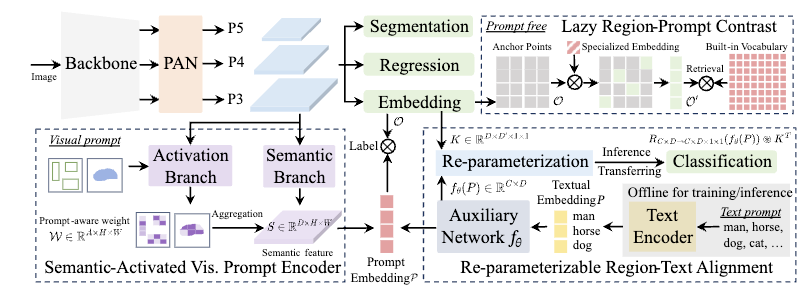
YOLOE：实时观察一切

一、简介

YOLOE（实时感知一切）专为开放词汇检测和分割而设计。与之前局限于固定类别的 YOLO 模型不同，YOLOE 使用文本、图像或内部词汇提示，能够实时检测任何目标类别。YOLOE 基于 YOLOv10 和YOLO-World，速度快、准确率高。

1. 在 LVIS 数据集上
   1. YOLOE 的 AP 值比 YOLO-Worldv2 提高了 3.5 倍
   2. 仅使用了三分之一的训练资源，推理速度提高了 1.4 倍。
2. YOLOE-v8-large比 YOLOv8-L高0.1mAP，训练时间减少了4 倍。
3. YOLOE原理介绍



YOLOE 保留了标准的 YOLO 结构，骨干网络CSP-Darknet、PAN-FPN，以及无锚点、解耦的检测头。

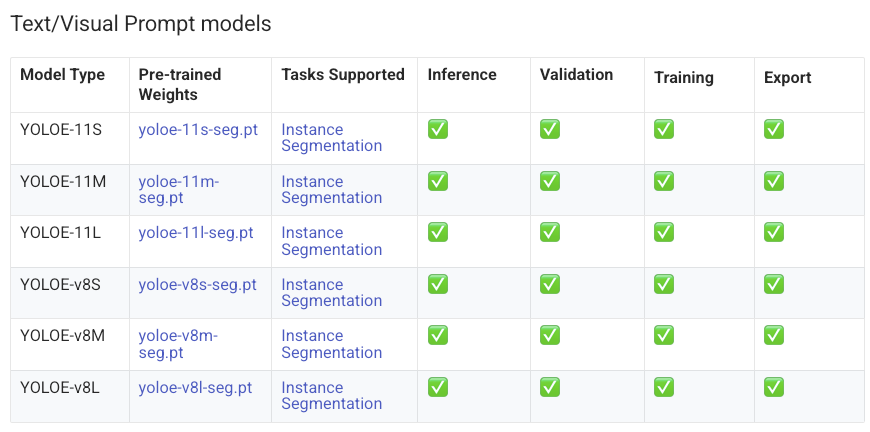
1. YOLOE 引入了三个用于开放词汇检测的全新模块：

* Re-parameterizable Region-Text Alignment ([RepRTA](https://zhida.zhihu.com/search?content_id=255923720&content_type=Article&match_order=1&q=RepRTA&zhida_source=entity" \t "/Users/ucasdragon/Documents\\x/_blank))：针对文本提示，YOLOE提出了RepRTA策略，通过一个可重新参数化的轻量级辅助网络来优化预训练的文本嵌入，从而提高视觉-文本对齐的准确性。在训练期间，辅助网络处理文本提示，而在推理和迁移时，辅助网络可以无缝地重新参数化到分类头部，从而实现零推理开销。
* Semantic-Activated Visual Prompt Encoder (SAVPE)：针对视觉提示，YOLOE设计了SAVPE，它通过语义分支和激活分支来高效地处理视觉线索。语义分支提取与提示无关的语义特征，而激活分支则通过与图像特征的交互生成分组的提示感知权重。这种设计在保持性能的同时，将计算复杂度降至最低。
* Lazy Region-Prompt Contrast (LRPC)：在无提示场景下，YOLOE引入了LRPC策略，它利用内置的大型词汇表和专用嵌入来识别所有物体，避免了依赖昂贵的语言模型。通过仅匹配锚点与已识别物体的类别名称，LRPC确保了高性能的同时降低了计算开销。

