arXiv:2603.01029v1 [cs.CV] 1 Mar 2026

面向道路异常分割的视觉-语言特征对齐

何卓林、唐嘉诚、浦健*、薛向阳

摘要——复杂环境中的安全自主系统需要稳健的道路异常分割，以识别未知障碍物。然而，现有方法通常依赖像素级统计来判断某个区域是否呈现异常。这种依赖会导致天空或植被等语义上正常的背景区域出现较高的误报率，同时对真正的分布外（OOD）实例的召回率较低，从而给机器人感知和决策带来安全风险。为应对这些挑战，我们提出了 VL-Anomaly，这是一种融入预训练视觉语言模型（VLM）语义先验的视觉-语言异常分割框架。具体而言，我们设计了一种由提示学习驱动的对齐模块，将 Mask2Former 的视觉特征与已知类别的 CLIP 文本嵌入对齐，从而有效抑制背景区域中虚假的异常响应。在推理阶段，我们进一步提出一种多源推理策略，融合文本引导相似度、基于 CLIP 的图文相似度和检测器置信度，利用互补信息源实现更可靠的异常预测。大量实验表明，VL-Anomaly 在包括 RoadAnomaly、SMIYC 和 Fishyscapes 在内的基准数据集上取得了当前最优性能。代码已发布于 https://github.com/NickHezhuolin/VL-aligner-Road-anomaly-segment。

I. 引言

语义分割在使自动驾驶系统 [1]–[3]和移动机器人 [4] 实现对周围环境的细粒度理解方面发挥着关键作用。通常，分割模型经过训练，用于识别一组固定的预定义类别 [5], [6]。然而，在真实场景中，这些模型往往会遇到分布外（OOD）物体，例如训练数据中未曾出现的意外障碍物。若缺乏适当的机制来识别此类异常，模型往往会将这些区域误分类为已知类别，从而产生不准确的分割掩码 [7]。这种误分类会削弱模型可靠性，更严重的是，会在自动驾驶[8]等安全关键领域带来重大安全风险。鉴于这一事实，对分布外异常的准确分割

对于在开放世界环境中构建稳健且可信的感知系统至关重要。

道路异常分割 [9], [10] 通过分割已知训练类别集合之外的道路物体来应对这一挑战。大多数现有方法仍遵循纯视觉范式 [11], [12], ，通常通过对像素级预测置信度进行阈值处理，或测量低级视觉特征的偏差来检测异常。

Z. He和X. Xue就职于复旦大学计算机科学学院，中国上海200433。J. Tang和J. Pu就职于复旦大学类脑智能科学与技术研究院，中国上海200433。*表示通讯作者。（电子邮箱：jianpu@fudan.edu.cn）。

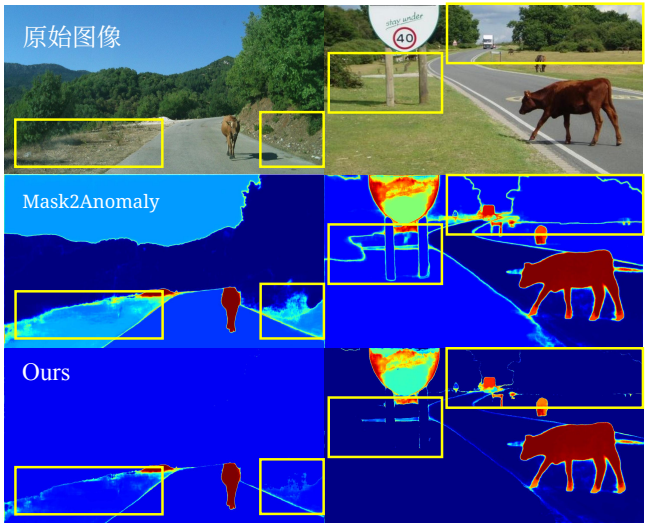

图1. 异常分数图对比。第一行展示原始图像，第二行呈现由Mask2Anomaly生成的异常分数图，第三行展示我们方法的结果。我们的方法通过抑制路面和植被等语义正常背景区域上的假阳性，得到更干净的图，同时更精确地突出动物等真实异常。

由于缺乏高级语义理解，这些模型受限于视觉相似性，其预测仅仅依据与特征原型或已知类别分布的相似程度。因此，天空和树木等背景区域在数据集中属于正常类别，但纹理或外观的变化（如云朵或颜色变化）可能导致它们被误检测为异常，如图 1 所示。这一现象导致假阳性率居高不下，并削弱了现有方法在真实世界自动驾驶场景中的鲁棒性和实用性。

近年来，视觉语言模型（视觉语言模型） [13], [14] 的出现为道路异常分割带来了新的视角。文本引导提供了显式语义，使模型能够识别与已知类别相对应的像素 [15]。与此同时，由视觉语言对齐得到的语义差异会自然地突出可能对应未知物体的区域 [16]。利用语义相似性不仅能抑制假阳性，还能强化稀有但信息丰富的分布外线索，从而提升开放世界场景中的检测精度与泛化能力。

基于视觉语言语义线索，我们提出了 VL-Anomaly，这是一种视觉-语言异常分割框架，可在训练和

推理过程中利用视觉语言模型的先验知识。其核心思想是将 CLIP [13] 中的类别级知识融入分割过程，从而增强模型区分分布内类别与分布外区域的能力。然而，直接将视觉语言模型应用于分割并非易事，因为它们并非专为像素级多类别预测而设计。核心挑战在于如何使视觉语言模型适应语义分割的多类别特性。尽管 MaskCLIP [17]等近期方法表明，可以在单个句子中匹配多个类别，但这些方法通常依赖手工设计或拼接的提示，且这些提示并未针对分割进行明确优化。为解决这一问题，我们设计了提示学习驱动对齐器（PL-Aligner）。该对齐器采用可学习的文本提示，引导骨干网络特征和解码器掩码查询进入视觉语言模型语义空间。对齐分两个阶段进行，即像素级和掩码级，二者共同弥合视觉表征与文本表征之间的差距。PL-Aligner 不依赖特定架构，可无缝集成到现有分割框架中，无需进行结构性修改。

在推理时，我们进一步采用多源推理策略来增强鲁棒性。具体而言，我们融合了以下信息：(i) 分割网络的检测器置信度，(ii) 从学习得到的提示中推导出的文本引导相似度，以及 (iii) 基于CLIP的图像-文本相似度，以发挥它们的互补优势。对这些互补信号进行融合能够提供可靠的异常预测，并减轻依赖单一信息源所带来的不足。通过这种多模态融合，VL-Anomaly 在多种道路异常数据集上实现了细粒度的异常分割，同时保持了较强的泛化能力。总结而言，我们作出以下贡献：

