VLPD: Context-Aware Pedestrian Detection via Vision-Language Semantic Self-Supervision

Mengyin Liu[1*] Jie Jiang[2*] Chao Zhu[1†] Xu-Cheng Yin[1]

[1]School of Computer and Communication Engineering,
University of Science and Technology Beijing, Beijing, China
[2]Data Platform Department, Tencent, Shenzhen, China

blean@live.cn, zeus@tencent.com, {chaozhu, xuchengyin}@ustb.edu.cn

Abstract

Detecting pedestrians accurately in urban scenes is significant for realistic applications like autonomous driving or video surveillance. However, confusing human-like objects often lead to wrong detections, and small scale or heavily occluded pedestrians are easily missed due to their unusual appearances. To address these challenges, only object regions are inadequate, thus how to fully utilize more explicit and semantic contexts becomes a key problem. Meanwhile, previous context-aware pedestrian detectors either only learn latent contexts with visual clues, or need laborious annotations to obtain explicit and semantic contexts. Therefore, we propose in this paper a novel approach via Vision-Language semantic self-supervision for context-aware Pedestrian Detection (VLPD) to model explicitly semantic contexts without any extra annotations. Firstly, we propose a self-supervised Vision-Language Semantic (VLS) segmentation method, which learns both fully-supervised pedestrian detection and contextual segmentation via self-generated explicit labels of semantic classes by vision-language models. Furthermore, a self-supervised Prototypical Semantic Contrastive (PSC) learning method is proposed to better discriminate pedestrians and other classes, based on more explicit and semantic contexts obtained from VLS. Extensive experiments on popular benchmarks show that our proposed VLPD achieves superior performances over the previous state-of-the-arts, particularly under challenging circumstances like small scale and heavy occlusion. Code is available at https://github.com/lmy98129/VLPD.

1. Introduction

With the recent advances of pedestrian detection, enormous applications benefit from such a fundamental perception technique, including person re-identification, video surveillance and autonomous driving. In the meantime, various challenges from the urban contexts, i.e., pedestrians and non-human objects, still hinder the better performances of detection. For example, confusing appearances of human-like objects often mislead the detector, as shown in Figure 1(a). Moreover, heavily occluded or small scale pedestrians have unusual appearances and cause missing detections as Figure 1(a) and (b). Apart from the object regions, the contexts are crucial to address these challenges.

Nevertheless, previous methods still make inadequate investigations on the contexts in urban scenarios. For in-

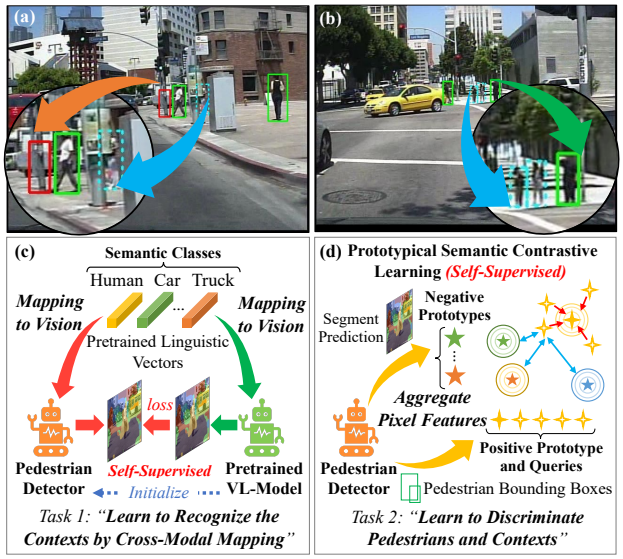

Figure 1. Illustration of the problems by previous works (top) and our proposed method to tackle them (bottom). (a) and (b) are predicted by [27]. Green boxes are correct, red ones are human-like traffic signs, and dashed blue ones are missing heavily occluded or small scale pedestrians. (c) and (d): We propose self-supervisions to recognize the contexts and discriminate them from pedestrians.

∗ Equal contribution. † Corresponding author.