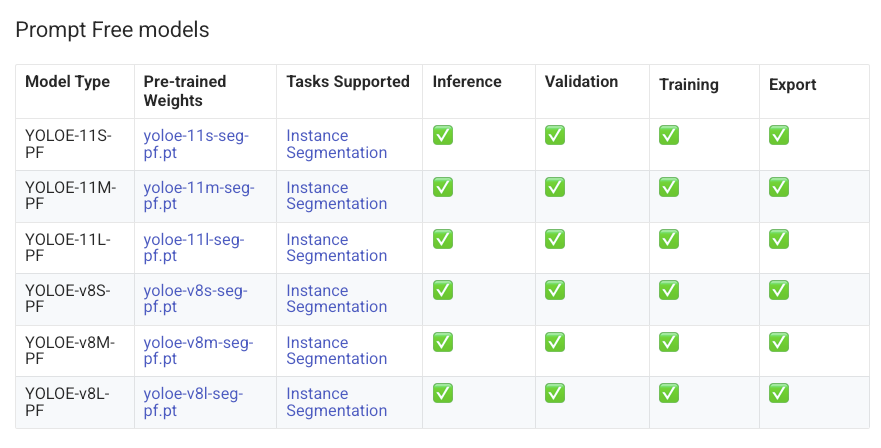
2、此外：

* YOLOE 通过在检测头中添加一个掩码预测分支（类似于 YOLACT 或 YOLOv8-Seg）来集成实时实例分割，从而最大限度地减少了开销。
* YOLOE 的开放世界模块在用作常规闭集 YOLO 时不会产生推理成本。训练后，YOLOE 参数可以重新参数化为标准 YOLO 头，同时保持相同的 FLOP 和速度（例如，与 YOLO11 完全匹配）。

1. 开源的预训练权重、支持的任务和操作模式
2. 文本提示、视觉提示的预训练权重。
   1. 主要是基于YOLO v11和YOLO v8，以及它们三个的不同参数量版本Small、Medium、Large。
   2. 文本提示、视觉提示都支持实例分割任务。
   3. ultralytics支持：训练、推理、验证、导出onnx等流程。



1. prompt free的预训练权重
   1. 主要是基于YOLO v11和YOLO v8，以及它们三个的不同参数量版本Small、Medium、Large。
   2. 支持实例分割任务。
   3. ultralytics支持：训练、推理、验证、导出onnx等流程。

  
四、使用示例

Ultralytics 提供Python API

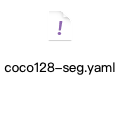
1. 训练的使用示例

在自定义的数据集上微调。

1. Fine-Tuning

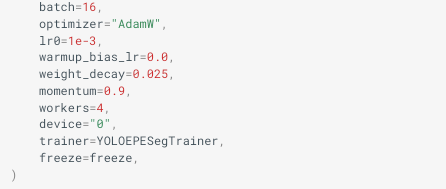


* 是一种 Full tuning方*式*，所有参数均可训练。
  + 查看了一下coco128-seg的数据格式：数据格式处理成COCO格式。
* coco128-seg YAML数据格式：



1. Linear Probing





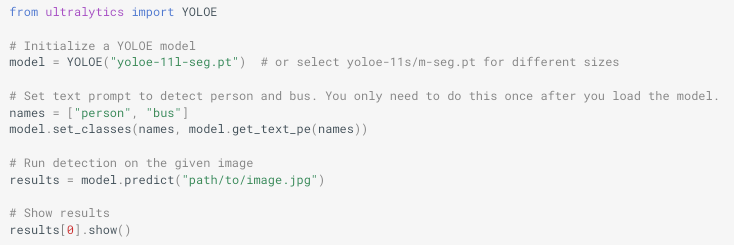
* 只有分类头是可学习的。它把模型绝大部分参数都冻结（不更新），只放开极少量的层（主要是最后的检测/分割头中的某些卷积层）进行训练，以快速适配新的任务或数据。
  + 先把 除最后一层外的所有 backbone + neck 层冻结。
  + 在检测/分割头（model.model.model[-1]）里，所有 名字里不含 cv3 的子模块 也冻结。 这些通常是 early-stage 的卷积或上采样分支，与最终输出关系不大。
  + 额外再冻结 cv3 分支里的前几层卷积（0、1、2 号尺度各自的前两个卷积），只留下 cv3.\*.2 及之后的层 不冻结。
  + 最终效果：只有 cv3.\*.2... 及之后的权重会被更新，其余全部冻结。