- 我们提出了 PL-Aligner，这是一种由提示驱动的对齐模块，在像素级和掩码级同时对齐特征，从而实现更鲁棒的文本引导异常分割。
- 我们提出了一种多源推理策略，融合文本引导相似度、基于 CLIP 的图文相似度和检测器置信度，以实现鲁棒的异常预测。
- 我们的方法在 RoadAnomaly [18], 、Fishyscapes [7]和 SMIYC[19],上均持续取得最先进的结果，展现出在多种数据集上的强大泛化能力。

II. 相关工作

A. 道路异常分割

道路异常分割旨在定位训练标签范围之外的意外障碍物，Fishyscapes [7]和 SMIYC [19]等基准测试均对此进行评估。

基于不确定性的方法通过从softmax/logit统计量或掩码级置信度推导异常分数来分割异常区域。由MSP [21], 扩展而来的MaxLogits [20], 已通过CAOS [22]基准应用于驾驶场景。为缓解置信度分数中的误校准，SML [23]对logit分布进行归一化并引入边界细化，以更好地定位异常。Mask2Anomaly [12]将异常分割重新表述为掩码分类任务，通过掩码级对比学习和细化策略实现了更高的一致性与召回率。尽管有这些改进，此类方法仍容易产生假阳性，尤其是在视觉上显著不同但语义上正常的区域，如树木或天空。

基于异常暴露的方法则利用辅助分布外样本，但其效果取决于所选异常数据集的多样性和相关性。最大化熵 [23]鼓励模型对代理分布外样本给出更高熵的预测，以防止模型在不确定区域过度自信；PEBAL [24] 则采用对比学习来区分分布外像素。类似地，RbA [11] 采用掩码分类范式，将被所有已知类别拒绝的区域定义为异常，从而在不损害分布内性能的情况下实现高质量的分布外分割。

除基于分数和异常暴露的方法外，生成式与混合方法还会对分布内特征分布建模，以获得基于似然的分布外分数。DenseHybrid[10] 进一步结合对分布内数据的生成建模，以及通过异常暴露对负样本进行判别式训练，以提升分布外分割效果。此外，基于图像再合成的方法通过重建差异凸显异常。例如，SynBoost [25]从语义图合成图像，并结合重建线索与基于不确定性的分数。

在本文中，我们提出 VL-Anomaly，该方法引入视觉语言先验以抑制假阳性。据我们所知，VL-Anomaly 是最早将多模态语义先验融入道路异常分割的方法之一，为自动驾驶中稳健的分布外感知提供了一种替代方案。

B. 用于语义分割的视觉语言模型

视觉语言模型（如 CLIP [13],）通过在共享语义空间中对齐图像和文本嵌入，展现了在开放世界任务中的出色泛化能力。近期研究将视觉语言模型融入语义分割，以增强类别扩展性和零样本能力。OpenSeg [26] 和 GroupViT [27] 等方法利用视觉语言模型，通过文本监督引导像素级或区域级预测，从而实现对未见类别的分割。MaskCLIP [17] 进一步将视觉语言模型特征与分割掩码对齐，以提升视觉定位能力。然而，这些工作主要聚焦于分布内领域中的开放词汇分割或零样本分割。近期，SimCLIP [28] 探究了工业异常检测中高层语言特征与细粒度视觉特征之间的不对齐问题，这促使我们利用视觉语言模型引导的语义先验来进行道路异常分割，以更好地捕捉未见物体。尽管大多数现有方法 [17], [26], [27] 仅在单一粒度上进行对齐——即像素级或掩码级对齐，CoupAlign [29] 则通过分层方式耦合二者，在概念上与我们的设计更为接近。

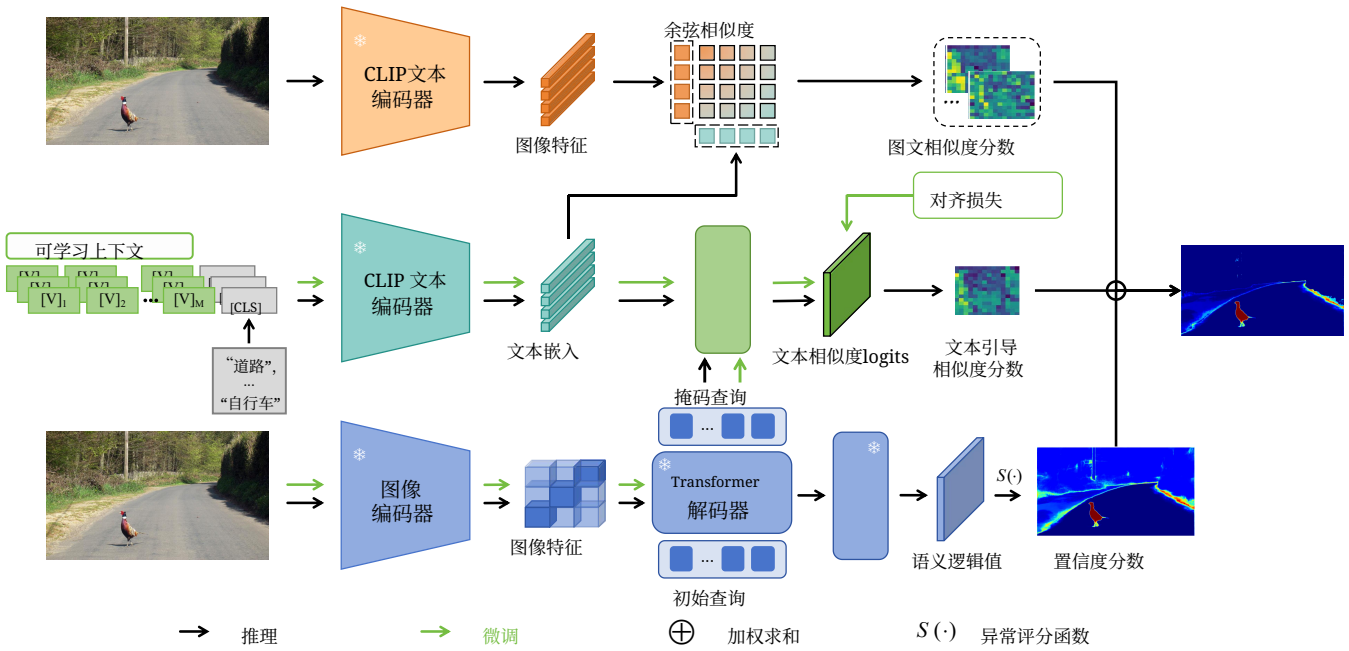

图 2. VL-Anomaly 的整体架构。该框架将分割骨干网络与基于 CLIP 的视觉-语言模块相结合。在训练期间，提示学习驱动对齐器（PL-Aligner）首先对骨干网络的视觉特征与 CLIP 的已知类别文本嵌入进行像素级对齐，随后进一步将解码器的掩码查询与之进行掩码级对齐。在推理期间，将分割模型输出、文本引导相似度和基于 CLIP 的图文相似度所产生的多源分数融合，以获得稳健的异常分割结果。