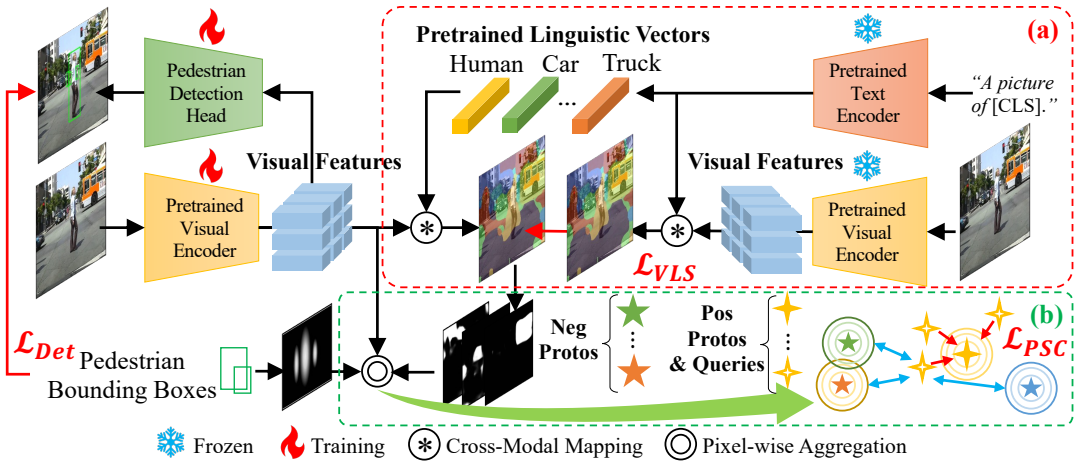

Figure 2. The overall architecture of our proposed VLPD approach. (a) Vision-Language Semantic (VLS) segmentation obtains pseudo labels via Cross-Modal Mapping, then the Pretrained Visual Encoder learns fully-supervised detection ($\mathcal{L}_{Det}$) and self-supervised segmentation to recognize semantic classes for explicit contexts without any annotations. (b) Prototypical Semantic Contrastive (PSC) learning lets the pixel-wise pedestrian features as queries closer to positive prototypes and further to negative ones based on Pixel-wise Aggregation.

stance, manual contextual annotations from CityScapes [6] boost SMPD [15] on the pedestrian benchmark CityPersons [43], because they share homologous image data. Besides, a semi-supervised model yields pseudo labels for the Caltech dataset [9]. However, both these two solutions require expensive fine-grained annotations, especially for training the semi-supervised model. Moreover, other methods learn regional latent contexts merely from limited visual neighborhood [47], or non-human local proposals as negative samples for contrastive learning [23]. Without an explicit awareness of semantic classes in the contexts, these methods thus still suffer from unsatisfactory performance.

Besides, some pedestrian detection methods also indirectly handle the contexts. For the occlusion problems, many part-aware methods [4, 13, 19, 20, 28, 33, 41, 44, 45] adopt visible annotations for the occluded pedestrians, which indicate the occlusion by other pedestrians or non-human objects in the contexts. Whereas, these labels still need heavy labors of human annotators. For scale variation [2, 8, 21, 39, 46], crowd occlusion [14, 25, 38, 40, 48, 50] or generic hard pedestrians [1, 24, 26, 27, 34], most previous works are intra-class, e.g., small pedestrians or crowded scenes, and thus irrelevant to context modeling problems.

Inspired by the vision-language models, we notice a more explicit context modeling without any annotations via cross-modal mapping. For instance, DenseCLIP [32] is initialized with vision-language pretrained CLIP model [31] to learn cross-modal mapping from pixel-wise features to linguistic vectors of human-annotated classes. Meanwhile, MaskCLIP [49] generates pseudo labels via cross-modal mapping and train another visual model. Hence, complementing the initialized mapping and pseudo labeling, we propose to recognize the semantic classes for explicit contexts via self-supervised Vision-Language Semantic (VLS) segmentation, as shown in Figure 1(c) and 2(a).

Furthermore, we consider that only pixel-wise scores are ambiguous to discriminate pedestrians and contexts. Due to the coarse-grained pseudo labels, some parts of pedestrians might have higher scores of other classes. Different from the regional contrastive learning [23], we introduce the concept of prototype [35, 51] for a global discrimination. Each pixel of pedestrian features is pulled closer to pixel-wise aggregated positive prototypes and pushed away from the negative ones of other classes based on the explicit contexts obtained from VLS. As illustrated in Figure 1(d) and 2(b), a novel contrastive self-supervision for pedestrian detection is proposed to better discriminate pedestrians and contexts.

In conclusion, we have observed a dilemma between the heavy burden of manual annotation for explicit contexts and local implicit context modeling. Hence, we propose a novel approach to tackle these problems via **V**ision-**L**anguage semantic self-supervision for **P**edestrian **D**etection (**VLPD**). The main contributions of this paper are as follows:

- Firstly, the Vision-Language Semantic (VLS) segmentation method is proposed to model explicit semantic contexts by vision-language models. With pseudo labels via cross-modal mapping, the visual encoder learns fully-supervised detection and self-supervised segmentation to recognize the semantic classes for explicit contexts. **To our best knowledge, this is the first work to propose such a vision-language extra-annotation-free method for pedestrian detection**.
- Secondly, we further propose the Prototypical Semantic Contrastive (PSC) learning method to better discriminate pedestrians and contexts. The negative and

positive prototypes are aggregated via the score maps of contextual semantic classes obtained from VLS and pedestrian bounding boxes, respectively. Each pixel of pedestrian features is pulled close to positive prototypes and pushed away from the negative ones, in order to strengthen the discrimination power of the detector.

- Finally, by the integration of VLS and PSC, our proposed approach VLPD achieves superior performances over the previous state-of-the-art methods on popular Caltech and CityPersons benchmarks, especially on the challenging small scale and occlusion subsets.

2. Related Works

2.1. Pedestrian Detection

In realistic applications, various circumstances are challenging for pedestrian detection, including occlusion, scale-variation and generic hard pedestrian handling. Here, we discuss these common problems as well as the context-aware methods which are specialized for these problems.

2.1.1 Occlusion Handling

As a research hot-spot of pedestrian detection, handling occlusion should make the best of limited information from visible parts of pedestrian, and also avoid the noisy one from occlusion by other pedestrians or non-human objects.

On the one hand, part-aware methods handle the visible parts with other parts occluded by contextual objects. For example, OR-CNN [44] re-scores parts to highlight the visible ones. PRNet++ [33] progressively refines the predicted visible and full-body boxes. Extra labels of less frequently occluded heads facilitate HBAN [28], JointDet [5] and PedHunter [4]. Some methods [2, 19, 20] handle the visible and full bodies by parallel branches. Moreover, DMSFLN [13] explores the feature distributions between both branches.

