arXiv:2304.03135v1 [cs.CV] 6 Apr 2023

VLPD：通过视觉语言语义自监督实现上下文感知行人检测

Mengyin Liu[1*] JieJiang[2*] ChaoZhu[1†] Xu-Cheng Yin[1]北京科技大学 1
计算机与通信工程学院，中国北京[2]腾讯数据平台部，中国深圳
blean@live.cn, zeus@tencent.com, {chaozhu, xuchengyin}@ustb.edu.cn

摘要

在城市场景中准确检测行人，对于自动驾驶或视频监控等实际应用具有重要意义。然而，容易混淆的人形目标常常会导致错误检测，而由于外观异常，小尺度或严重遮挡的行人也很容易被漏检。为应对这些挑战，仅依靠目标区域是不够的，因此，如何充分利用更明确、更具语义的上下文成为一个关键问题。与此同时，现有的上下文感知行人检测器要么仅利用视觉线索学习潜在上下文，要么需要繁琐的标注来获取显式且具有语义的上下文。因此，本文提出一种基于视觉语言语义自监督的新方法，用于上下文感知行人检测（*VLPD*），无需任何额外标注即可对语义上下文进行显式建模。首先，我们提出一种自监督视觉语言语义（*VLS*）分割方法，通过视觉语言模型生成的语义类别自生成显式标签，同时学习全监督行人检测和上下文分割。此外，我们提出一种自监督原型语义对比（*PSC*）学习方法，基于从 *VLS* 中获得的更明确、更具语义的上下文，更好地区分行人与其他类别。在多个常用基准数据集上的大量实验表明，我们提出的 *VLPD* 的性能优于以往的最先进方法，尤其是在小尺度和严重遮挡等具有挑战性的情况下。代码可从 *https://github.com/lmy98129/VLPD* 获取。

1. 引言

随着行人检测领域的最新进展，众多应用都受益于这样一种基础的感知-

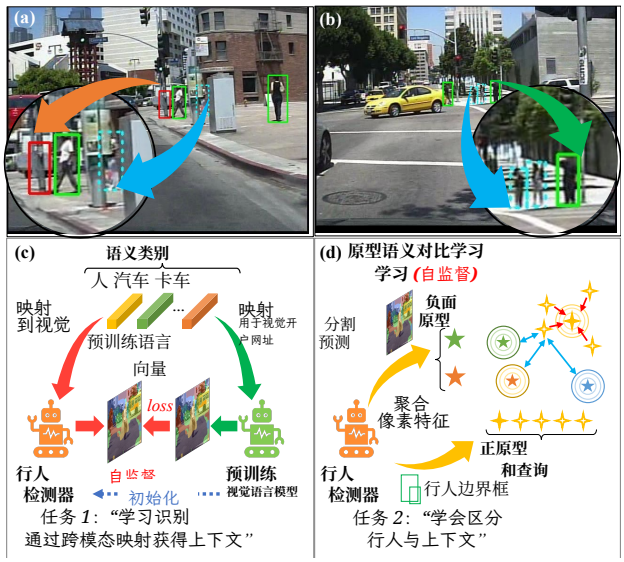

图 1. 以往工作的不足（上）及我们提出的解决方法（下）示意图。(a) 和 (b) 由 [27]预测。绿色框表示正确检测结果，红色框表示类人交通标志，蓝色虚线框表示漏检的严重遮挡或小尺度行人。(c) 和 (d)：我们提出自监督方法来识别上下文，并将其与行人区分开。

测技术，包括行人重识别、视频监控和自动驾驶。与此同时，来自城市环境的各种挑战，即行人和非人类物体，仍然阻碍着检测性能的提升。例如，如图1(a)所示，类人目标外观相似，常常会误导检测器。此外，严重遮挡或尺寸较小的行人外观异常，导致漏检，如图1(a)和(b)所示。除了目标区域，上下文对于应对这些挑战也至关重要。

然而，以往方法对城市场景中的上下文仍缺乏充分探究。对于

*共同贡献。 † 通讯作者。

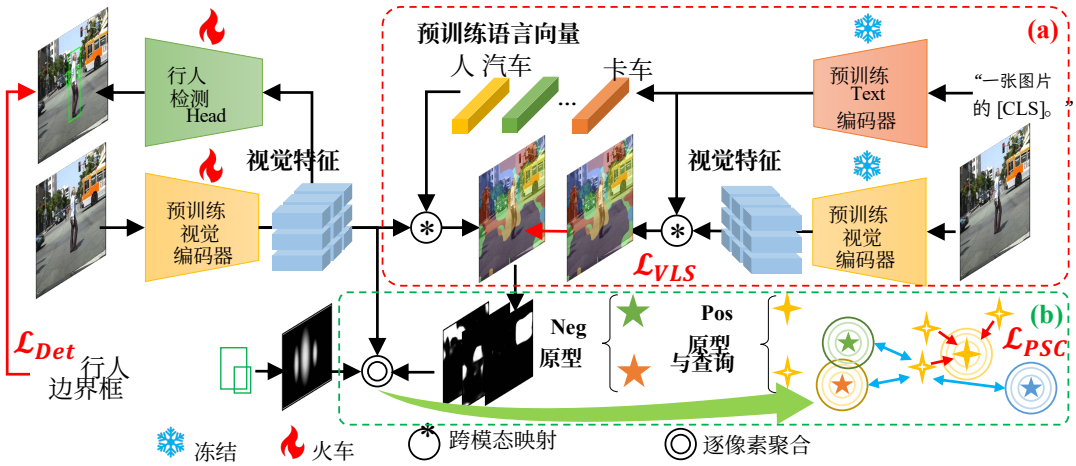

图2. 我们提出的 VLPD 方法的整体架构。(a) 视觉语言语义（VLS）分割通过跨模态映射获取伪标签，随后预训练视觉编码器学习全监督检测（$\mathcal{L}_{Det}$）和自监督分割，在无需任何标注的情况下识别显式上下文中的语义类别。(b) 原型语义对比（PSC）学习基于逐像素聚合，使逐像素行人特征作为查询更接近正原型，并远离负原型。

立场，来自 CityScapes 的人工上下文标注 [6]提升 SMPD [15] 在行人基准数据集 CityPersons上[43], 因为它们共享同源图像数据。此外，一个半监督模型为 Cal-tech 数据集生成伪标签 [9]。然而，这两种解决方案都需要昂贵的细粒度标注，尤其是在训练半监督模型时。此外，其他方法仅从有限的视觉邻域中学习区域潜在上下文 [47], 或将非人类局部提议作为对比学习的负样本 [23]。由于缺乏对上下文中语义类别的明确感知，这些方法的性能仍然不尽如人意。