不同于使用文本提示作为分类器来扩展标签空间的开放词汇分割，我们的目标并非识别新类别，而是利用视觉语言模型的先验作为语义正则化，以抑制异常分割中的假阳性。

III. 方法

本节首先概述问题定义，然后结合异常分割回顾一种通用的掩码 Transformer 架构，最后介绍我们提出的框架及其独特组件，如图 2 所示。

A. 预备知识

形式化地，设 $\mathcal{X} \subset \mathbb{R}^{3\times H\times W}$ 表示 RGB 图像空间，其中 H 和 W 分别表示图像高度和宽度，$\mathcal{Y} \subset \mathbb{R}^{C_k\times H\times W}$ 表示语义标签空间，其中每个像素被赋予 C_k 个预定义已知类别之一。给定训练集 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{D}$，其中 $x_i \in \mathcal{X}$ 和 $y_i \in \mathcal{Y}$，异常分割任务旨在学习一个映射：

$$f : \mathcal{X} \to \mathbb{R}^{H\times W} \tag{1}$$

为每张输入图像生成异常分数图。

在逐像素架构和掩码 Transformer 架构中，$f(\cdot)$ 最终用于为每个像素或预测掩码赋值，然后将其与固定阈值进行比较，以区分分布外区域和分布内（ID）区域。尽管这类置信度框架在某些情况下有效，但它们本质上将分布外检测性能与模型学习到的分布内表征绑定在一起，因此在陌生场景中容易发生误分类。

为克服这一局限，我们将视觉语言模型先验作为语义知识的外部“观察者”引入。与训练好的分割模型受限于任务特定表征的“内部视角”不同，视觉语言模型能够提供丰富的开放世界语义，为在异常分割中区分分布内区域与分布外区域提供额外视角。以下各节将详细介绍我们方法的技术贡献。

B. 文本提示构建

与每次仅处理一个类别描述的典型视觉语言模型应用 [13] 不同，语义分割需要同时处理多个类别，因为单张图像通常包含多个语义类别。与人工编写的自然语言句子相比，采用可学习提示范式不仅可以避免歧义和冗余，还能通过联合训练自动适应分割任务，从而实现更稳健的跨模态对齐 [30]。

为了与所有已知类别并行对齐，我们为每个类别 $c_i \in \mathcal{C}$ 构建专用提示，其中 $\mathcal{C} = \{c_1, c2, \ldots, c_{C_k}\}$ 表示 C_k 个已知语义类别的集合（例如，Cityscapes [31] 中的 $C_k = 19$）。这一设计使模型能够在一次前向传播中计算类别级视觉语言相似度，从而实现高效的密集预测。遵循可学习提示范式 [30], , 每个提示均采用统一的上下文形式：

$$\mathbf{p}_i = [\mathbf{V}]_1 [\mathbf{V}]_2 \ldots [\mathbf{V}]_M [\mathrm{CLS}], \tag{2}$$

其中 $[\mathbf{V}]_i (i \in \{1, \ldots, M\})$ 表示一个可学习上下文令牌，其维度 d 与视觉语言模型词嵌入相同，

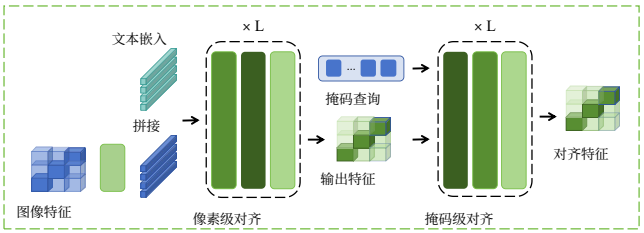

图 3. PL-Aligner 的架构。第一层将骨干网络提取的像素级视觉特征与文本嵌入对齐；第二层则将解码器中的掩码查询与第一层提取的像素级特征对齐，从而实现掩码级对齐。为清晰起见，省略了归一化和激活函数等标准操作。

M 是上下文词元的数量，[CLS] 是类别 c_i 的文本名称。上下文词元在各类别间共享相同形式，并与分割模型联合优化，以捕获任务特定语义，同时保留视觉语言模型的开放世界先验。

给定构造的提示$\mathbf{p}_i$，通过视觉语言模型的文本编码器获得特定类别的文本嵌入：

$$\mathbf{t}_i = \text{TextEncoder}(\mathbf{p}_i), \tag{3}$$

其中，$\mathbf{t}_i \in \mathbb{R}^d$ 表示类别c_i的文本嵌入。

C. 提示学习驱动对齐器

分割模型与视觉语言模型（VLM）的特征空间天然不对齐，因此我们提出了可学习的基于提示的对齐机制。现有基于 VLM 的分割方法通常采用单一粒度策略，例如像素级特征对齐 [15] 或基于区域特征的掩码级对齐 [17]。受 CoupAlign [29], 的启发——该方法展示了结合像素级与掩码级对齐的优势——我们提出 PL-Aligner，同时实现细粒度的像素一致性和结构化的掩码级语义对齐，见图 3。

a) 像素级对齐：对于输入图像，视觉编码器首先生成一个密集特征图，其中每个空间位置 i 都对应一个特征向量 $\mathbf{v}_i$。为使这些特征与文本提示对齐，图像特征图会被重新投影到与文本嵌入相同的嵌入维度，与提示特征拼接，并通过注意力机制和前馈网络进行联合处理。随后，使用像素级对比损失，将得到的像素对齐特征与相应的文本嵌入 $\mathbf{t}_{y_i}$对齐：

$$\mathcal{L}_{\text{pixel}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(\mathbf{v}_i,\mathbf{t}_{y_i})/\tau\right)}{\sum_{k=1}^{C_k}\exp\left(\text{sim}(\mathbf{v}_i,\mathbf{t}_k)/\tau\right)}, \tag{4}$$

其中，$\text{sim}(\cdot,\cdot)$表示余弦相似度，τ 是可学习的温度参数，y_i 是真实标签。此阶段直接对齐像素级视觉特征与文本语义。

b) 掩码级对齐：基于像素对齐特征，我们进一步在掩码变换器解码器之后引入掩码级对齐。具体而言，像素对齐特征用作键和值，而解码器掩码查询用作查询。通过注意力机制和前馈网络，解码器查询被细化为语义一致的表征，随后以相同的对比方式与文本嵌入对齐：

$$\mathcal{L}_{\text{mask}} = -\frac{1}{N_q}\sum_{m=1}^{N_q}\log\frac{\exp\left(\text{sim}(\mathbf{q}_m,\mathbf{t}_{y_m})/\tau\right)}{\sum_{k=1}^{C_k}\exp\left(\text{sim}(\mathbf{q}_m,\mathbf{t}_k)/\tau\right)}, \tag{5}$$