On the other hand, crowd-aware methods are specialized to intra-class occlusion without context modeling. Some post-processing methods [14, 25, 29] focus on the over-suppression of dense predictions in crowd scenes, and the others [48] handle the under-suppression of sparse ones. For the heavily overlapped pedestrians, loss-based methods [36, 38, 40] identify them by learning representations.

Differently, our proposed VLPD discriminates non-human occluders and pedestrians via the contrastive learning of PSC, on the basis of the self-supervised learning via self-generated explicit labels of semantic classes from VLS.

2.1.2 Scale-Variation Handling

Scale-variation is another problem that potentially related to modeling the context. Due to the distance, the blurry and noisy appearances of both small pedestrians and non-human objects often confuse the detectors. Multiple branches [8, 21, 46] are popular for modeling different scales. With the powerful FPN [22] architecture, LBST [2] detects smaller pedestrians with the fusion of bottom-up and top-down features. Differently, SML [39] pushes the features of small-scaled pedestrians towards the distribution of large ones.

Unfortunately, these works focus on the small-scale pedestrians, due to no labels for small non-human objects. Hence, our proposed VLPD uses label-free explicit contexts including the latter and performs better on different scales.

2.1.3 Generic Hard Pedestrian Handling

The central issue to handle hard pedestrians is accurate localization. Plenty of previous works [1, 26] introduce multi-phase spatial refinements. Following the anchor-free style from generic object detection, the CSP [18, 27] series decrease hyper-parameters with an adaptive prediction. AP^{2}M [24] matches proper parameters for different hard samples.

For the context modeling, SMPD [15] adopts extra segmentation annotations, EGCL [23] uses contrastive learning by local proposals, and FC-Net [47] learns latent features of the local contexts. Without any extra labels, our proposed VLPD can recognize explicit contextual objects via pseudo labels of VLS, and then discriminates them with pedestrians via more global positive and negative prototypes of PSC.

2.2. Segmentation by Vision-Language Pretraining

Recent progress of vision-language pretraining CLIP [31] has facilitated more powerful segmentation methods. For example, cosine similarity, i.e., cross-modal mapping, is calculated between visual features and linguistic vectors to obtain segmentation results. DenseCLIP [32] and LSeg [17] initialize the model with a pretrained CLIP visual encoder, and then learn mapping features to annotated classes via linguistic vectors. For self-supervision, MaskCLIP [41] obtains the pseudo labels via the mapping and learns a new vision model, which is evaluated to be sub-optimal by [32].

Differently, complementing the cross-modal mapping and pseudo labeling, our proposed novel self-supervised VLS recognizes semantic classes as explicit contexts without any extra labels for context-aware pedestrian detection.

2.3. Prototypical Contrastive Learning

Due to the spatial resolution of images, purely pixel-wise dense contrastive learning [37] leads to heavy computational burden, and only discriminates locally regardless of the global image. Hence, previous works introduce "Prototypes" [35, 51] as the alternatives for pixel features of each semantic class. Differently, our proposed PSC maintain the pixels of pedestrians as queries to keep their inner variance, which learns better discrimination between pedestrians and other classes for the contextual-aware pedestrian detection.

3. Proposed Method

As illustrated in Figure 2 and 3, our proposed approach **V**ision-**L**anguage semantic self-supervision for **P**edestrian **D**etection (**VLPD**) is an anchor-free detection framework following the baseline CSP [27]. A pretrained visual encoder extracts features at different stages from S3 to S5. As shown in Figure 3, they are concatenated into "Detection Features" for the Detection Head to make predictions.

To achieve the explicit semantic context modeling without any extra labels, our architecture comprises two key components: Vision-Language Semantic (VLS) segmentation and Prototypical Semantic Contrastive (PSC) learning. VLS leverages vision-language models to recognize the explicit contexts, where the visual encoder learns both fully-supervised pedestrian detection and segmentation via self-generated explicit labels of semantic classes by cross-modal mapping. PSC supervises the detector to better discriminate pedestrians and contextual semantic classes based on VLS. More details will be introduced in the following sections.

3.1. Vision-Language Semantic Segmentation

Benefiting from self-supervised cross-modal contrastive learning, vision-language models map the visual and linguistic vectors with similar meanings closer to each other into a unified feature space. Thus, it is possible to obtain the existences of semantic classes in an image via linguistic vectors. However, previous works initialize a model with cross-modal mapping for full-supervision [17, 32], or re-train a new model by pseudo labels via the mapping [41].

Therefore, as shown in Figure 2(a), we propose Vision-Language Semantic (VLS) segmentation as the complement of both initialized mapping and pseudo labeling. Labels are generated by frozen pretrained models based on cross-modal mapping, thus the unfrozen visual encoder is supervised to predict the segmentation of explicit semantic classes, which serve as more global contexts rather than previous local latent ones [23, 47]. More details of our proposed VLS will be provided in the following sections.