此外，一些行人检测方法也会间接处理上下文。针对遮挡问题，许多部件感知方法 [4, 13, 19, 20, 28, 33, 41, 44, 45]会为被遮挡的行人采用可见区域标注，这些标注指出了上下文中其他行人或非人类物体造成的遮挡。然而，这些标签仍需人工标注者付出大量劳动。对于尺度变化 [2,8,21,39,46],人群遮挡 [14,25,38,40,48,50]或一般的困难行人 [1, 24, 26, 27, 34], , 大多数以往工作都局限于类内情形，例如小尺寸行人或拥挤场景，因此与上下文建模问题无关。

受视觉语言模型启发，我们发现，通过跨模态映射，无需任何标注即可进行更明确的上下文建模。例如，DenseCLIP [32] 使用视觉-语言预训练CLIP模型 [31]进行初始化，以学习从像素级特征到人工标注类别的语言向量的跨模态映射。同时，MaskCLIP [49] 通过跨模态映射生成伪标签，并训练另一个视觉模型。因此，作为对初始化映射和伪标签生成的补充，我们提出识别语义类别，以实现明确的上下文——通过自监督视觉语言语义（VLS）分割得到的文本，如图1(c)和2(a)所示。

此外，我们认为，只有像素级分数难以判别行人与上下文。由于粗粒度的伪标签，行人的某些部分可能会获得其他类别的更高分数。不同于区域对比学习 [23], , 我们引入了原型的概念 [35, 51] , 以实现全局判别。基于从VLS中获得的显式上下文，将行人特征的每个像素拉近至像素级聚合的正原型，并推远至其他类别的负原型。如图1（d）和2（b）所示，我们提出了一种用于行人检测的新型对比自监督，以更好地区分行人与上下文。

总之，我们观察到显式上下文的繁重人工标注负担与局部隐式上下文建模之间存在两难困境。因此，我们提出一种新方法，通过视觉语言语义自监督来解决这些问题，用于行人检测（VLPD）。本文的主要贡献如下:

- 首先，提出了视觉语言语义（VLS）分割方法，通过视觉语言模型对显式语义上下文进行建模。借助跨模态映射生成的伪标签，视觉编码器学习全监督检测和自监督分割，从而识别显式上下文的语义类别。**据我们所知，这是首个提出此类无需视觉语言额外标注的行人检测方法的工作**。

- 其次，我们进一步提出原型语义对比（PSC）学习方法，以更好地区分行人和上下文。负样本和

正原型分别通过从VLS和行人边界框中获得的上下文语义类别的分数图进行聚合。行人特征的每个像素都被拉近正原型并推离负原型，以增强检测器的判别能力。

- 最后，通过融合 VLS 和 PSC，我们提出的 VLPD 方法在广受欢迎的 Caltech 和 CityPersons 基准数据集上，性能优于以往的最先进方法，尤其是在具有挑战性的小尺度和遮挡子集上。

2. 相关工作

2.1. 行人检测

在实际应用中，多种情况都会给行人检测带来挑战，包括遮挡、尺度变化和通用困难行人处理。本文将讨论这些常见问题，以及针对这些问题的上下文感知方法。

2.1.1 遮挡处理

作为行人检测领域的研究热点，遮挡处理应充分利用行人的可见部分所提供的有限信息，同时避免受到其他行人或非人类物体遮挡所引入的噪声影响。

另一方面，部件感知方法处理被上下文物体遮挡的其他部分所对应的可见部分。例如，OR-CNN [44] 重新评估各部分，以突出显示可见部分。PRNet++ [33] 逐步细化预测的可见部分框和全身框。较少被遮挡的头部的额外标签促进了 HBAN [28], JointDet [5] 和 Ped- Hunter [4]。一些方法 [2,19,20]通过并行分支处理可见部分和完整人体。此外，DMSFLN [13]探索两个分支之间的特征分布。

另一方面，人群感知方法专门处理缺乏上下文建模的类内遮挡。一些后处理方法 [14, 25, 29] 专注于抑制人群场景中密集预测结果时的过度抑制问题，而其他方法 [48] 则处理对稀疏预测结果抑制不足的问题。对于高度重叠的行人，基于损失的方法 [36,38,40] 通过学习表征来识别它们。

不同的是，我们提出的 VLPD 基于通过 VLS 自生成的语义类别显式标签进行自监督学习，并通过 PSC 的对比学习来区分非人类遮挡物与行人。

2.1.2 尺度变化处理

尺度变化是另一个可能与上下文建模相关的问题。由于距离的影响，小尺寸行行人和非人类物体模糊且带有噪声的外观常常会使检测器混淆。多分支 [8,21,46] 在建模不同尺度方面十分常见。借助强大的特征金字塔网络 [22] 架构，LBST [2]通过融合自底向上和自顶向下的特征来检测更小的行人。不同的是，S ML [39]将小尺寸行人的特征推向大尺寸行人的分布。

遗憾的是，由于缺乏针对小型非人类物体的标签，这些工作主要关注小尺度行人。因此，我们提出的 VLPD 使用包含后者的无标签显式上下文，并在不同尺度上取得了更好的表现。

2.1.3 通用困难行人处理

处理困难行人的核心问题是实现精准定位。大量先前工作 [1,26]引入了多阶段空间细化。CSP [18, 27] 系列遵循通用目标检测中的无锚框风格，通过自适应预测减少超参数。AP^{2}M[24]为不同困难样本匹配适当的参数。

在上下文建模方面，SMPD [15] 采用额外的分割标注，EGCL [23]通过局部提议进行对比学习，而 FC-Net [47]则学习局部上下文的潜在特征。无需任何额外标签，我们提出的 VLPD 可以通过 VLS 的伪标签识别显式上下文物体，然后利用 PSC 的更全局正负原型将其与行人区分开来。

2.2. 视觉语言预训练分割

近期视觉语言预训练 CLIP[31] 的发展推动了更强大的分割方法。例如，通过计算视觉特征与语言向量之间的余弦相似度（即跨模态映射），可获得分割结果。DenseCLIP [32]和 LSeg[17] 使用预训练的 CLIP 视觉编码器初始化模型，然后通过语言向量学习映射特征与标注类别之间的对应关系。在自监督方面，MaskCLIP [41] 通过映射获得伪标签，并据此训练新的视觉模型； [32]的评估表明，这种方法效果并不理想。