其中，$\mathbf{q}_m$ 表示投影后的掩码查询，y_m 是掩码m的真实类别标签。此阶段确保掩码级表征不仅与像素级对齐特征保持一致，还与视觉语言模型的语义空间对齐。

c) 对齐损失：最终的对齐损失结合了分割监督与两个对齐目标：

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{pixel}}\mathcal{L}_{\text{pixel}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}, \tag{6}$$

其中，$\mathcal{L}_{\text{seg}}$ 表示遵循 Mask2Former [12],的标准分割损失；在我们的微调阶段，λ_{pixel} 和 λ_{mask} 均固定为 0.5。

D. 多源推理策略

基于上一节中 PL-Aligner 在训练期间引入文本感知能力的做法，我们进一步在推理时引入来自 CLIP 的外部语义先验，以便更好地区分分布内区域与分布外区域。为此，我们提出一种多源推理策略，融合三种互补分数：(i) 检测器置信度、(ii) 文本引导的相似度，以及 (iii) 基于CLIP的图像-文本相似度。该策略缓解了依赖单一信息源的不足。

a) 检测器置信度：受 [12]，掩码 Transformer 架构的启发，解码器为每个掩码索引m输出类别分数$\mathbf{c}_m \in \mathbb{R}^{C_k}$ 和掩码逻辑值$\mathbf{m}_m \in \mathbb{R}^{H\times W}$。类别$k$ 的检测器置信度图计算如下：

$$S_{\text{conf}} = \max_{k\in\{1,\ldots,C_k\}} \text{softmax}(\mathbf{c}_m)_k \cdot \sigma(\mathbf{m}_m), \tag{7}$$

其中，$\sigma(\cdot)$ 表示逐元素的 Sigmoid 函数。

b) 文本引导相似度：我们重复使用训练期间获得的学习得到的提示嵌入$\mathbf{t}_k$ 。将 PL-Aligner 输出的最终对齐特征$\mathbf{v}^{align}$ 与$\mathbf{t}_k$进行比较，以计算文本引导相似度分数：

$$S_{\text{text}}^{(k)} = \text{sim}(\mathbf{v}_{align}, \mathbf{t}_k). \tag{8}$$

该分数用于量化对齐后的视觉表示与每个学习得到的类别特定提示嵌入之间的语义一致性。

c) 基于 *CLIP* 的图文相似度：我们还使用冻结的 CLIP 图像编码器，计算输入图像x 与每个类别提示$\mathbf{t}_k$ 之间的图像级相似度分数：

$$S_{\text{img}}^{(k)} = \text{sim}(\text{ImageEncoder}(x), \mathbf{t}_k), \tag{9}$$

它提供了独立于分割模型预测结果的全局语义先验。

表 I 在 RoadAnomaly [18], SMIYC-RA21 [19] 和 SMIYC-RO21 [19]上的评估。

方法	RoadAnomaly			SMIYC-RA21			SMIYC-RO21			平均值		
	$FPR_{95}\downarrow$	AuPRC↑	AuROC↑	$FPR_{95}\downarrow$	AuPRC↑	AuROC↑	$FPR_{95}\downarrow$	AuPRC↑	AuROC↑	$FPR_{95}\downarrow$	AuPRC↑	AuROC↑
MSP [21]	71.4	15.7	-	72.0	28.0	-	16.6	15.7	-	53.3	19.8	-
熵 [21]	68.2	15.7	-	-	-	-	-	-	-	68.2	15.7	-
马哈拉诺比斯 [32]	81.1	14.4	-	87.0	20.0	-	13.1	20.1	-	60.4	18.2	-
SML [23]	70.7	17.5	-	39.5	46.8	-	36.8	3.4	-	49.0	22.6	-
JSRNet [33]	**9.2**	**94.4**	-	43.9	33.6	-	28.9	28.1	-	27.3	52.0	-
SynBoost [9]	64.8	38.2	-	61.9	56.4	-	3.2	71.3	-	43.3	55.3	-
最大熵 [34]	31.8	48.9	-	15.0	85.5	-	0.8	85.1	-	15.9	73.2	-
密集混合 [10]	64.0	31.4	-	9.8	78.0	-	**0.2**	87.1	-	24.7	65.5	-
PEBEL [24]	44.6	45.1	-	40.8	49.1	-	12.7	5.0	-	32.7	33.1	-
ODIN† [35]	32.9	55.0	91.4	3.6	91.5	98.5	1.1	47.8	99.2	12.5	64.8	96.4
Mask2Anomaly† [12]	13.2	79.7	96.2	14.6	88.7	**99.7**	**0.2**	**93.3**	99.1	9.3	87.2	98.3
VL-Anomaly（我们的方法）	12.9	79.2	**96.8**	**3.5**	**95.1**	**99.7**	0.6	91.0	**99.7**	**5.7**	**88.4**	**98.7**

注意：↑表示越高越好，↓表示越低越好。最佳和次佳结果分别以**粗体**和下划线标示。

d) 分数融合：将这三个分数合并为一个统一的异常分数：

$$S_{\text{final}} = 1 - \max_{k\in\{1,\dots,C_k\}} \left(\alpha \cdot S^{(k)}_{\text{conf}} + \beta \cdot S^{(k)}_{\text{text}} + \gamma \cdot S^{(k)}_{\text{img}}\right), \quad (10)$$

其中 α 、β、 γ 分别设为0.7、0.2和0.1。较高的S_{final} 表示属于分布外区域的可能性更大。

IV. 实验

在本节中，我们首先介绍实验设置，包括对异常推理数据集、评估指标和实现细节的介绍。然后，我们将我们的方法与其他出色的基线进行比较。此外，我们对VL-Anomaly模块以及PL-Aligner的层设计进行了广泛的消融实验。

A. 设置

数据集。异常推理模型首先在Cityscapes [31] 上微调（2,975张训练图像、500张验证图像和1,525张测试图像），然后按照 [12] 使用由MS-COCO [36]生成的离群辅助数据集进行微调。我们在三个基准测试上进行评估：RoadAnomaly [18]、Segment Me If You Can（SMIYC） [19] 和Fishyscapes [7]。RoadAnomaly包含60张来自互联网的验证图像，其中的异常具有不同尺度。SMIYC包含两个子集：RoadAnomaly21（RA21，10张验证图像和100张测试图像）和RoadObstacle21（RO21，30张验证图像和327张测试图像）。Fishyscapes有两个子集：Static（30张验证图像和1,000张测试图像）以及Lost & Found（L&F，100张验证图像和275张测试图像）。在推理过程中，对于RoadAnomaly、RA21、RO21和Fishyscapes，仅可获取验证集真实标注；测试集真实标注仍然不可用。