3.1.1 Cross-Modal Mapping for Pseudo Labeling

As one of the most popular vision-language models, CLIP [31] is capable of mapping image and text with similar meanings into closer vectors, based on its visual and linguistic encoders pretrained by self-supervised cross-modal contrastive learning. Although it is impossible to recover pixel-wise contextual information of an image from its visual vector after an attention-based pooling of CLIP [31], this pooling operation can be modified into projections to keep visual regions, following [49]. As shown in Figure 4, cosine similarities $S_i^c = ((L^c)^\top(V_i))/(\|L^c\|\|V_i\|) \in S$ are calculated between pretrained linguistic vector $L^c \in \mathbb{R}^{\mathrm{D}'}$ of each class $c \in C$ and the projected vision features $V_i \in \mathbb{R}^{\mathrm{D}'}$

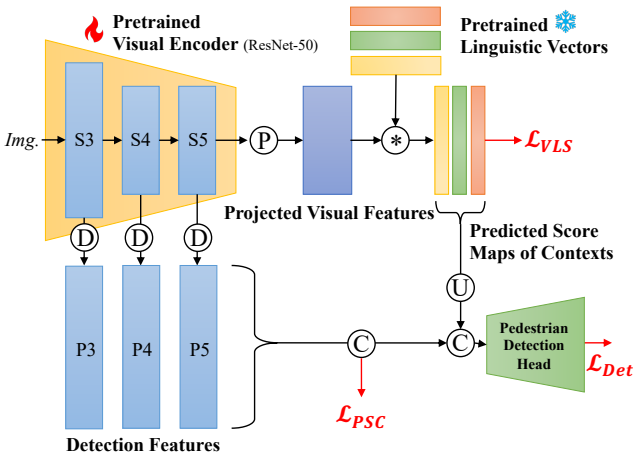

Figure 3. The detailed network architectures of our proposed VLPD. The visual encoder is ResNet-50 [12] from the vision-language pretrained CLIP model [31]. Following the baseline CSP [27], S3∼S5 are Deconvolved ("D") as "Detection Features" for $\mathcal{L}_{Det}$, which are further supervised by our $\mathcal{L}_{PSC}$. The Projections ("P") before cross-modal mapping are adopted like [49]. Concatenation and Up-sampling ("C" and "U") are used for prediction as [32]. These details are omitted in Figure 2 for simplicity.

of each pixel $i = 1, 2, ..., \mathrm{H}'\mathrm{W}'$, which means the existence of each contextual class at each pixel of an image.

To obtain the pseudo labels for self-supervised learning, as illustrated in Figure 2(a), images are feed into the frozen CLIP visual encoder to obtain visual features, and the linguistic vectors of classes are generated by frozen text encoder via the prompted sentence "*A picture of* [CLS]".

3.1.2 Self-Supervised Learning for VLS

Evaluated by the experiments of [32], initialization with CLIP pretrained visual encoder contributes the maintenance of Cross-Modal Mapping, which allows the model to fully utilize the similarity between each pixel and each linguistic vector. While the ImageNet [7] pretrained one as [49] needs re-training for an adaption to the mapping and thus is sub-optimal. Therefore, we propose to embrace both the advantage of both pseudo labeling and initialized mapping.

In details, as shown in the Figure 2(a) and 3, the visual encoder of pedestrian detector is initialized by CLIP pretrained parameters, which is identical to the frozen visual encoder for pseudo labeling as a self-supervision. Since the Cross-Modal Mapping is also performed, the linguistic vectors can be generated just once. The predicted $\bar{S} \in \mathbb{R}^{\mathrm{H}'\times\mathrm{W}'\times\mathrm{N}}$ are supervised by pseudo labels S based on Smooth L1 Loss [10] $\mathcal{L}_{VLS}$ that is robust to noisy labels:

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

where class count $\mathrm{N} = |C|$. More than the self-supervision

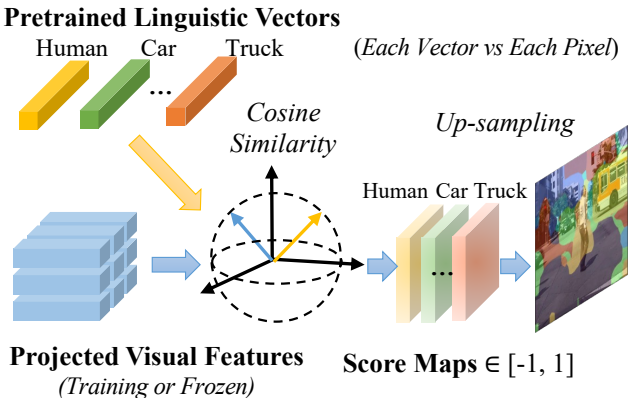

Figure 4. Cross-Modal Mapping for our proposed VLS. Visual features are projected as [49] rather than a global vector in CLIP pretraining [31]. Cosine similarity is computed between each linguistic vector of semantic classes and each pixel of visual features, which constitutes the score maps of these classes in contexts.

in Eq.1, the visual encoder is also trained to detect pedestrians as the baseline [27]. Consequently, the model learns to model the contexts during detecting within a unified pipeline. The predicted $\bar{S}$ are fed into the Detection Head in Figure 3 for an explicit contextual reference like [32].

3.1.3 Compacted Class Policy

The complicated environments in the urban scenarios for pedestrian detection require a proper contextual class set. Meanwhile, urban dataset CityScapes [6] already defines various classes. However, experiments (in the following sections) reveal a negative effect of it. Inspired by imbalanced class frequency statistics in the CityScapes paper [6], we propose Compacted Class Policy in Table 1 to decide whether classes should be kept, compacted or discarded.