与跨模态映射和伪标签生成不同，我们提出的新型自监督 VLS 无需任何额外标签，即可将语义类别识别为显式上下文，用于上下文感知行人检测。

2.3. 原型对比学习

由于图像的空间分辨率限制，纯像素级密集对比学习 [37]会带来沉重的计算负担，并且只进行局部判别，而不考虑全局图像。因此，先前的工作引入“原型” [35,51] 作为各语义类别像素特征的替代方案。不同的是，我们提出的 PSC 将行人像素保留为查询，以保留其内部方差，从而为上下文感知行人检测学习行人与其他类别之间更强的判别能力。

3. 提出的方法

如图所示2 和 3所示，我们提出的方法 **V**ision-**L**anguage语义自监督，用于 **P**edestrian**D**etection（**VLPD**）是一种遵循基线 CSP 的无锚检测框架 [27]。预训练视觉编码器从 S3 到 S5 的不同阶段提取特征。如图3所示，这些特征被拼接成“检测特征”，供检测头进行预测。

为实现无需任何额外标签的显式语义上下文建模，我们的架构由两个关键组件构成：视觉语言语义（ VLS）分割和原型语义对比（PSC）学习。VLS 利用视觉语言模型识别显式上下文，视觉编码器则通过跨模态映射生成的语义类别显式标签，同时学习全监督行人检测与分割。PSC 基于 VLS 对检测器进行监督，使其能够更好地区分行人与上下文语义类别。更多细节将在以下章节中介绍。

3.1. 视觉语言语义分割

得益于自监督跨模态对比学习，视觉语言模型会将含义相近的视觉向量和语言向量映射到统一特征空间中，使它们彼此更加接近。因此，我们可以通过语言向量获取图像中语义类别的存在情况。然而，以往的工作会先使用跨模态映射初始化模型，再进行全监督训练 [17, 32], ，或者通过映射得到的伪标签重新训练一个新模型 [41]。

因此，如图2(a)所示，我们提出了视觉语言语义（VLS）分割，作为初始化映射和伪标签生成的补充。标签由基于跨模态映射的冻结预训练模型生成，从而监督未冻结的视觉编码器预测显式语义类别的分割结果，这些类别充当更全局的上下文，而非先前的局部潜在上下文 [23, 47]。关于我们提出的 VLS 的更多细节将在后续章节中介绍。

3.1.1 用于伪标签生成的跨模态映射

CLIP[31] 作为最受欢迎的视觉语言模型之一，能够基于通过自监督跨模态对比学习预训练的视觉编码器和语言编码器，将含义相近的图像和文本映射到距离更近的向量。尽管无法恢复

通过对 CLIP 进行基于注意力的池化，从图像的视觉向量中获取图像的逐像素上下文信息 [31],该池化操作可以修改为投影，以保留视觉区域，遵循 [49]。如图4所示，余弦相似度$S_i^c = ((L^c)^\top(V_i))/(\|L^c\|\|V_i\|) \in S$ 是在每个类别的预训练语言向量 $L^c \in \mathbb{R}^{D'}$ 与投影视觉特征 $c \in C$ 之间计算的 $V_i \in \mathbb{R}^{D'}$

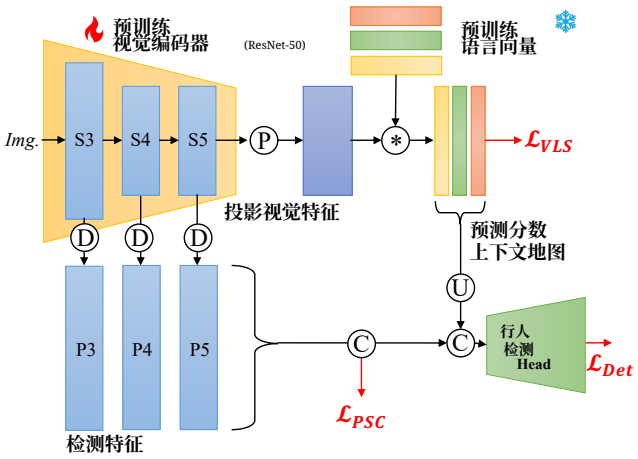

图3. 我们提出的 VLPD 的详细网络架构。视觉编码器为 ResNet-50 [12] ，来自视觉-语言预训练 CLIP 模型 [31]。遵循基线 CSP [27], S3∼S5 经过反卷积（“D”）后作为“检测特征”，用于 $\mathcal{L}_{Det}$ ，并进一步由我们的 $\mathcal{L}_{PSC}$ 进行监督。跨模态映射前的投影（“P”）采用类似 [49] 的方式。拼接和上采样（“C”和“U”）用于预测，如 [32]。为简洁起见，图2中省略了这些细节。

每个像素的 $i = 1, 2, \ldots\ldots, H'W'$，这意味着图像中每个像素处都存在相应的上下文类别。

为了获得自监督学习所需的伪标签，如图 2(a) 所示，将图像输入冻结的CLIP视觉编码器以提取视觉特征，并通过冻结的文本编码器，利用提示句“一张 [CLS] 的图片”。生成类别的语言向量。

3.1.2 VLS 的自监督学习

通过实验评估发现， [32], 使用 CLIP 预训练视觉编码器进行初始化，有助于保持跨模态映射，使模型能够充分利用每个像素与每个语言向量之间的相似性。而 ImageNet [7] 预训练的视觉编码器 [49]则需要重新训练以适应映射，因此性能并非最优。因此，我们提出兼顾伪标签生成与映射初始化各自的优势。

具体而言，如图2(a) 和3所示，行人检测器的视觉编码器使用 CLIP 预训练参数进行初始化，与用于自监督伪标签生成的冻结视觉编码器完全相同。由于也执行了跨模态映射，语言向量只需生成一次。预测的¯W N 使用基于平滑 L1 损失的伪标签进行监督， [10] $\mathcal{L}_{VLS}$ 该损失对噪声标签具有鲁棒性：

$$\mathcal{L}_{VLS} = \frac{1}{H'W'N} \sum_{i,c} Smooth\mathrm{L1}(\bar{S}_i^c, S_i^c), \quad (1)$$