评估指标。遵循 [12], [24], [37],，我们在像素级评估中报告受试者工作特征曲线下面积（AuROC）、95%真正率下的假阳性率（FPR_{95}）以及精确率-召回率曲线下面积（AuPRC）。然而，像素级指标可能会忽略较小的异常，并偏向较大的异常区域。因此，我们还采用组件级指标，包括平均逐组件 F1 分数（$F1^{*}$）、阳性预测值（PPV）和逐组件交并比（sIoU）。

基线方法。除了与先前工作 [9], [10], [21], [23], [24], [32]–[34], 报告的结果进行比较外，我们还进一步复现了 ODIN † [35] 和 Mask2Anomaly† [12]，以进行更公平、全面的比较。对于 ODIN[35], ，温度参数设为 $T = 3.0$，遵循 SMIYC 基准[19] 代码库提供的实现。对于 Mask2Anomaly [12], ，训练期间不使用细化掩码阶段。前者是基于置信度的分布外检测领域的经典工作，后者则是我们在掩码级异常分割方面的主要基线，因此纳入二者有助于体现比较的合理性。值得注意的是，所有复现的基线方法均未对 Mask2Former 模型架构作任何修改，从而确保了比较的公平性。**实现细节**。我们的掩码分割网络基于 Mask2Anomaly [12],，后者以带有 ResNet-50 [38]骨干网络的 Mask2Former [5] 为基础。微调期间，分割网络和 CLIP ViT-L-16 的权重均被冻结。默认情况下，我们采用 L = 1层像素级和掩码级注意力。对于提示学习，我们为每个类别采用 $M = 16$ 个可学习上下文词元，每个词元的维度为 $d = 512$，并类似于 MaskCLIP [17]，利用 CLIP 的词嵌入从预定义模板初始化。分割损失 $\mathcal{L}_{\text{seg}}$ 遵循 Mask2Former 的标准配置，由分类损失、Dice [39]损失和掩码损失加权求和得到，权重系数分别为 1。0，1。0，20。0。对比损失中的温度参数 τ 被视为可学习标量，并遵循对比学习 [13]中的常见做法，将其初始化为 0.07。优化时，使用 AdamW [40] 对模型进行微调，初始学习率为 1e−4，权重衰减为 0.05，批量大小为 8，共训练 5,000 次迭代。训练图像会随机裁剪至 380×760，并从 0.1 到 2.0 之间随机采样缩放比例。验证期间，我们使用真实分布外掩码，分别计算分布内像素和分布外像素的异常分数，与先前工作 [12]保持一致。所有实验均在单张 NVIDIA 3090 GPU 上进行。

B. 评估结果

1) RoadAnomaly 和 *SMIYC-RA21/RO21* 上的结果：表 I 总结了 RoadAnomaly 以及

表 II Fishyscapes静态和FISHYSCAPES LOST & FOUND 验证数据集上的评估。

方法	FS 静态			FS L&F		
	FPR_{95}↓	AuPRC↑	AuROC↑	FPR_{95}↓	AuPRC↑	AuROC↑
MSP [21]	39.8	12.9	-	44.8	1.3	-
熵 [21]	39.7	15.4	-	44.8	2.9	-
SML [23]	20.5	52.1	-	21.9	31.7	-
SynBoost [9]	15.8	43.2	-	18.8	**72.6**	-
最大熵 [34]	8.6	86.5	-	35.1	29.9	-
密集混合 [10]	5.9	80.2	-	3.9	47.1	-
PEBEL [24]	**1.7**	**92.4**	-	7.6	44.2	-
ODIN† [35]	62.4	82.3	89.4	81.1	1.4	86.3
Mask2Anomaly† [12]	1.9	90.4	98.3	**4.4**	46.0	93.6
VL-Anomaly（我们的方法）	2.3	90.7	**98.7**	8.4	69.5	**96.0**

注：↑表示越高越好，↓表示越低越好。最佳和次佳结果分别以**粗体**和下划线标示。

表III SMIYC [19] 基准测试结果。

方法	SMIYC-RA21			SMIYC-RO21		
	sIoU↑	PPV↑	$F1^*$↑	sIoU↑	PPV↑	$F1^*$↑
MSP [21]	15.5	15.3	5.4	19.7	15.9	6.3
马哈拉诺比斯 [32]	14.8	10.2	2.7	13.5	21.8	4.7
SML [23]	26.0	24.7	12.2	5.1	13.3	3.0
JSRNet [33]	20.2	29.3	13.7	18.6	24.5	11.0
Mask2Former [5]	25.2	18.2	15.3	5.0	21.9	4.8
最大熵 [34]	49.2	39.5	28.7	47.9	62.6	48.5
SynBoost [9]	34.7	17.8	9.9	44.3	41.7	37.5
密集混合 [10]	54.1	24.1	31.0	45.7	50.1	50.7
PEBEL [24]	38.8	27.2	14.4	29.9	7.5	5.5
Rba [11]	55.7	**52.1**	46.8	58.4	58.8	60.9
Maskomaly [16]	55.4	51.5	**49.9**	-	-	-
Mask2Anomaly† [12]	**60.4**	45.7	48.6	**61.4**	70.3	69.8
VL-Anomaly（我们的方法）	59.6	50.1	48.7	61.2	**73.4**	**70.1**

注：↑表示数值越高越好。最佳和次佳结果分别以**粗体**和下划线标示。

SMIYC 基准的两个子集 RA21 和 RO21 上的结果。在采用相同 Mask2Anomaly 骨干网络的情况下，VL-Anomaly 的表现始终优于现有方法。在 RoadAnomaly 上，其 AuROC 达到 96.8，比 Mask2Anomaly 高 +0.6，且将 FPR_{95} 从 13.2 降至 12.9。在 RA21 上，与 Mask2Anomaly 相比，VL-Anomaly 的 AuPRC 提高了 +6.4。在 RO21 上，其 AuROC 达到最高值 99.7，比 Mask2Anomaly 高 +0.6。在这三个数据集上的平均表现中，VL-Anomaly 整体性能最佳。这与我们的核心理念相符，因为 FPR_{95} 的显著降低以及 AuPRC 的提升，主要源于语义上属于正常场景的背景区域中的假阳性减少。