For the most classes, we adopt the 2^{nd} level of classes in CityScapes higher than the original ones, as illustrated in Table 1. Furthermore, the higher variance inside the "vehicle" class leads to performance loss in experimental trials. Thus, neither the 1^{st} nor 2^{nd} level of classes is applicable.

Therefore, we make statistics on frequencies in pixel-wise annotations of CityScapes for the images shared with CityPersons [43], which are computed via not only pixel-wise counting but also re-weighting by image-wise occurrence times. Less frequent tail classes are omitted by thresholding and the head ones are kept as the bottom of Table 1.

In conclusion, our proposed VLS leverages the powerful vision-language model CLIP [31] to perform self-supervision based on Cross-Modal Mapping and Compacted Class Policy, which obtains pseudo labels of semantic classes for explicit contexts to learn recognizing them during detection for better discriminations.

Table 1. Compacted Class Policy for our proposed VLS.

Original →	Compacted	Used
{road, sidewalk}	ground	✓
{building, wall, fence}	building	✓
{vegetation, terrain}	tree	✓
{person, rider}	human	✓
{pole, traffic light, traffic sign}	traffic sign	✓
{car, bicycle, bus, truck, motorcycle, train}	vehicle	×
	{car, bicycle, bus, truck}	✓

3.2. Prototypical Semantic Contrastive Learning

Due to the coarse-grained characteristics of pseudo labels by our proposed VLS, some visible parts of the pedestrians might have higher scores of other classes from VLS, which are annotated by the bounding boxes for the detection tasks. Since there are no manual annotations available, explicit refinement to the pseudo labels is rather difficult.

Inspired by self-supervised contrastive learning [35, 37, 51] for discriminative representations of positive and negative samples, we introduce this powerful technique and propose a novel Prototypical Semantic Contrastive (PSC) learning, which learns a better discrimination of pedestrians and other semantic classes without any extra labels.

In order to decrease the heavy computations by dense pixel-wise methods [37], the concept of "Prototype" is embraced. It means a representative feature which represents all the features belong to same semantic class. In this paper, prototype of pedestrian is positive, and others are negative.

Take the negative prototypes as examples. The predicted score maps $\bar{S} \in \mathbb{R}^{\text{H}' \times \text{W}' \times (\text{N}-1)}$ are adopted as the indicator of the spatial existences of all the non-human $\text{N}-1$ classes, except the "Human" which is overlapped with pedestrian. Here, we denote C as non-human classes for simplicity, where $|C| = \text{N} - 1$. $\dot{S} \in \mathbb{R}^{\text{H} \times \text{W} \times (\text{N}-1)}$ are up-sampled from $\bar{S}$. A SoftMax function δ with a temperature τ' is applied to normalize $\dot{S} \in [-1, 1]$ into $\hat{S} \in [0, 1]$:

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

Since we should not disturb the self-supervised learning of VLS and only aim to improve the detection, "Detection Features" $E \in \mathbb{R}^{\text{D} \times \text{H} \times \text{W}}$ are supervised. As shown in Figure 5, prototypes of each class are obtained via aggregating E pixel-wisely by $\hat{S}$, denoted as Pixel-wise Aggregation:

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

where $\cdot$ is matrix multiplication, and $P^- \in \mathbb{R}^{\text{D} \times (\text{N}-1)}$ are

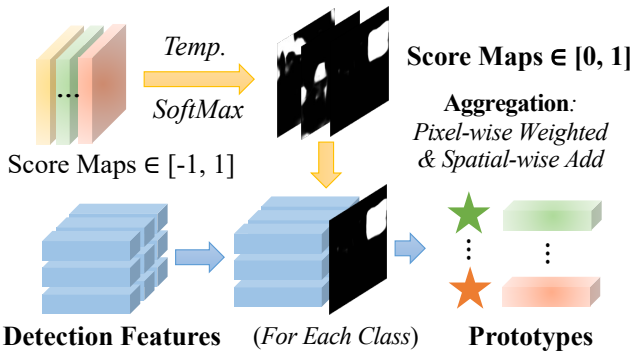

Figure 5. Pixel-wise Aggregation for our proposed PSC. "Detection Features" are pixel-wise weighted by predicted score maps of VLS and then spatially added into prototypes as an aggregation.

negative prototypes of $\mathrm{N}-1$ non-human classes with channels D. Weighted by $\hat{S}_i^c \in \mathbb{R}$, feature $E_i \in \mathbb{R}^D$ at each position $i = 1, 2, ..., \mathrm{HW}$ is aggregated. Similarly, 2D Gaussians of pedestrian positions $G \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 1}$ from the baseline CSP [27] can replace the $\hat{S}$ for positive prototype P^+.

Finally, each pixel-wise E_j of the annotated pedestrians is supervised as "query" of contrastive loss function $\mathcal{L}_{PSC}$ in Figure 1(d) and 2(b), which pulls them close to P^+ and pushes them away from each P^{c-}. E_j is located by the > 0 positions j of G and $|G^{>0}| = \mathrm{M}$. $\mathcal{L}_{PSC}$ is formulated as:

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

where self-normalization of the features and prototypes, e.g., $E_j/\|E_j\|$, is omitted for simplicity. Similar to some contrastive pretraining methods [3, 11], negative prototypes are expanded to all images $b \in B$ inside the mini-batch B.