其中类别数量 N $= |C|$。超过自监督

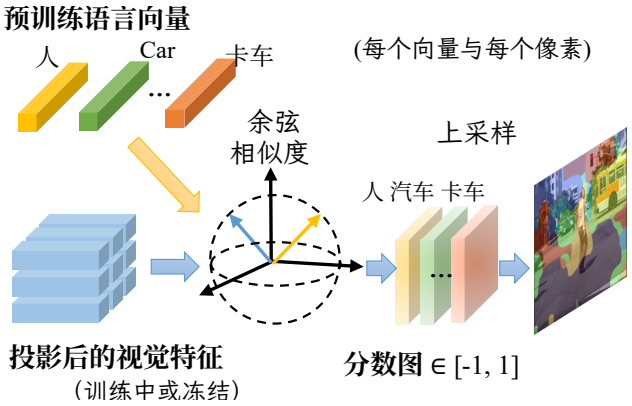

图4. 我们提出的 VLS 的跨模态映射。视觉特征被投影为 [49]，而非 CLIP 预训练中的全局向量 [31]。计算语义类别的每个语言向量与视觉特征中每个像素之间的余弦相似度，由此构成这些类别在上下文中的分数图。

在公式1，视觉编码器也经过训练以检测行人，作为基线 [27]。因此，模型学会在检测过程中对上下文进行建模，并在统一的流程中。预测的 S 被输入图中的检测头3，以获得明确的上下文参考，例如 [32]。

3.1.3 紧凑类别策略

行人检测所面临的复杂城市场景要求使用恰当的上下文类别集合。同时，城市数据集 CityScapes [6] 已经定义了多种类别。然而，实验（见后续章节）表明，这会带来负面影响。受 CityScapes 论文中类别频率统计不均衡现象的启发 [6],我们在表 1 中提出紧凑类别策略，以决定哪些类别应予以保留、合并或舍弃。

对于大多数类别，我们采用 2nd CityScapes 中高于原始类别层级的类别层级，如表 1所示。此外，“车辆”类别内部较大的方差会导致实验结果性能下降。因此，1st 和 2nd 类别层级均不适用。

因此，我们针对 CityPersons 与 CityScapes 图像重合部分的像素级标注统计了各类别的频率 [43]，，统计方式不仅包括逐像素计数，还按照图像级出现次数进行重新加权。通过设置阈值省略频率较低的尾部类别，并保留频率较高的类别，如表 1底部所示。

总而言之，我们提出的 VLS 利用强大的视觉语言模型CLIP [31] 基于跨模态映射和紧凑类别策略执行自监督学习，从而为显式上下文获取语义类别的伪标签，以便在检测过程中学习识别这些上下文，进而提升判别能力。

表1. 我们提出的 VLS 的紧凑类别策略。

原始 →	紧凑的	Used
{道路, 人行道}	地面	✓
{建筑物、墙、围栏}	建筑物	✓
{植被，地形}	tree	✓
{人，骑行者}	人	✓
{杆，交通信号灯，交通标志}	交通标志	✓
{汽车、自行车、公交车、卡车、摩托车，火车}	车辆	×
	{汽车，自行车，公交车, 卡车}	✓

3.2. 原型语义对比学习

由于我们提出的 VLS 生成的伪标签具有粗粒度特征，行人的某些可见部分可能会从VLS 中获得其他类别的更高分数，而这些部分会通过边界框进行标注，以用于检测任务。由于没有可用的人工标注，对伪标签进行显式细化相当困难。

受自监督对比学习的启发，该方法旨在获得正样本和负样本的判别性表征 [35, 37,51]。我们引入这一强大技术，并提出一种新颖的原型语义对比（PSC）学习方法，无需任何额外标签，即可更好地判别行人和其他语义类别。

为降低密集逐像素方法带来的大量计算开销 [37], ，我们采用“原型”这一概念。原型是能够代表同一语义类别所有特征的代表性特征。本文中，行人的原型为正原型，其他类别的原型则为负原型。

以负原型为例，预测得分图 $\bar{S} \in \mathbb{R}^{\mathrm{H}' \times \mathrm{W}' \times (\mathrm{N}-1)}$ 用作指示所有非人类类别在空间中的存在情况，但与行人重叠的“人”类别除外。为简便起见，我们将 C 记作非人类类别，其中 $|C| = \mathrm{N} - 1$。 $\dot{S} \in \mathbb{R}^{\mathrm{H} \times \mathrm{W} \times (\mathrm{N}-1)}$ 通过对 S 进行上采样得到。我们采用温度为 τ' 的 Softmax 函数 δ，将 $S \in [-1, 1]$ 归一化为 $S \in [0, 1]$: ˙

$$\hat{S} = \delta(\dot{S}) = \{\frac{\exp(\dot{S}_i^c/\tau')}{\sum_{d \in C} \exp(\dot{S}_i^d/\tau')} \mid c \in C, i = 1, 2, ..., \mathrm{HW}\}. \tag{2}$$

由于我们不应干扰 VLS 的自监督学习，而只旨在改进检测，“检测特征” $E \in \mathbb{R}^{\mathrm{D} \times \mathrm{H} \times \mathrm{W}}$ 受到监督。如图 5 所示，各类别的原型通过聚合 ˆE 逐像素地由 S 表示为逐像素聚合:

$$P^- = E \cdot \hat{S} = \{P^{c-} = \sum\nolimits_i E_i \cdot \hat{S}_i^c \mid c \in C\}, \tag{3}$$

其中，$\cdot$ 是矩阵乘法，$P^- \in \mathbb{R}^{\mathrm{D} \times (\mathrm{N}-1)}$ 是

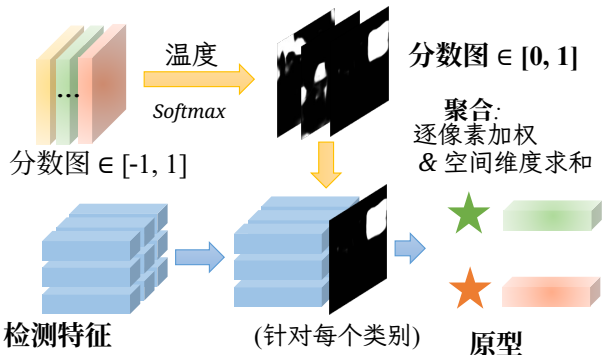

图 5. 我们提出的 PSC 的逐像素聚合。“检测特征”根据 VLS 预测的分数图进行逐像素加权，然后在空间维度上相加，聚合为原型。

非人类类别的负原型 $\mathrm{N}-1$ 个具有通道 D。由 $\hat{S}_i^c \in \mathbb{R}$加权后，特征 $E_i \in \mathbb{R}^D$ 在每个位置 $i=1$、2、.....、HW 处进行聚合。同样，来自基线 CSP $G \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 1}$ 的行人位置 [27] 的二维高斯分布可以替代正原型的 $\hat{S}$。 P^+。