*2）Fishyscapes*上的结果：Fishyscapes基准上的结果见表II。在Static子集上，VL-Anomaly的AuROC达到98.7，比Mask2Anomaly高出 +0.4，同时保持具有竞争力的FPR_{95} 为2.3。在更具挑战性的Lost and Found子集上，VL-Anomaly表现出明显提升，将AuPRC从46.0提高到69.5（+23.5），并将AuROC从93.6提高到96.0（+2.4）。这些结果证实，VL-Anomaly不仅在较简单的设置中表现优异，也能很好地泛化到复杂场景。

表 IV 管线设计的消融实验。

设置	FPR_{95} ↓	AuPRC ↑	AuROC ↑	每秒帧数 ↑
基线	4.4	46.0	93.6	**8.3**
+ 多层感知机 & 手工设计提示	13.4	60.1	95.1	8.3
→ PL-Aligner 和 S_{text}	9.2	64.0	95.1	7.9
→ 可学习提示 [30]	9.2	**69.7**	95.4	7.9
+S^{img} (完整设计)	**8.4**	69.5	**96.0**	6.7

注：以基线为起点，我们逐步加入提示、PL-Aligner 和多源推理组件（S_{text}：文本引导相似度，S^{img}：基于 CLIP 的图文相似度）。↑表示越高越好，↓表示越低越好。

表 V PL-ALIGNER 层消融实验。

对齐	对齐器	FPR_{95} ↓	AuPRC ↑	AuROC ↑
仅像素	MLP	13.4	60.1	93.6
仅像素	交叉注意力 [26]	9.2	69.4	95.4
仅像素	PL-Aligner	9.1	69.4	95.6
仅掩码	PL-Aligner	16.7	55.4	78.6
像素+掩码	PL-Aligner	**8.4**	**69.5**	**96.0**

注意：我们比较了仅进行像素级对齐、仅进行掩码级对齐，以及同时使用两个层级（像素+掩码）的效果。↑表示越高越好，↓表示越低越好。

*3）SMIYC*上的结果：表III给出了SMIYC 基准中RA21和RO21子集上的结果。MSP、马哈拉诺比斯和SML等传统的基于置信度或基于距离的方法表现较差，在RA21上的$F1^*$ 分数低于15，在RO21上低于12，显示出它们在复杂驾驶场景中的局限性。Rba和Mask2Anomaly等基于掩码的分割框架取得了显著改进，其中Mask2Anomaly在RA21上的sIoU达到60.4，在RO21上的$F1^*$ 达到69.8。在这一强基线之上，VL-Anomaly取得了进一步提升。在RA21上，它取得了最佳的$F1^*$ ，为48.7。在RO21上，它将PPV从70.3提高到73.4（+3.1），并将$F1^*$ 从69.8提高到70.1（+0.3），同时保持相当的sIoU为61.2。总体而言，VL-Anomaly在RA21和RO21上均取得了最佳或接近最佳的性能。

C. 消融研究

本节报告的所有消融研究结果均来自 Fishyscapes Lost and Found 验证数据集 [7]。

1) 管线设计： 我们在表 IV 中进行了消融研究，以评估各模块的贡献。从基线开始，加入基于多层感知机的像素级对齐，并结合S_{text} ，可提升检测性能，尽管FPR95有所增加。将多层感知机替换为我们的 PL-对齐器后，结果进一步提升，且 FPR95有所降低。引入提示并从手工设计转为可学习提示，带来了持续的增益，AuPRC 达到 69.7，同时保持每秒 7.9 帧。最后，将S^{img} 纳入多源推理后，整体性能达到最佳，为每秒 6.7 帧；与基线相比，每秒帧数仅适度下降 1.6，表明我们的创新仅带来极小的效率成本。

2) PL-Aligner 层设计：表 V 展示了针对不同 PL-Aligner 配置的消融研究。仅在像素级进行对齐时，简单的多层感知机投影只能带来有限的改进，而加入交叉

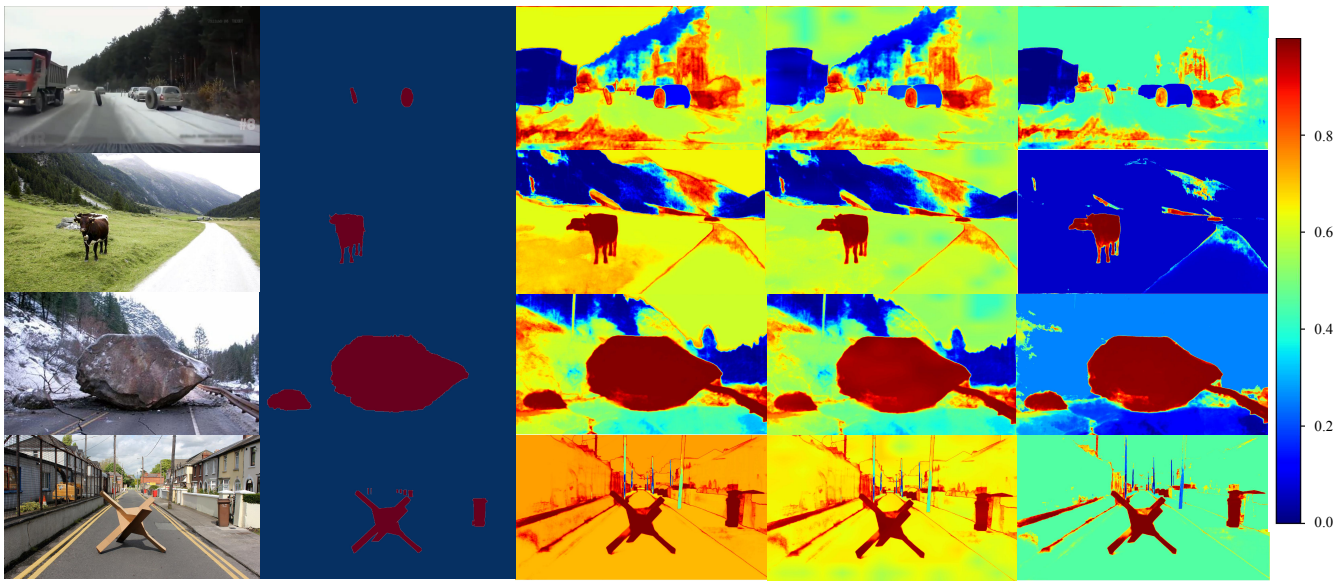

图 4. Road Anomaly 数据集 [18]上的异常分割结果定性比较。为进行公平比较，我们将本方法预测的异常分数图与使用相同骨干网络的 MSP [21] 和 Mask2Anomaly [12], 生成的结果进行比较。为便于可视化，所有分数均归一化至相同范围。本方法能更有效地抑制语义上正常的背景区域中的假阳性，而其他方法往往会在这些区域产生模糊或虚假的激活。