In brief, based on the contrastive self-supervision, PSC trains the detector to better discriminate the pedestrians and other classes via the positive and negative prototypes, owing to the explicit and semantic contexts obtained from VLS. By the integration of the VLS and PSC, our proposed VLPD is supervised by $\mathcal{L}_{Det}$ [27], $\mathcal{L}_{VLS}$ and $\mathcal{L}_{PSC}$ simultaneously:

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

3.3. Detection Head

Following the anchor-free style of our baseline CSP [27], Detection Head in Figure 2 and 3 firstly decreases the channels of "Detection Features" via convolution layers. Then multiple branches predict result maps: "Center Heatmap" to classify the centers of pedestrians, "Scale Map" to predict heights with a fixed aspect ratio 0.41 for widths, and "Offset Map" to adjust the localization horizontally and vertically. Finally, these maps are assembled into bounding boxes of pedestrians. More details can be found in the CSP paper.

4. Experiments

In this section, extensive experiments are conducted on two popular benchmarks for pedestrian detection, i.e., Caltech and CityPersons, to evaluate our proposed VLPD method. Ablation study is performed on key components VLS and PSC. Furthermore, we also report the state-of-the-art comparisons on both benchmarks. For more experiments and visualizations, please refer to supplementary materials.

4.1. Datasets

The Caltech pedestrian dataset [9] comprises 2.5-hour video data captured on the urban areas of Los Angeles, with 4024 images for testing. Over 70% of pedestrians are less than 100 pixels high, including particularly small pedestrians that are less than 50 pixels. By fixing the inconsistency and box misalignment, Zhang et. al. [42] has released a new version of the annotations. For fair comparisons, all the following evaluations are performed based on the new version.

CityPersons [43] is a more recently published large-scale pedestrian detection dataset. 2975 images are split for training and 500 images for validation. The standard evaluation metric is as follows: log miss rate is averaged over the false positive per image (FPPI) $\in [10^{-2}; 100]$, denoted as MR^{-2}. All tests are applied on the original data ($1\times$) without resizing and any extra visible or head labels for fair comparisons.

4.2. Implementation Details

Our proposed method is based on a powerful pedestrian detector CSP [27], which is re-implemented on PyTorch [30] framework from the original Keras one. Adam [16] is adopted for optimization. The backbone network is ResNet-50 [12] pretrained on ImageNet [7] by fully-supervised image classification or WIT [31] by self-supervised vision-language contrastive learning. For Caltech, one Nvidia 3090 GPU is utilized for training with 10^{-4} learning rate. For CityPersons, two 3090 GPUs are used with 2×10^{-4}. Batch sizes are set following [27]. All tests are conducted on a single 3090 GPU. The size of training images is 336×448 for Caltech and 640×1280 for CityPersons. For our proposed VLS, its loss weight $\lambda_1 = 100$. For PSC, its weight $\lambda_2 = 10^{-4}$ for Caltech and 10^{-3} for CityPersons. Temperatures $\tau' = 10^{-3}$ and $\tau = 7\times10^{-2}$ following [11].

4.3. Ablation Study

The ablation study is firstly performed on the popular CityPersons dataset. Comprehensive subset Reasonable, more challenging ones Small and HO (Heavy Occlusion, visible rate $\in [0.2, 0.65]$) are widely-used for comparisons.

Table 2 illustrates the overall ablation study for each key components of our proposed VLPD. We provide the original results of CSP [27] and our re-implemented one by PyTorch [30]. Under the CLIP [31] initialized visual encoder

Figure 6. Qualitative analysis on Caltech [9] between the baseline CSP [27] (top) and our proposed VLPD (bottom). Green are correct detections, red are wrong detections, and dashed blue are missing detections. With the powerful vision-language semantic self-supervision, our proposed VLPD is context-aware and thus more robust to human-like objects, inter-class occlusion and ambiguous small pedestrians.

Table 2. Overall ablation study for key components of our proposed VLPD, including VLS and PSC. **Bolden** are the best results.

Method	Reasonable	Small	HO
CSP [27]	11.0	16.0	-
CSP (our re-imp.)	10.96	16.05	40.59
CSP w/ CLIP	10.13	12.59	38.97
+VLS	9.70	12.57	36.50
+VLS+PSC=**VLPD**	**9.41**	**10.93**	**34.88**

as a precondition of VLS, the improvements are limited because merely vision-language pretraining cannot fully handle the context modeling. With VLS as well as PSC based on VLS, our proposed VLPD gains significant boosts especially on the context-related subsets Small and HO.

In Table 3, different policies of the class set for our proposed VLS are evaluated. Full CityScapes policy adopts all the classes of the CityScapes dataset [6], and Full Compacted uses 2^{nd} level of classes. Both the too scattered and concentrated sets lead to performance losses. Instead, our proposed policy for VLS handles the largest "vehicle" classes via frequency statistics for better context modeling.

Meanwhile, sub-items of this policy are evaluated at the bottom of Table 3 via recovering to the 1^{st} column of Table 1. Table 4 shows that our PSC in Eq.4 with cross-image negatives and inner-image positives of "Detection Features" E without disturbing the $\dot{S}$ from VLS performs the best.