最后，每个像素级的 E_j 标注的行人的像素都被作为对比损失函数的“查询”进行监督 $\mathcal{L}_{PSC}$见图 1(d) 和 2(b)，将它们拉近到 P^+ 并将它们彼此推离 P^{c-}。 E_j 由 >0 的位置 j 确定 G 和 $|G^{>0}|=\mathrm{M}$。 $\mathcal{L}_{PSC}$ 其公式如下:

$$-\frac{1}{\mathrm{M}}\sum_{j\in G^{>0}}\log\frac{\exp(E_j\cdot P^+/\tau)}{\exp(E_j\cdot P^+/\tau)+\sum_{c,b}\exp(E_j\cdot P_b^{c-}/\tau)}, \tag{4}$$

为简洁起见，省略了特征和原型的自归一化，例如 $E_j/\|E_j\|$。与一些对比预训练方法 [3, 11]，类似，负原型会扩展到小批次 $b\in B$ 中的所有图像 B。

简而言之，基于对比自监督，PSC 借助正负原型训练检测器，使其能够更好地区分行人与其他类别，这得益于从 VLS 获得的显式语义上下文。通过整合 VLS 和 PSC，我们提出的 VLPD 同时接受 $\mathcal{L}_{Det}$ [27]、$\mathcal{L}_{VLS}$ 和 $\mathcal{L}_{PSC}$ 的监督:

$$\mathcal{L}=\mathcal{L}_{Det}+\lambda_1\mathcal{L}_{VLS}+\lambda_2\mathcal{L}_{PSC}. \tag{5}$$

3.3. 检测头

遵循基线 CSP 的无锚框风格 [27],图 2 和 3 中的检测头首先通过卷积层减少“检测特征”的通道数。随后，多分支预测结果图：“中心热力图”用于分类行人中心，“尺度图”用于预测高度，并按固定宽高比 0.41 推算宽度，“偏移图”则用于水平和垂直方向上的定位调整。最后，将这些图组合成行人的边界框。更多详情请参阅 CSP 论文。

4. 实验

本节在两个广受欢迎的行人检测基准数据集 Caltech 和 CityPersons 上进行了大量实验，以评估我们提出的 VLPD 方法。我们针对关键组件 VLS 和 PSC 开展了消融研究。此外，我们还报告了在这两个基准数据集上的最新对比结果。更多实验和可视化内容请参阅补充材料。

4.1. 数据集

Caltech 行人数据集 [9] 包含在洛杉矶市区拍摄的 2.5 小时视频数据，其中有 4024 张测试图像。超过 70% 的行人高度不足 100 像素，其中包括高度不足 50 像素的特别小的行人。通过修正标注不一致和边界框错位问题，Zhang 等人 [42] 发布了新版标注。为公平比较，以下所有评估均基于该新版进行。

CityPersons [43]是一个较新发布的大规模行人检测数据集。2975 张图像用于训-练，500 张图像用于验证。标准评估指标如下：在每图假阳性数（FPPI）$\in[10^{-2}$范围内对数漏检率的平均值；$100]$, 记为 MR^{-2}。为公平比较，所有测试均在未缩放的原始数据（1×）上进行，且不使用任何额外的可见或头部标签。

4.2. 实现细节

我们提出的方法基于强大的行人检测器 CSP [27], ，该检测器在 PyTorch[30]框架中由原始 Keras 版本重新实现。我们采用 Adam [16] 进行优化。骨干网络为在 ImageNet [7] 上通过全监督图像分类或在 WIT [31] 上通过自监督视觉-语言对比学习进行预训练的 ResNet-50 [12]。对于 Caltech，使用一块 Nvidia 3090 GPU 进行训练，学习率为 10^{-4}。对于 CityPersons，使用两块 3090 GPU，参数为 2×10^{-4}。批量大小按照 [27]设置。所有测试均在单块 3090 GPU 上进行。训练图像尺寸在 Caltech 上为336×448，在 CityPersons 上为 640×1280 。对于我们提出的 VLS，其损失权重为 $\lambda_1=100$。对于 PSC，其权重在 Caltech 上为 $\lambda_2=10^{-4}$，在 CityPersons 上为 10^{-3}。温度参数为 $\tau'=10^{-3}$ 和 $\tau=7\times10^{-2}$，按照 [11]设置。

4.3. 消融研究

首先在广泛使用的 CityPersons 数据集上进行消融研究。常用的比较子集包括综合子集“合理”，以及更具挑战性的 Small 和 HO（严重遮挡，可见率为 $\in[0.2, 0.65]$）。

表 2 展示了我们提出的VLPD 各个关键组件的整体消融研究。我们提供了 CSP 的原始结果 [27]以及我们使用 Py-Torch 重新实现的结果 [30]。在 CLIP [31]初始化的视觉编码器



图 6. Caltech 上的定性分析 [9] 基线CSP [27]（上）与我们提出的VLPD（下）的对比。绿色表示正确检测，红色表示错误检测，蓝色虚线表示漏检。借助强大的视觉-语言语义自监督，我们提出的VLPD具备上下文感知能力，因此对类人目标、类间遮挡和具有歧义的小尺寸行人更加鲁棒。

表 2. 我们提出的VLPD 的关键组件（包括 VLS 和 PSC）的总体消融研究。**加粗显示** 表示最佳结果。

方法	合理	小	HO
CSP [27]	11.0	16.0	-
CSP（我们的重新实现）	10.96	16.05	40.59
CSP 与 CLIP	10.13	12.59	38.97
+VLS	9.70	12.57	36.50
+VLS+PSC=**VLPD**	**9.41**	**10.93**	**34.88**

作为VLS的前提条件，其改进受到限制，因为仅凭视觉语言预训练无法完全处理上下文建模。借助VLS以及基于VLS的PSC，我们提出的VLPD取得了显著提升，尤其是在与上下文相关的Small和HO子集上。

表 3评估了我们提出的 VLS 所采用的不同类别集策略。完整 CityScapes 策略采用 CityScapes 数据集的所有类别 [6], ，而完整紧凑策略使用 2nd 级别的类别集。过于分散或集中过度的类别集都会导致性能下降。相反，我们提出的 VLS 策略通过频率统计处理数量最多的“车辆”类别，以实现更好的上下文建模。

与此同时，表 3 底部通过恢复至 1st 表的列1，评估了该策略的子项。表 4显示，式4 中的 PSC 采用“检测特征”的跨图像负样本和图像内正样本E，且不干扰 S VLS 所提供的内容时，性能最佳。