1. 预测使用示例

YOLOE 支持文本和视觉提示，只需将提示传递给 predict 方法即可，如下所示：

1. 文本提示

* 用文本描述指定要检测的类别，如下所示检测图像中的人物和公交车：

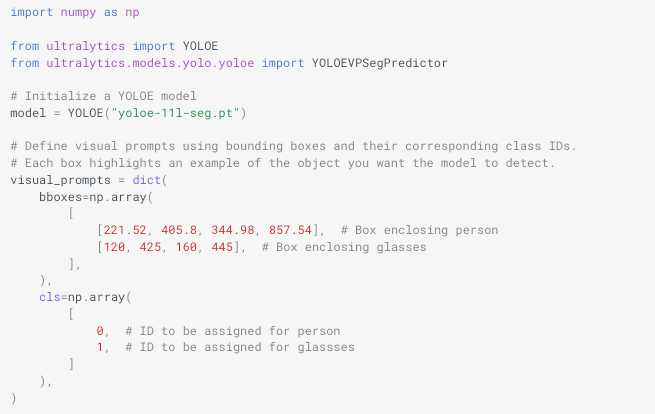


1. 视觉提示

* 通过向模型展示目标类别的视觉示例（而不是用文本描述），来引导模型。
* visual\_prompts 参数接受一个包含两个键的字典：bboxes 和 cls。
  + bboxes 中的每个边界框都应紧密包围您希望模型检测的对象的一个示例，
  + cls 中的相应条目则指定该边界框的类别标签。

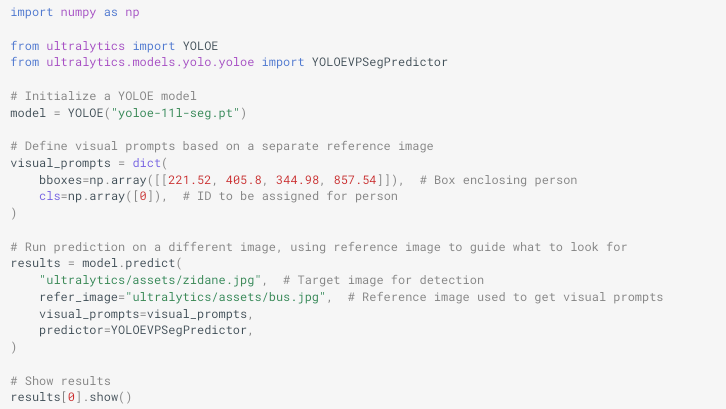
这种配对告诉模型：“这就是 X 类的样子——现在找到更多类似的。”

* visual\_prompts 中的类别 ID (cls) 用于将每个边界框与提示中的特定类别关联起来。它们不是固定的标签，而是您为每个示例分配的临时标识符。唯一的要求是类别 ID 必须是连续的，从 0 开始。这有助于模型正确地将每个边界框与其各自的类别关联起来。、
* 使用示例如下：





