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

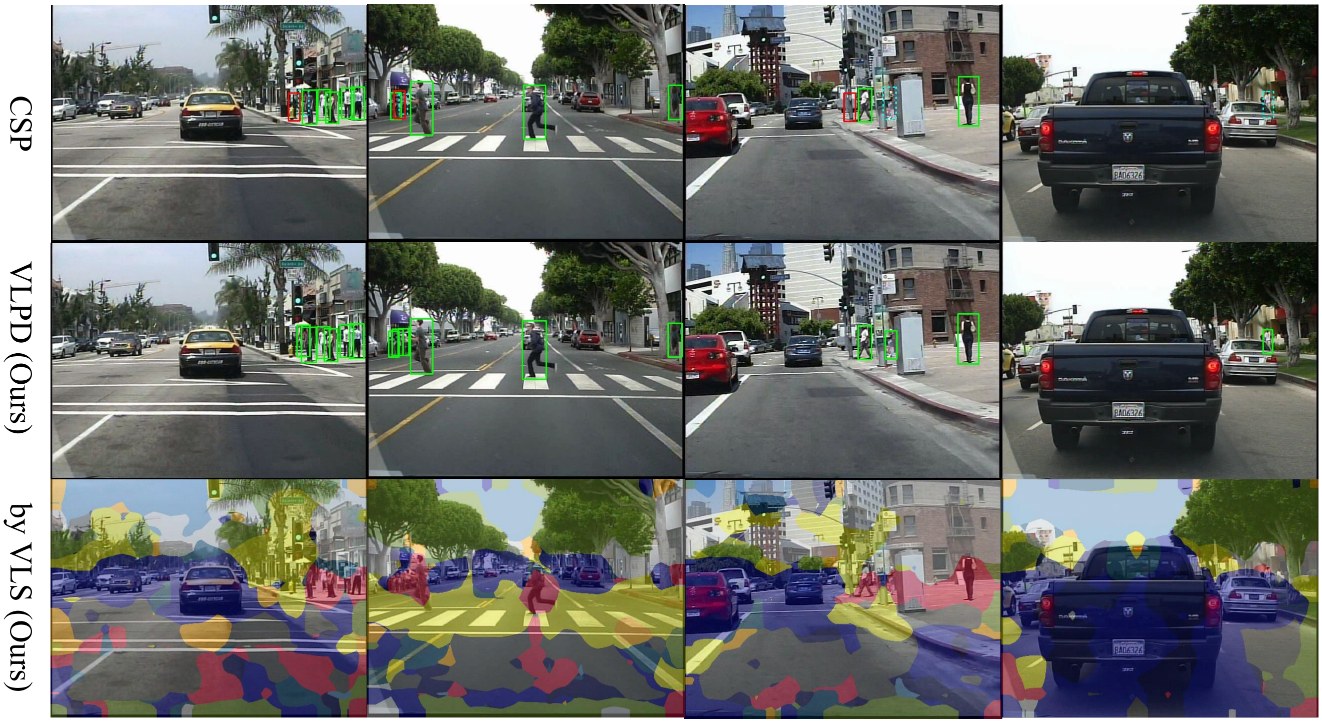

Figure 7. Qualitative analysis on Caltech [3] between our proposed VLPD and the baseline CSP [4]. Green bounding boxes are correct detections, red are wrong detections, and dashed blue are missing detections. Semantic classes of explicit contexts are obtained from our proposed VLS and colored following [1], e.g., "human" is red, "traffic sign" is yellow and "car" is blue.

6. Qualitative Analysis

In this section, we provide more qualitative analysis of our proposed **V**ision-**L**anguage semantic self-supervision for **P**edestrian **D**etection (**VLPD**) on challenging benchmarks Caltech [3] and CityPersons [9]. Various forms of visualizations are illustrated including detection results, contexts by our VLS and t-SNE of prototypes by our PSC.

6.1. Comparisons on Caltech Benchmark

In addition to our main paper, more qualitative analysis on Caltech [3] benchmark is provided in Figure 7 and 8 of this section. Note that different colors in Figure 9 indicate the semantic classes of explicit contexts from our proposed Vision-Language Semantic (VLS) segmentation.