4.4. Comparisons with the State-of-the-arts

For CityPersons, we compare our proposed VLPD with various state-of-the-art methods: AMSCNN [46], DHRNet [8] and SML [39] for scale-variation; RepLoss [36], Adaptive NMS [25], PBM+R^{2}NMS [14], CaSe [40], NMS-Ped [29] and MAPD [38] for intra-class occlusion; OR-CNN [44], HBAN [28] and PRNet++ [33] for part-aware occlusion handling; LBST [2], ALFNet [26], CSP [27], AP^{2}M

Table 3. Different policies of the class set for our proposed VLS.

Method	Reasonable	Small	HO
CSP w/ CLIP	10.13	12.59	38.97
+ Full CityScapes	10.40	12.87	37.14
+ Full Compacted	10.47	13.30	40.49
+ VLS (ours)	**9.70**	12.57	**36.50**
w/o ground	10.51	13.53	37.48
w/o building	10.42	12.84	37.70
w/o tree	10.34	12.66	38.39
w/o human	10.24	12.52	38.47
w/o {car, bicycle, bus, truck}	10.61	13.61	38.84
w/o traffic sign	10.11	**12.27**	37.22

Table 4. Different prototypes and features for our proposed PSC.

Method	Reasonable	Small	HO
VLPD (w/ PSC, ours)	**9.41**	**10.93**	**34.88**
Cross $\rightarrow$ Inner-img Neg.	10.12	12.67	37.90
Inner $\rightarrow$ Cross-img Pos.	10.58	13.14	38.08
$E \rightarrow \text{Concate}(E, \dot{S})$	10.04	12.93	38.21

Table 5. Comparison with the state-of-the-arts on CityPersons.

Methods	R	Hea.	Partial	Bare	Small
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 (ours)	**9.4**	**43.1**	**8.8**	**6.1**	**10.9**

[24] and BGCNet [18] for generic hard pedestrian detection. Note that the context-related methods are: SMPD [15] with segmentation annotation, EGCL [23] with proposal-wise contrastive learning and FC-Net [47] with neighbor region modeling. As illustrated in Table 5, our VLPD outperforms them comprehensively among all the subsets.

In details, we denote Reasonable subset as "R" and the occlusion one Heavy (visible rate $\in [0, 0.65]$) as "Hea." in Table 5. We also compare our proposed VLPD with the methods on other occlusion subsets in Table 6, i.e., R+HO (Reasonable+HO, visible rate $\in [0.2, 1]$) and HO. Our method keeps the best under these setting changes.

For Caltech, additional state-of-the-art methods are compared: AR-Ped [1] for generic hard pedestrian handling; JointDet [5], PedHunter [4] and DMSFLN [13] with visible or head labels. In Table 7, without any extra labels, our proposed method VLPD also surpasses them significantly. Its Reasonable 2.27% is better than 2.31% of [4]. Context-related challenging subsets Heavy Occlusion 37.7% and All 52.4% are especially better than other methods.

In conclusion, our proposed VLPD has become a new state-of-the-art on both benchmarks especially in context-related subsets, which sufficiently validates its power of vision-language semantic self-supervision to explicitly model semantic contexts without any extra labels and better discriminate pedestrians from other contextual classes.

Table 6. Comparison on other occlusion subsets on CityPersons.

Methods	Reasonable	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 (ours)	**9.4**	**21.7**	**34.9**

Table 7. Comparison with the state-of-the-arts on Caltech.

Methods	Reasonable	All	Heavy
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	-	-
AP^{2}M [24]	3.3	55.9	42.2
JointDet [5]	3.0	-	-
DMSFLN [13]	2.7	-	-
PedHunter [4]	2.3	-	-
VLPD (ours)	**2.3**	**52.4**	**37.7**

5. Conclusion

In this paper, we have proposed a novel pedestrian detection method VLPD for explicit contexts modeling towards challenging problems, e.g., human-like objects and small scale or heavily occluded pedestrians. It tackles these challenges via vision-language semantic self-supervision with two key components: VLS is proposed to leverage vision-language models to recognize semantic classes for explicit contexts, which learns fully-supervised pedestrian detection and self-supervised segmentation via pseudo labels by cross-modal mapping. PSC is proposed to adopt contrastive self-supervision for better discriminating pedestrians and semantic classes based on explicit contexts from VLS. By the integration of VLS and PSC, our VLPD achieves the new cutting-edge performances on two challenging benchmarks Caltech and CityPersons, especially on the very difficult circumstances of small scale and heavy occlusion.

6. Acknowledgments

This work was supported by National Key Research and Development Program of China (2020AAA0109701), National Natural Science Foundation of China (62072032, 62076024), and National Science Fund for Distinguished Young Scholars (62125601).

References

[1] Garrick Brazil and Xiaoming Liu. Pedestrian detection with autoregressive network phases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7231–7240, 2019. 2, 3, 8

[2] Jiale Cao, Yanwei Pang, Jungong Han, Bolin Gao, and Xuelong Li. Taking a look at small-scale pedestrians and occluded pedestrians. *IEEE transactions on image processing*, 29:3143–3152, 2019. 2, 3, 7, 8

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 6

[4] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Pedhunter: Occlusion robust pedestrian detector in crowded scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10639–10646, 2020. 2, 3, 8

[5] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Relational learning for joint head and human detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10647–10654, 2020. 3, 8

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 5, 7

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 4, 6

[8] Mengyuan Ding, Shanshan Zhang, and Jian Yang. Learning a dynamic high-resolution network for multi-scale pedestrian detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9076–9082. IEEE, 2021. 2, 3, 7, 8