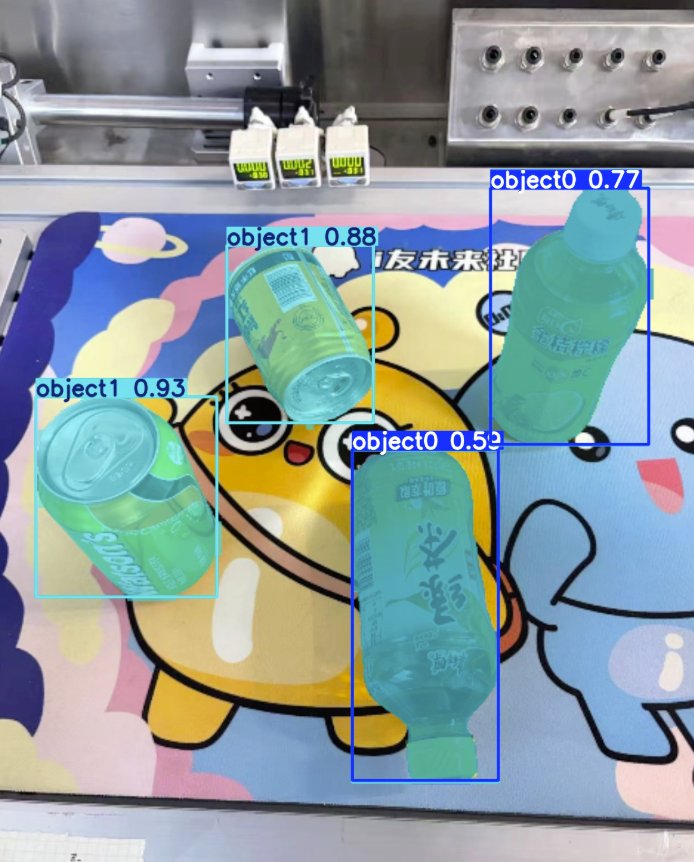
* 或者，可以使用 refer\_image 参数提供来自单独参考图像。在这种情况下，visual\_prompts 中的 bboxes 和 cls 应该描述参考图像中的对象，而不是正在预测的图像：



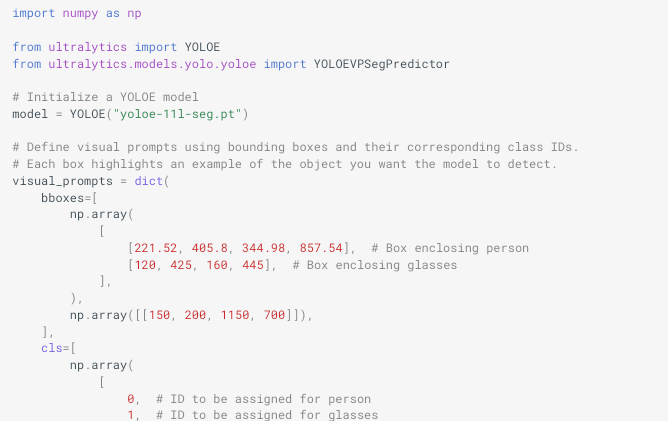
* 参考图像示例：类别0-柠檬，类别1-红牛，类别2-Watson。



* 预测的图像及其结果：柠檬✅ 红牛✅ Watson❌绿茶为未见过的物体，识别为了柠檬。



* 还可以传递多个目标图像来运行预测：

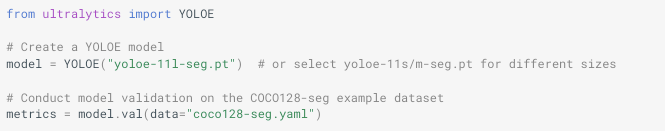




1. 无提示

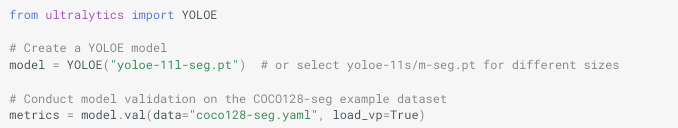
* YOLOE 内置词汇表的无提示版本。模型无需任何提示，其工作原理与传统的 YOLO 模型相同。它不依赖用户提供的标签或视觉示例，而是根据 Recognize Anything Model Plus (RAM++) 中使用的标签集，从预定义的 4,585 个类别中检测对象。