4.4. 与最先进方法的比较

对于 CityPersons，我们将提出的 VLPD 与多种最先进的方法进行比较：AMSCNN [46], DHRNet [8] 以及 SML [39]，用于尺度变化；RepLoss [36], 、Adaptive NMS [25], 、PBM+、R^2NMS [14], 、CaSe [40], 、NMS-Ped[29] 和 MAPD [38]，用于类内遮挡；OR-CNN[44], 、HBAN [28]和 PRNet++ [33]，用于部件感知的遮挡处理；LBST [2], 、ALFNet [26], 、CSP [27], 、AP^2M

表 3. 我们提出的 VLS 的类别集采用的不同策略。

方法	合理	小	HO
带 CLIP 的 CSP	10.13	12.59	38.97
+完整 CityScapes	10.40	12.87	37.14
+完全压缩生物系	10.47	13.30	40.49
+VLS（我们的方法）	**9.70**	12.57	**36.50**
不含地面	10.51	13.53	37.48
无建筑物	10.42	12.84	37.70
不使用树	10.34	12.66	38.39
无行人	10.24	12.52	38.47
不含 {汽车、自行车，公交车，卡车}	10.61	13.61	38.84
w/o 交通标志	10.11	**12.27**	37.22

表 4. 我们提出的 PSC 所采用的不同原型和特征。

方法	合理	小	HO
VLPD（含 PSC，我们的方法）	**9.41**	**10.93**	**34.88**
交叉 $\rightarrow$ 内向图像负样本	10.12	12.67	37.90
内部 $\rightarrow$跨向图像正样本	10.58	13.14	38.08
$E \rightarrow$拼接(E, $\dot{S}$)	10.04	12.93	38.21

表 5. 在 CityPersons 上与最先进方法的比较。

方法	R	Hea.	部分	Bare	小型
AMSCNN [46]	14.0	-	-	-	12.6
FC-Net [47]	13.9	46.8	-	-	-
RepLoss [36]	13.2	56.9	16.8	7.6	42.6
OR-CNN [44]	12.8	55.7	15.3	6.7	42.3
LBST [2]	12.6	48.7	18.6	-	-
SML [39]	12.3	-	-	-	19.3
ALFNet [26]	12.0	51.9	11.4	8.4	19.0
AdaNMS [25]	11.9	55.2	12.6	6.2	-
PR^{2}NMS [14]	11.1	53.3	-	-	-
CSP [27]	11.0	49.3	10.4	7.3	16.0
CaSe [40]	11.0	50.3	-	-	-
HBAN [28]	10.9	47.0	-	-	-
EGCL [23]	10.9	46.4	11.6	7.4	-
PRNet++ [33]	10.7	51.2	9.9	6.9	-
AP^{2}M [24]	10.4	48.6	9.7	6.2	15.3
DHRNet [8]	10.4	-	-	-	13.4
NMS-Ped [29]	10.1	-	-	-	-
SMPD [15]	9.9	45.6	9.0	6.5	-
MAPD [38]	9.7	46.4	9.9	6.1	-
BGCNet [18]	9.4	45.9	9.0	6.4	-
VLPD（我们的方法）	**9.4**	**43.1**	**8.8**	**6.1**	**10.9**

[24] 以及 BGCNet [18]，用于通用困难行人检测。请注意，与上下文相关的方法包括：SMPD [15]，使用分割标注；EGCL [23]，采用逐提议对比学习；以及 FC-Net [47]，采用邻域建模。如表5所示，在所有子集上，我们的 VLPD 均全面优于这些方法。

具体而言，我们将合理子集记为“R”，并将遮挡子集 Heavy（可见率 $\in [0, 0.65]$）记为“Hea.”，如表 5所示。我们还在表 6中将所提出的 VLPD 与其他遮挡子集上的方法进行了比较，即R+HO（合理+HO，可见率 $\in [0.2, 1]$）和 HO。我们的方法在这些设置变化下始终保持最佳性能。

对于 Caltech，还比较了其他先进方法：AR-Ped [1]（用于通用困难行人处理）；JointDet [5], PedHunter [4]和 DMSFLN [13]（使用可见区域或头部标签）。在表 7中，无需任何额外标签，我们提出的方法 VLPD 也显著优于这些方法。其合理子集上的 2.27% 优于 [4]的 2.31%。与上下文相关的挑战性子集严重遮挡（37.7%）和全部（52.4%）的表现尤其优于其他方法。

总之，我们提出的 VLPD 已在两个基准数据集上，尤其是在与上下文相关的子集上，达到新的最先进水平，这充分验证了其视觉语言语义自监督的能力：无需任何额外标签即可显式建模语义上下文，并更好地区分行人与其他上下文类别。

表6。在 CityPersons 上其他遮挡子集的比较。

方法	合理	R+HO	HO
FC-Net [47]	13.9	29.6	-
ALFNet [26]	12.0	26.3	43.8
EGCL [23]	10.9	24.8	39.3
PRNet++ [33]	10.7	25.4	40.9
SMPD [15]	9.9	-	36.6
VLPD（我们的方法）	**9.4**	**21.7**	**34.9**

表 7. 在 Caltech 上与最先进方法的比较。

方法	合理	All	重型
ALFNet [26]	6.1	59.1	51.0
RepLoss [36]	5.0	59.0	47.9
CSP [27]	4.5	56.9	45.8
AR-Ped [1]	4.4	-	-
BGCNet [18]	4.1	-	42.0
DHRNet [8]	3.4	-	-
平均精度^{2}M [24]	3.3	55.9	42.2
JointDet [5]	3.0	-	-
DMSFLN [13]	2.7	-	-
PedHunter [4]	2.3	-	-
VLPD（我们的方法）	**2.3**	**52.4**	**37.7**

5. 结论

在本文中，我们提出了一种新颖的行人检测方法 VLPD，通过显式上下文建模来应对具有挑战性的问题，例如类人目标以及小尺度或严重遮挡的行人。该方法通过视觉语言语义自监督来解决这些挑战，包含两个关键组件：VLS旨在利用视觉语言模型识别显式上下文中的语义类别，并通过跨模态映射，利用伪标签学习全监督行人检测和自监督分割。PSC旨在采用对比自监督，根据来自 VLS 的显式上下文，更好地区分行人和语义类别。通过整合 VLS 和 PSC，我们的 VLPD 在两个具有挑战性的基准数据集 Caltech 和 CityPersons上取得了新的顶尖性能，尤其是在小尺度和严重遮挡这一极其困难的场景下。