图 5. 图像特征与构建的文本提示之间相似度的可视化。高亮区域显示了模型将图像区域与特定语义类别相关联的位置，体现了我们的提示学习策略在引导跨模态对齐方面的有效性。

注意力则显著降低了 FPR95 并提升了 AuPRC，证实了对视觉特征与文本特征之间的交互进行建模的必要性。我们的像素级 PL-Aligner 取得了与交叉注意力相当的增益，但 AuROC 略高，表明它是一种有效的轻量级替代方案。相比之下，仅在掩码级进行对齐的效果较差，这说明单靠高层查询不足以捕捉细粒度语义线索。融合像素级对齐和掩码级对齐的完整设计取得了最佳综合结果，表明这两个层级提供了互补信息：像素级对齐锚定细粒度语义，而掩码级对齐则强化结构化类别的一致性。

D. 定性结果

图4展示了ODIN、Mask2Anomaly和VL-Anomaly的异常图。我们的方法有效抑制了树木和植被等语义正常区域中的假阳性，生成更干净的热力图，并减少了虚假激活。图5进一步说明我们的提示学习策略的作用。高亮区域对应于与文本提示对齐的正常背景类别，而异常物体则表现为低强度区域。这表明基于提示的跨模态对齐引入了额外的语义先验，使模型能够更好地将意外模式与背景区分开来。

V. 结论

在本研究中，我们提出了 VL-Anomaly，这是一种视觉语言引导的异常分割框架，通过利用语义先验来增强道路异常的分割效果。通过引入 PL-Aligner，我们将视觉特征与 CLIP 文本嵌入显式对齐，从而有效抑制语义正常区域中的假阳性。此外，我们的多源推理策略融合了检测器置信度、文本引导相似度和基于 CLIP 的图像相似度，从而生成更可靠的异常预测。在 RoadAnomaly、SMIYC 和 Fishyscapes 基准测试上开展的大量实验表明，与最先进的基线方法相比，我们的方法具有优越性和良好的泛化能力。尽管结果令人鼓舞，我们的多源推理策略仍受限于基于多个数据集手动调整的权重，这可能会限制其可扩展性和自动化程度。自适应或数据驱动的权重学习将是更为合理的替代方案。未来的工作将着重开发自动权重优化策略，以进一步提升该方法在多种应用场景中的鲁棒性和实用性。