1. 验证使用示例
2. 文本提示

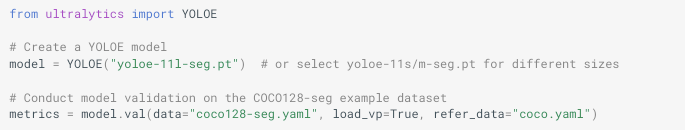


1. 视觉提示

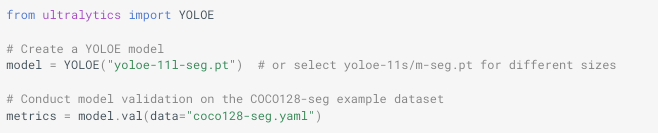
* 默认情况下，它使用提供的数据集为每个类别提取视觉嵌入。



* 或者，可以将另一个数据集用作参考数据集，以提取每个类别的视觉嵌入。请注意，此参考数据集的类别应与提供的数据集完全相同。



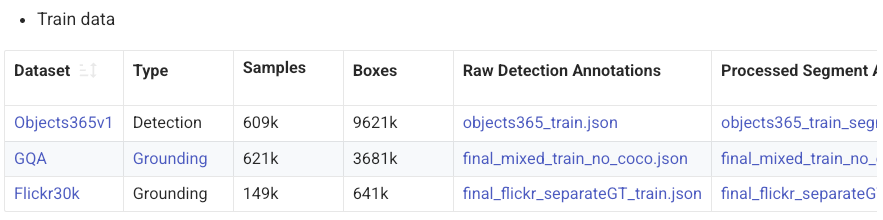
1. 无提示



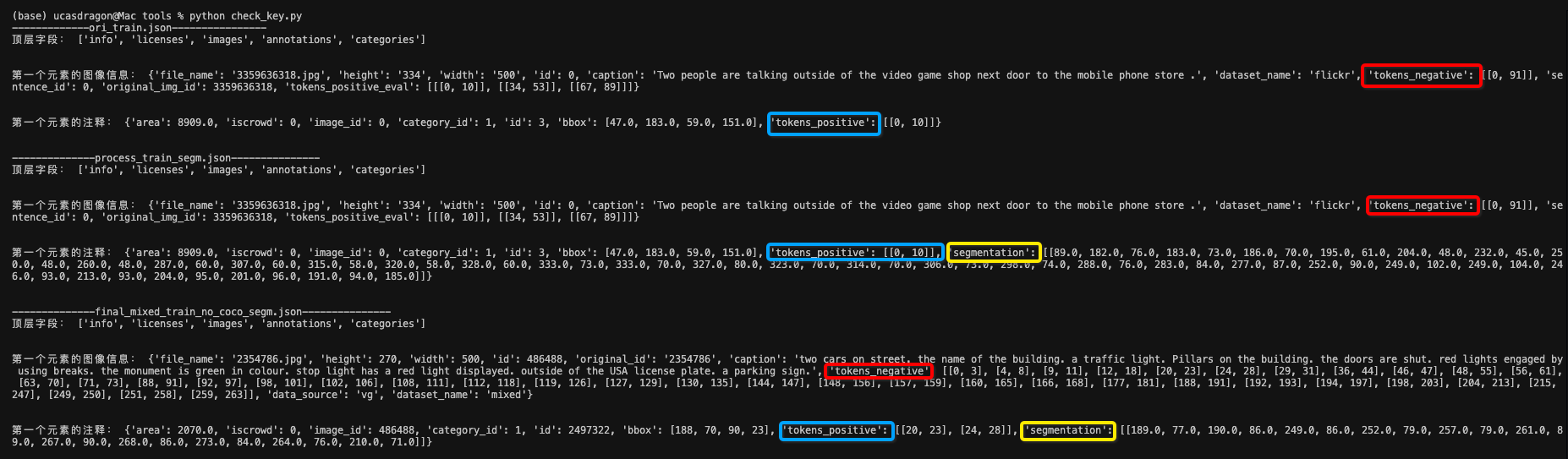
五、训练官方模型

1、准备数据集

* 训练官方 YOLOE 模型需要训练数据的分割标注，官方团队基于 SAM2.1 模型，提供将数据集转换为分割标注的脚本。

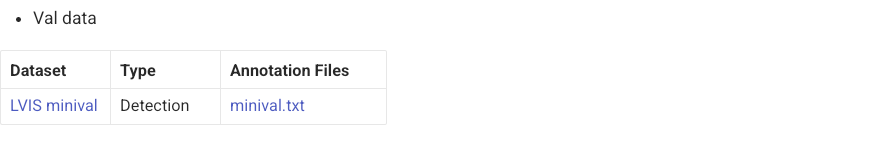


* 处理前、后数据对比：



主要是增加了segmentation字段、tokens\_positive字段。

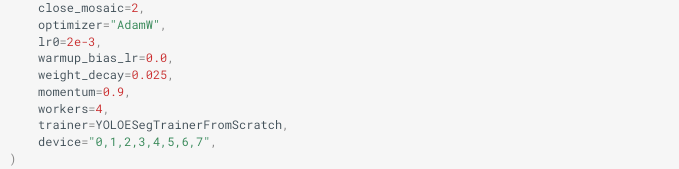
* 验证数据：

2、从头开始训练

视觉提示模型是在训练好的文本提示模型的基础上进行微调的。

1. 文本提示训练：





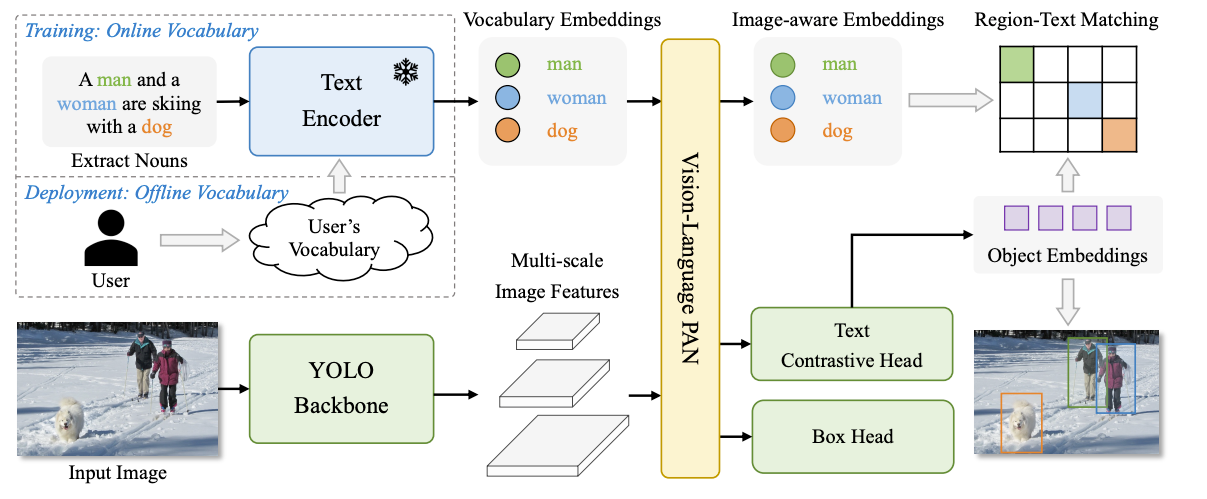
1. 视觉提示训练。