6. 致谢

本研究得到国家重点研发计划（2020年AAA0109701）、国家自然科学基金（62072032、62076024）以及国家杰出青年科学基金（62125601）的资助。

References

[1] Garrick Brazil 和 XiaomingLiu。使用自回归网络阶段进行行人检测。在IEEE计算机视觉与模式识别会议论文集中，第7231–7240页，2019年。2, 3, 8[2] Jiale Cao、Yanwei Pang、JungongHan、BolinGao 和 Xue-longLi。关注小尺度行人和被遮挡行人。IEEE图像处理汇刊，29:3143–3152，2019年。2, 3, 7, 8[3] Xinlei Chen、Haoqi Fan、RossGirshick和KaimingHe。通过动量对比学习改进基线。arXiv预印本 arXiv:2003.04297，2020年。6[4] ChengChi、Shifeng Zhang、Junliang Xing、Zhen Lei、StanZ Li和Xudong Zou。PedHunter：拥挤场景中的遮挡鲁棒行人检测器。在 AAAI 人工智能会议论文集，第34卷，第10639– 10646页，2020年。2,3, 8[5] ChengChi、Shifeng Zhang、Junliang Xing、ZhenLei、Stan Z Li 和 Xudong Zou。用于联合头部和人检测的关系学习。在 AAAI 人工智能会议论文集，第34卷，10654，2020年 10647 3, 8 页 – 。[6] Marius Cordts、Mohamed Omran、Sebastian Ramos、Timo Rehfeld、Markus Enzweiler、Rodrigo Benenson、Uwe Franke、StefanRoth 和 Bernt Schiele。用于语义城市场景理解的 CityScapes数据集。在IEEE计算机视觉与模式识别会议论文集，第3213–3223页，2016年。2, 5, 7[7] Jia Deng、Wei Dong、RichardSocher、Li-JiaLi、KaiLi 和 Li Fei-Fei。ImageNet：一个大规模分层图像数据库。在2009年IEEE计算机视觉与模式识别会议，第248–255页。IEEE，2009年。4, 6[8] MengyuanDing、Shanshan Zhang 和 JianYang。学习用于多尺度行人检测的动态高分辨率网络。在2020年第25届国际模式识别会议（ICPR），第9076–9082页。IEEE，2021年。2, 3, 7, 8[9] Piotr Doll´ar、Christian Wojek、Bernt Schiele 和Pietro Per-ona。行人检测：一个基准。在2009年 IEEE计算机视觉与模式识别会议，第304–311页。IEEE，2009年。2, 6, 7[10] Ross Girshick。Fast r-cnn。在 IEEE国际计算机视觉会议论文集，第1440 1448页，2015年4 – 。[11] Kaiming He、Haoqi Fan、Yuxin Wu、SainingXie和 Ross Girshick。用于无监督视觉表征学习的动量对比。在IEEE/CVF计算机视觉与模式识别会议论文集，第9729–9738页，2020年。6[12] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和Jian Sun。用于图像识别的深度残差学习。在 IEEE计算机视觉与模式识别会议论文集，第770–778页，2016年。4, 6[13] YeHe、Chao Zhu和Xu-Cheng Yin。通过基于分布的互监督特征学习进行遮挡行人检测。IEEE 智能交通系统汇刊，2021年。2, 3, 8

[14] Xin Huang、Zheng Ge、Zequn Jie 和 Osamu Yoshie。基于代表性区域的 NMS：通过候选框配对实现拥挤场景行人检测。在 IEEE/CVF计算机视觉与模式识别会议论文集，第 10750–10759 页，2020年。2、3、7、8[15] Hangzhi Jiang、Shengcai Liao、Jinpeng Li、V´eronique Prinet 和 Shiming Xiang。基于城市场景的行人检测语义调制。神经计算，474 12，2022年 1 2、3、8：–。[16] Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。在第三届国际学习表征会议（ICLR 2015）论文集，2015年 5 月 7–9 日，美国加利福尼亚州圣迭戈，会议轨道论文集，2015年。6[17] Boyi Li、Kilian Q. Weinberger、Serge J. Belongie、Vladlen Koltun 和 Ren´e Ranftl。语言驱动的语义分割。在第十届国际学习表征会议（ICLR 2022）论文集，2022年 29，2022年 25 3、4，虚拟会议，4 月 -。[18] Jinpeng Li、Shengcai Liao、Hangzhi Jiang 和 Ling Shao。框引导卷积用于行人检测。在第 28 届 ACM 国际多媒体会议论文集，第 1615–1624 页，2020年。3、8[19] Qiming Li、Yuquan Bi、Rongsheng Cai 和 Jun Li。通过无锚网络中的双中心预测进行遮挡行人检测。神经计算，507：199–207，2022年。2、3[20] Qiming Li、Yijing Su、Yin Gao、Feng Xie 和 Jun Li。Oaf-net：面向人群场景行人检测的遮挡感知无锚网络。IEEE 智能交通系统汇刊，2022年。2、3[21]Chunze Lin、Jiwen Lu、Gang Wang 和 Jie Zhou。面向粒度感知的行人检测深度特征学习。在欧洲计算机视觉会议（ECCV）论文集，第 732–747 页，2018年。2、3[22] Tsung-Yi Lin、Piotr Doll´ar、Ross Girshick、Kaiming He、Bharath Hariharan 和 Serge Belongie。用于目标检测的特征金字塔网络。在 IEEE 计算机视觉与模式识别会议论文集，第2117–2125 页，2017年。3[23] Zebin Lin、Wenjie Pei、Fanglin Chen、David Zhang 和 Guangming Lu。通过示例引导的对比学习进行行人检测。IEEE 图像处理汇刊，2022年。2、3、4、8[24] Mengyin Liu、Chao Zhu、Jun Wang 和 Xu-Cheng Yin。用于鲁棒行人检测的自适应模式参数匹配。在 AAAI 人工智能会议论文集，第 35 卷，第2154–2162 页，2021年。2、3、8[25] Songtao Liu、Di Huang 和 Yunhong Wang。自适应非极大值抑制：改进拥挤场景中的行人检测。在 IEEE/CVF计算机视觉与模式识别会议论文集，第 6459–6468 页，2019年。2、3、7、8[26] Wei Liu、Shengcai Liao、Weidong Hu、Xuezhi Liang 和 Xiao Chen。通过渐近定位拟合学习高效的单阶段行人检测器。在欧洲计算机视觉会议（ECCV）论文集，第 618–634 页，2018年。2、3、7、8