参考文献

[1] J. Chen、J. Lu、X. Zhu 和 L. Zhang，《生成式语义分割》，载于 IEEE/CVF 计算机视觉与模式识别会议（ CVPR ）论文集，2023 年 6 月，第 7111–7120 页。[2] Z. He、X. Li、J. Tang、S. Qiu、W. Wang、X. Xue 和 J. Pu，《面向自动驾驶场景的摄像头开放集三维目标检测》，IEEE智能交通系统汇刊，第 26 卷，第 12 期，第 23 190–23 201 页，2025年。[3] Y. Xu、J. Cui、F. Cai、Z. Zhu、H. Shang、S. Luan、M. Xu、N. Zhang、Y. Li、J. Cai 等，《WAM-Flow：通过离散流匹配实现自动驾驶中的并行由粗到细运动规划》，arXiv 预印本 arXiv:2512.06112，2025年。[4] Y. Liao、S. Kang、J. Li、Y. Liu、Y. Liu、Z. Dong、B. Yang 和 X. Chen，《Mobile-Seed：面向移动机器人的语义分割与边界检测联合方法》，IEEE机器人与自动化快报，2024年。[5] B. Cheng、I. Misra、A. G. Schwing、A. Kirillov 和 R. Girdhar，《用于通用图像分割的掩码注意力掩码变换器》，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。[6] D. Bo、W. Pichao 和 F. Wang，《Afformer：基于线性变换器的无头轻量级语义分割》，载于 AAAI人工智能会议论文集，2023年。[7] H. Blum、P.-E. Sarlin、J. Nieto、R. Siegwart 和 C. Cadena，《Fishyscapes基准：测量语义分割中的盲区》，国际计算机视觉期刊，第 129 卷，第 11 期，第 3119–3135 页，2021年。[8] D. Bozhinoski 、 D. Di Ruscio 、 I. Malavolta 、 P. Pelliccione 和 I. Crnkovic，《移动机器人系统安全：从软件工程视角开展的系统性映射研究》，《系统与软件杂志》，第 151 卷，第 150–179 页，2019年。[9] G. Di Biase、H. Blum、R. Siegwart 和 C. Cadena，《复杂驾驶场景中的像素级异常检测》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第 16 918–16 927 页。[10] M. Grci´c、P. Bevandi´c 和 S. ˇ Segvi´c，《DenseHybrid：用于密集开放集识别的混合异常检测》，载于欧洲计算机视觉会议。施普林格，2022年，第 500–517 页。[11] N. Nayal、M. Yavuz、J. F. Henriques 和 F. G¨uney，《Rba：分割被所有方法拒绝的未知区域》，载于 IEEE/CVF国际计算机视觉会议论文集，2023年，第 711–722 页。[12] S. N. Rai、F. Cermelli、D. Fontanel、C. Masone 和 B. Caputo，《揭示道路场景分割中的异常》，载于 IEEE/CVF国际计算机视觉会议论文集，2023年，第 4037–4046 页。[13] 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。[14] J. Li、D. Li、C. Xiong 和 S. Hoi，《BLIP：通过自举式语言-图像预训练实现统一的视觉语言理解与生成》，载于国际机器学习会议。PMLR，2022年，第 12 888–12 900 页。[15] C. Jia、Y. Yang、Y. Xia、Y.-T. Chen、Z. Parekh、H. Pham、Q. Le、Y.-H. Sung、Z. Li 和 T. Duerig，《利用噪声文本监督扩展视觉与视觉语言表征学习》，载于国际机器学习会议。PMLR，2021年，第 4904–4916 页。[16] J. Ackermann、C. Sakaridis 和 F. Yu，《Maskomaly：零样本掩码异常分割》，载于英国机器视觉会议（BMVC），2023年。[17] X. Dong、J. Bao、Y. Zheng、T. Zhang、D. Chen、H. Yang、M. Zeng、W. Zhang、L. Yuan、D. Chen 等，《MaskCLIP：掩码自蒸馏促进对比式语言-图像预训练》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2023年，第 10 995–11 005 页。[18] K. Lis、K. Nakka、P. Fua 和 M. Salzmann，《通过图像再合成检测意外情况》，载于 IEEE/CVF国际计算机视觉会议论文集，2019年，第 2152–2161 页。[19] R. Chan、K. Lis、S. Uhlemeyer、H. Blum、S. Honari、R. Siegwart、P. Fua、M. Salzmann 和 M. Rottmann，《SegmentMeIfYouCan：异常分割基准》，载于第三十五届神经信息处理系统会议数据集与基准赛道（第二轮），2021年。[20] D. Hendrycks、S. Basart、M. Mazeika、A. Zou、J. Kwon、M. Mostajabi、J. Steinhardt 和 D. Song，《扩展分布外检测的载于 IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第 16 918–16 927 页。[10] M. Grci´c、P. Bevandi´c 和 S. ˇ Segvi´c，《DenseHybrid：用于密集开放集识别的混合异常检测》，载于欧洲计算机视觉会议。施普林格，2022年，第 500–517 页。[11] N. Nayal、M. Yavuz、J. F. Henriques 和 F. G¨uney，《Rba：分割被所有方法拒绝的未知区域》，载于 IEEE/CVF国际计算机视觉会议论文集，2023年，第 711–722 页。[12] S. N. Rai、F. Cermelli、D. Fontanel、C. Masone 和 B. Caputo，《揭示道路场景分割中的异常》，载于 IEEE/CVF国际计算机视觉会议论文集，2023年，第 4037–4046 页。[13] 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。[14] J. Li、D. Li、C. Xiong 和 S. Hoi，《BLIP：通过自举式语言-图像预训练实现统一的视觉语言理解与生成》，载于国际机器学习会议。PMLR，2022年，第 12 888–12 900 页。[15] C. Jia、Y. Yang、Y. Xia、Y.-T. Chen、Z. Parekh、H. Pham、Q. Le、Y.-H. Sung、Z. Li 和 T. Duerig，《利用噪声文本监督扩展视觉与视觉语言表征学习》，载于国际机器学习会议。PMLR，2021年，第 4904–4916 页。[16] J. Ackermann、C. Sakaridis 和 F. Yu，《Maskomaly：零样本掩码异常分割》，载于英国机器视觉会议（BMVC），2023年。[17] X. Dong、J. Bao、Y. Zheng、T. Zhang、D. Chen、H. Yang、M. Zeng、W. Zhang、L. Yuan、D. Chen 等，《MaskCLIP：掩码自蒸馏促进对比式语言-图像预训练》，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2023年，第 10 995–11 005 页。[18] K. Lis、K. Nakka、P. Fua 和 M. Salzmann，《通过图像再合成检测意外情况》，载于 IEEE/CVF国际计算机视觉会议论文集，2019年，第 2152–2161 页。[19] R. Chan、K. Lis、S. Uhlemeyer、H. Blum、S. Honari、R. Siegwart、P. Fua、M. Salzmann 和 M. Rottmann，《SegmentMeIfYouCan：异常分割基准》，载于第三十五届神经信息处理系统会议数据集与基准赛道（第二轮），2021年。[20] D. Hendrycks、S. Basart、M. Mazeika、A. Zou、J. Kwon、M. Mostajabi、J. Steinhardt 和 D. Song，《扩展分布外检测的载于世界场景”，国际机器学习大会，2022年。[23] S. Jung、J. Lee、D. Gwak、S. Choi与J. Choo，“标准化最大逻辑值：一种在城市场景分割中识别意外道路障碍物的简单而有效方法”，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第15 425–15 434页。[24] Y. Tian、Y. Liu、G. Pang、F. Liu、Y. Chen与G. Carneiro，“面向复杂城市驾驶场景异常分割的像素级能量偏置弃权学习”，载于《欧洲计算机视觉会议》。施普林格，2022年，第246–263页。[25] G. Di Biase、H. Blum、R. Siegwart与C. Cadena，“复杂驾驶场景中的像素级异常检测”，载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，2021年6月，第16 918–16 927页。[26] G. Ghiasi、X. Gu、Y. Cui与T.-Y. Lin，“利用图像级标签扩展开放词汇图像分割”，载于《欧洲计算机视觉会议》。施普林格，2022年，第540–557页。[27] J. Xu、S. De Mello、S. Liu、W. Byeon、T. Breuel、J. Kautz与X. Wang，“GroupViT：语义分割从文本监督中涌现”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第18 134–18 144页。[28] C. Deng、H. Xu、X. Chen、H. Xu、X. Tu、X. Ding与Y. Huang，“SimCLIP：利用简单提示优化图文对齐，用于零样本/少样本异常检测”，载于《第32届ACM国际多媒体大会论文集》，2024年，第1761–1770页。[29] Z. Zhang、Y. Zhu、J. Liu、X. Liang与W. Ke，“CoupAlign：将词-像素对齐与句子-掩码对齐相结合用于指代图像分割”，《神经信息处理系统进展》，第35卷，第14 729–14 742页，2022年。[30] K. Zhou、J. Yang、C. C. Loy与Z. Liu，“学习为视觉语言模型生成提示”，《国际计算机视觉期刊》（国际计算机视觉期刊），2022年。[31] M. Cordts、M. Omran、S. Ramos、T. Rehfeld、M. Enzweiler、R. Benenson、U. Franke、S. Roth与B. Schiele，“用于语义城市场景理解的Cityscapes数据集”，载于《IEEE计算机视觉与模式识别会议论文集》，2016年，第3213–3223页。[32] K. Lee、K. Lee、H. Lee与J. Shin，“一种用于检测分布外样本与对抗攻击的简单统一框架”，《神经信息处理系统进展》，第31卷，2018年。[33] T. Vojir、T. ˇSipka、R. Aljundi、N. Chumerin、D. O. Reino与J. Matas，“通过与分割耦合的部分图像重建进行道路异常检测”，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第15 651–15 660页。[34] R. Chan、M. Rottmann与H. Gottschalk，“面向语义分割中分布外检测的熵最大化与元分类”，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第5128–5137页。[35] S. Liang、Y. Li与R. Srikant，“提升神经网络中分布外图像检测的可靠性”，载于《国际学习表征大会》，2018年。[36] T.-Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan、P. Doll´ar与C. L. Zitnick，“Microsoft COCO：上下文中的常见物体”，载于《欧洲计算机视觉会议》。施普林格，2014年，第740–755页。[37] Y. Liu、C. Ding、Y. Tian、G. Pang、V. Belagiannis、I. Reid与G. Carneiro，“面向语义分割中像素级分布外检测的残差模式学习”，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第1151–1161页。[38] K. He、X. Zhang、S. Ren与J. Sun，“面向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。[39] F. Milletari、N. Navab与S.-A. Ahmadi，“V-Net：用于体积医学图像分割的全卷积神经网络”，载于2016年第四届三维视觉国际会议（3DV）。电气电子工程师学会，2016年。[40] I. Loshchilov与F. Hutter，“解耦权重衰减正则化”，arXiv预印本 arXiv:1711.05101，2017年。