分割的简洁高效设计》，载于《神经信息处理系统进展》（NeurIPS），2021年。[26] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar，《用于通用图像分割的掩码注意力掩码Transformer》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第1290–1299页。[27] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum，《DINO：采用改进去噪锚框的DETR，实现端到端目标检测》，arXiv预印本 arXiv:2203.03605，2022。[28] I. Robinson, P. Robicheaux, M. Popov, D. Ramanan, and N. Peri，《Rf-detr：面向实时检测Transformer的神经架构搜索》，arXiv预印本 arXiv:2511.09554，2025。[29] S. Huang, Y. Hou, L. Liu, X. Yu, and X. Shen，《实时目标检测遇上DINOv3》，2025年。[30] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby 等，《DINOv2：无需监督地学习鲁棒视觉特征》，2023年。[31] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen 等，《简易开放词汇目标检测》，载于《欧洲计算机视觉会议》。Springer，2022年，第728–755页。[32] S. Liu 等，《Grounding DINO》，载于《欧洲计算机视觉会议论文集》（ECCV），2024年，arXiv:2303.05499。[33] X. Dong 等，《MaskCLIP》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），2023年。[34] Q. Yu 等，《卷积难以消亡》，载于《神经信息处理系统进展》（NeurIPS），2023年。[35] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello，《基于文本到图像扩散模型的开放词汇全景分割》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2023年，第2955–2966页。[36] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu，《基于掩码适配CLIP的开放词汇语义分割》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2023年，第7061–7070页。[37] M. H. Asad, M. M. Asim, M. N. M. Awan, and M. H. Yousaf，《使用无人机影像语义分割进行自然灾害损害评估》，载于2023年国际机器人与工业自动化会议（ICRAI）。IEEE，2023年，第1–7页。[38] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu，《CCNet：用于语义分割的交叉注意力》，载于《IEEE/CVF国际计算机视觉会议论文集》，2019年，第603–612页。[39] K. He, X. Zhang, S. Ren, and J. Sun，《用于图像识别的深度残差学习》，载于《IEEE计算机视觉与模式识别会议论文集》，2016年，第770–778页。[40] K. Li, S. Zhang, Y. Deng, Z. Wang, D. Meng, and X. Cao，《Segearth-ov3：探索SAM 3在遥感图像开放词汇语义分割中的应用》，arXiv预印本 arXiv:2512.08730，2025。[41] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang 等，《SAM 3：借助概念实现Segment Anything》，arXiv预印本 arXiv:2511.16719，2025。[42] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee，《YOLO26：实时目标检测的关键架构改进与性能基准测试》，arXiv预印本 arXiv:2509.25164，2025。[43] A. Wang, L. Liu, H. Chen, Z. Lin, J. Han, 和 G. Ding，《YOLOE：实时看见万物》，载于 IEEE/CVF国际计算机视觉会议论文集，2025年，第 24 591–24 602 页。[44] G. Jocher, J. Qiu, 和 A. Chaurasia，《Ultralytics YOLO（版本 8.0.0）[计算机软件]》，Ultralytics，2023年，网站：https://ultralytics.com。[在线]。可访问：https://github.com/ultralytics/ultralytics [45] X. Huang, Y.-J. Huang, Y. Zhang, W. Tian, R. Feng, Y. Zhang, Y. Xie, Y. Li, 和 L. Zhang，《采用多粒度文本监督的开放集图像标注》，载于第33届 ACM国际多媒体会议论文集，2025年，第 4117–4126 页。[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, 和 B. Guo，《Swin Transformer：采用移位窗口的分层视觉Transformer》，载于 IEEE/CVF国际计算机视觉会议论文集，2021年，第 10 012–10 022 页。[47] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, 和 S. Xie，《面向2020年代的卷积网络》，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2022年。[48] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, 和 A. Zisserman，《PASCAL视觉物体类别（VOC）挑战赛》，《国际计算机视觉期刊》，第88卷，第2期，第 303–338 页，2010年。[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, 和 C. L. Zitnick，《微软COCO：上下文中的常见物体》，载于欧洲计算机视觉会议。Springer，2014年，第 740–755 页。[50] M. Linardakis, I. Varlamis, 和 G. T. Papadopoulos，《基于视觉输入的手势识别综述》，IEEE Access，2025年。[51] K. Foteinos, M. Linardakis, P. Radoglou-Grammatikis, V. Argyriou, P. Sarigiannidis, I. Varlamis, 和 G. T. Papadopoulos，《深度学习视觉手势识别：方法、数据集、挑战与未来研究方向的全面综述》，arXiv预印本 arXiv:2507.04465，2025年。[52] M. Linardakis, I. Varlamis, 和 G. T. Papadopoulos，《使用多个智能体和最优目标分配进行分布式迷宫探索》，IEEE Access，第12卷，第 101 407–101 418 页，2024年。[53] G. T. Papadopoulos, M. Antona, 和 C. Stephanidis，《面向开放且可扩展的大规模多智能体人机协作学习认知人工智能架构》，IEEE Access，第9卷，73 909，2021年 73 890页。–。[54] G. T. Papadopoulos, A. Leonidis, M. Antona, 和 C. Stephanidis，《在联邦式人机协作环境中，由用户画像驱动的大规模多智能体示范学习》，载于国际人机交互会议。Springer，2022年，第 548–563 页。[55] M. Moutousi, A. El Saer, N. Nikolaou, A. ,Sanfeliu, A. Garrell, L. Bl´aha, M. ˇCech, E. K. Markakis, I. Kefaloukos, M. Lagomarsino 等，《Tornado：面向处理小型、柔软且可变形物体的机器人的基础模型》，载于2025年第6届国际电子工程与信息技术会议（EEITE）。IEEE，2025年，第 1–13 页。[56] I. Mademlis, M. Mancuso, C. Paternoster, S. Evangelatos, E. Finlay, J. Hughes, P. Radoglou-Grammatikis, P. Sarigiannidis, G. Stavropoulos, K. Votis 等，《看不见的军备竞赛：非法商品贩运的数字趋势与人工智能赋能的应对措施》，IEEE《技术与社会汇刊》，第6卷，第2期，第 181–199 页，2024年。[57] S. Evangelatos, E. Veroni, V. Efthymiou, C. Nikolopoulos, G. T. Papadopoulos, 和 P. Sarigiannidis，《探索能量景观以获得最简反事实解释：网络安全及其他领域的应用》，IEEE《人工智能汇刊》，2025年。