[9] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE conference on computer vision and pattern recognition*, pages 304–311. IEEE, 2009. 2, 6, 7

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 6

[13] Ye He, Chao Zhu, and Xu-Cheng Yin. Occluded pedestrian detection via distribution-based mutual-supervised feature learning. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 2, 3, 8

[14] Xin Huang, Zheng Ge, Zequn Jie, and Osamu Yoshie. Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10750–10759, 2020. 2, 3, 7, 8

[15] Hangzhi Jiang, Shengcai Liao, Jinpeng Li, Véronique Prinet, and Shiming Xiang. Urban scene based semantical modulation for pedestrian detection. *Neurocomputing*, 474:1–12, 2022. 2, 3, 8

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[17] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. 3, 4

[18] Jinpeng Li, Shengcai Liao, Hangzhi Jiang, and Ling Shao. Box guided convolution for pedestrian detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1615–1624, 2020. 3, 8

[19] Qiming Li, Yuquan Bi, Rongsheng Cai, and Jun Li. Occluded pedestrian detection through bi-center prediction in anchor-free network. *Neurocomputing*, 507:199–207, 2022. 2, 3

[20] Qiming Li, Yijing Su, Yin Gao, Feng Xie, and Jun Li. Oaf-net: An occlusion-aware anchor-free network for pedestrian detection in a crowd. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 2, 3

[21] Chunze Lin, Jiwen Lu, Gang Wang, and Jie Zhou. Graininess-aware deep feature learning for pedestrian detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 732–747, 2018. 2, 3

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[23] Zebin Lin, Wenjie Pei, Fanglin Chen, David Zhang, and Guangming Lu. Pedestrian detection by exemplar-guided contrastive learning. *IEEE transactions on image processing*, 2022. 2, 3, 4, 8

[24] Mengyin Liu, Chao Zhu, Jun Wang, and Xu-Cheng Yin. Adaptive pattern-parameter matching for robust pedestrian detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2154–2162, 2021. 2, 3, 8

[25] Songtao Liu, Di Huang, and Yunhong Wang. Adaptive nms: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6459–6468, 2019. 2, 3, 7, 8

[26] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–634, 2018. 2, 3, 7, 8

[27] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[28] Ruiqi Lu, Huimin Ma, and Yu Wang. Semantic head enhanced pedestrian detection in a crowd. *Neurocomputing*, 400:343–351, 2020. 2, 3, 7, 8

[29] Zekun Luo, Zheng Fang, Sixiao Zheng, Yabiao Wang, and Yanwei Fu. Nms-loss: learning with non-maximum suppression for crowded pedestrian detection. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 481–485, 2021. 3, 7, 8

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 5, 6

[32] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 2, 3, 4, 5

[33] Xiaolin Song, Binghui Chen, Pengyu Li, Biao Wang, and Honggang Zhang. Prnet++: Learning towards generalized occluded pedestrian detection via progressive refinement network. *Neurocomputing*, 482:98–115, 2022. 2, 3, 7, 8

[34] Fiseha B Tesema, Hong Wu, Mingjian Chen, Junpeng Lin, William Zhu, and Kaizhu Huang. Hybrid channel based pedestrian detection. *Neurocomputing*, 389:1–8, 2020. 2

[35] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021. 2, 3, 5

[36] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7774–7783, 2018. 3, 7, 8

[37] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 3, 5

[38] Yang Wang, Chong Han, Guangle Yao, and Wanlin Zhou. Mapd: an improved multi-attribute pedestrian detection in a crowd. *Neurocomputing*, 432:101–110, 2021. 2, 3, 7, 8

[39] Jialian Wu, Chunluan Zhou, Qian Zhang, Ming Yang, and Junsong Yuan. Self-mimic learning for small-scale pedestrian detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2012–2020, 2020. 2, 3, 7, 8

[40] Jin Xie, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Mubarak Shah. Count-and similarity-aware r-cnn for pedestrian detection. In *European Conference on Computer Vision*, pages 88–104. Springer, 2020. 2, 3, 7, 8

[41] Jin Xie, Yanwei Pang, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection. *IEEE transactions on image processing*, 30:3872–3884, 2020. 2, 3, 4

[42] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1259–1267, 2016. 6

[43] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017. 2, 5, 6

[44] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653, 2018. 2, 3, 7, 8

[45] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6995–7003, 2018. 2

[46] Shan Zhang, Xiaoshan Yang, Yanxia Liu, and Changsheng Xu. Asymmetric multi-stage cnns for small-scale pedestrian detection. *Neurocomputing*, 409:12–26, 2020. 2, 3, 7, 8

[47] Tianliang Zhang, Qixiang Ye, Baochang Zhang, Jianzhuang Liu, Xiaopeng Zhang, and Qi Tian. Feature calibration network for occluded pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 2, 3, 4, 8

[48] Anlin Zheng, Yuang Zhang, Xiangyu Zhang, Xiaojuan Qi, and Jian Sun. Progressive end-to-end object detection in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 857–866, 2022. 2, 3

[49] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 2, 4, 5

[50] Penghao Zhou, Chong Zhou, Pai Peng, Junlong Du, Xing Sun, Xiaowei Guo, and Feiyue Huang. Noh-nms: Improving pedestrian detection by nearby objects hallucination. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1967–1975, 2020. 2

[51] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2582–2593, 2022. 2, 3, 5