由于训练期间只需更新 SAVPE 模块，因此需要将训练好的 Text Prompt 模型转换为检测模型，并采用训练成本更低的检测流程。此步骤是可选的，也可以直接从分割开始。  
（3）无提示训练。

与视觉提示训练类似，对于无提示模型，训练期间只需更新专门的提示嵌入。将训练好的文本提示模型转换为检测模型，并采用训练成本更低的检测流程。此步骤是可选的，也可以直接从分割开始。

1. YOLOE 与 YOLO-World 有何不同？

* YOLOE 和 YOLO-World 都支持开放词汇检测。
* YOLOE 在 LVIS 上的准确率比 YOLO-Worldv2 高出 3.5 AP，同时训练资源减少了 3 倍，运行速度提高了 1.4 倍。
* YOLOE 还支持三种提示模式（文本、视觉和内部词汇），而 YOLO-World 主要侧重于文本提示。
* 此外，YOLOE 内置实例分割功能，无需额外开销即可为检测到的对象提供像素级精度的掩码。



* 将传统的YOLO检测器提升到新的开放词汇领域，通过大规模数据集进行区域文本对比学习，YOLO-World拥有丰富的区域文本对，展现出强大的开放词汇检测能力，更多训练数据可以带来更大的开放词汇能力提升。
* YOLO-World提出了一种新的可重新参数化的视觉语言路径聚合网络 (RepVL-PAN) 和区域文本对比损失函数，以促进视觉和语言信息之间的交互，从而获得更好的视觉语义表示。在推理过程中，可以移除文本编码器，并将文本向量重新参数化为 RepVLPAN 的权重，以实现高效部署。
* 探索了一种“提示-检测”范式，以进一步提升开放词汇目标检测在实际场景中的效率。