[27] Wei Liu、Shengcai Liao、Weiqiang Ren、Weidong Hu 和 Yinan Yu。高级语义特征检测：行人检测的新视角。在 IEEE 计算机视觉与模式识别会议论文集，第 5187–5196 页，2019年。1、2、3、4、5、6、7、8[28] Ruiqi Lu、Huimin Ma 和 Yu Wang。人群场景中的语义头增强行人检测。神经计算，400：343–351，2020年。2、3、7、8[29] Zekun Luo、Zheng Fang、Sixiao Zheng、Yabiao Wang 和 Yanwei Fu。NMS损失：利用非极大值抑制进行拥挤场景行人检测学习。在 2021年国际多媒体检索会议论文集，第 481–485 页，2021年。3、7、8[30] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga 等。PyTorch：一种命令式、高性能深度学习库。神经信息处理系统进展，32，2019年。6[31]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。在国际机器学习会议论文集，第 8748–8763 页。机器学习研究论文集，2021年。2、3、4、5、6[32] Yongming Rao、Wenliang Zhao、Guangyi Chen、Yansong Tang、Zheng Zhu、Guan Huang、Jie Zhou 和 Jiwen Lu。DenseCLIP：利用上下文感知提示进行语言引导的密集预测。在 IEEE/CVF计算机视觉与模式识别会议论文集，第 18082–18091 页，2022年。2、3、4、5[33] Xiaolin Song、Binghui Chen、Pengyu Li、Biao Wang 和 Honggang Zhang。Prnet[33] ：通过渐进式细化网络实现通用遮挡行人检测。在神经计算，482：98–115，2022年。2、3、7、8[34] Fiseha B Tesema、Hong Wu、Mingjian Chen、Junpeng Lin、William Zhu 和 Kaizhu Huang。基于混合通道的行人检测。神经计算，389：1–8，2020年。2[35] Wenguan Wang、Tianfei Zhou、Fisher Yu、Jifeng Dai、Ender Konukoglu 和 Luc Van Gool。探索用于语义分割的跨图像像素对比。在 IEEE/CVF 国际计算机视觉会议论文集，第 7303–7313 页，2021年。2、3、5[36] Xinlong Wang、Tete Xiao、Yuning Jiang、Shuai Shao、Jian Sun 和 Chunhua Shen。排斥损失：检测人群中的行人。在 IEEE 计算机视觉与模式识别会议论文集，第 7774–7783 页，2018年。3、7、8[37] Xinlong Wang、Rufeng Zhang、Chunhua Shen、Tao Kong 和 Lei Li。用于自监督视觉预训练的密集对比学习。在 IEEE/CVF计算机视觉与模式识别会议论文集，第 3024–3033 页，2021年。3、5[38] Yang Wang、Chong Han、Guangle Yao 和 Wanlin Zhou。MAPD：改进的人群场景多属性行人检测。神经计算，432：101–110，2021年。2、3、7、8

[39] Jialian Wu、Chunluan Zhou、Qian Zhang、Ming Yang 和 Junsong Yuan。用于小尺度行人检测的自模仿学习。载于第 28 届 ACM 多媒体国际会议论文集，第 2012–2020 页，2020年。2, 3, 7, 8[40] Jin Xie、Hisham Cholakkal、Rao Muhammad Anwer、Fahad Shahbaz Khan、Yanwei Pang、Ling Shao 和 Mubarak Shah。用于行人检测的计数与相似性感知 R-CNN。载于欧洲计算机视觉会议论文集，第 88–104 页。Springer，2020年。2, 3, 7, 8[41]Jin Xie、Yanwei Pang、Muhammad Haris Khan、Rao Muhammad Anwer、Fahad Shahbaz Khan 和 Ling Shao。用于遮挡行人检测的掩码引导注意力网络与遮挡敏感型困难样本挖掘。《IEEE 图像处理汇刊》，30：3872–3884，2020年。2, 3, 4[42] Shanshan Zhang、Rodrigo Benenson、Mohamed Omran、Jan Hosang 和 Bernt Schiele。我们距离解决行人检测问题还有多远？载于 IEEE 计算机视觉与模式识别会议论文集，第 1259–1267 页，2016年。6[43] Shanshan Zhang、Rodrigo Benenson 和 Bernt Schiele。CityPersons：用于行人检测的多样化数据集。载于 IEEE 计算机视觉与模式识别会议论文集，第 3213–3221 页，2017年。2, 5, 6[44] Shifeng Zhang、Longyin Wen、Xiao Bian、Zhen Lei 和 Stan Z Li。感知遮挡的 R-CNN：在人群中检测行人。载于欧洲计算机视觉会议（ECCV）论文集，第

637 653, 2018 2, 3, 7, 8 – 页。

[45] Shanshan Zhang、Jian Yang 和 Bernt Schiele。通过 CNN 中的引导注意力检测遮挡行人。载于 IEEE 计算机视觉与模式识别会议论文集，第 6995–7003 页，2018年。2[46] Shan Zhang、Xiaoshan Yang、Yanxia Liu 和 Changsheng Xu。用于小尺度行人检测的非对称多阶段 CNN。神经计算，409：12–26，2020年。2, 3, 7, 8[47] Tianliang Zhang、Qixiang Ye、Baochang Zhang、Jianzhuang Liu、Xiaopeng Zhang 和 Qi Tian。用于遮挡行人检测的特征校准网络。《IEEE 智能交通系统汇刊》，2020年。2, 3, 4, 8[48] Anlin Zheng、Yuang Zhang、Xiangyu Zhang、Xiaojuan Qi 和 Jian Sun。拥挤场景中的渐进式端到端目标检测。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 857–866 页，2022年。2, 3[49] Chong Zhou、Chen Change Loy 和 Bo Dai。从 CLIP 中提取自由密集标签。载于欧洲计算机视觉会议论文集，第 696–712 页。Springer，2022年。2, 4, 5[50] Penghao Zhou、Chong Zhou、Pai Peng、Junlong Du、Xing Sun、Xiaowei Guo 和 Feiyue Huang。Noh-NMS：通过幻化附近物体改进行人检测。载于第 28 届 ACM 多媒体国际会议论文集，第 1967–1975 页，2020年。2[51] Tianfei Zhou、Wenguan Wang、Ender Konukoglu 和 Luc Van Gool。重新思考语义分割：原型视角。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 2582–2593 页，2022年。2, 3, 5