As illustrated in Figure 7, the 1st and 2nd columns demonstrate typical crowded scenes. For example, the dashed blue boxes in the 1st rows of them are missing pedestrians by the baseline CSP [4] due to heavy occlusion. Highlighted by the red color for "human" class by our proposed VLS, these regions are focused on by our proposed VLPD and thus handled properly for accurate detections.

In the 3rd and 4th columns, the missing pedestrians are occluded by non-human objects, whose appearances are

∗ Equal Contribution. † Corresponding Author.

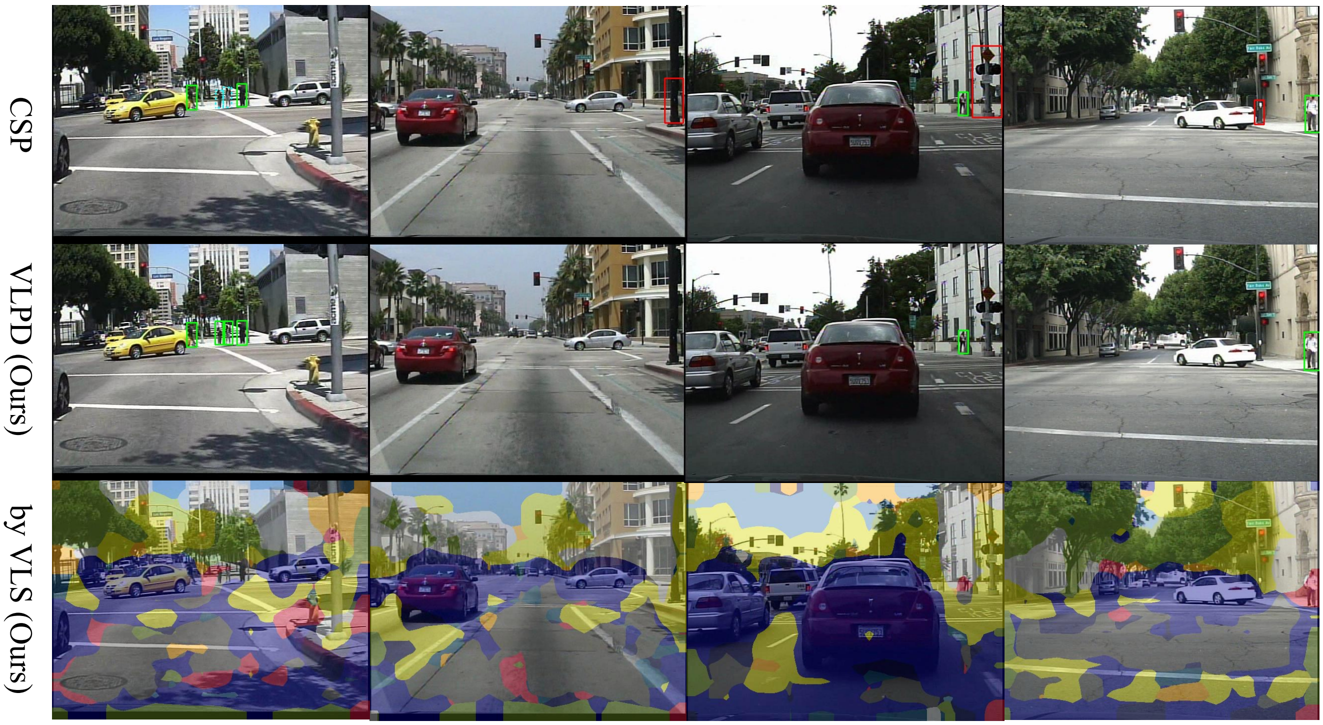

Figure 8. Qualitative analysis on Caltech [3] between our proposed VLPD and the baseline CSP [4]. Green bounding boxes are correct detections, red are wrong detections, and dashed blue are missing detections. Semantic classes of explicit contexts are obtained from our proposed VLS and colored following [1], e.g., "human" is red, "traffic sign" is yellow and "car" is blue.

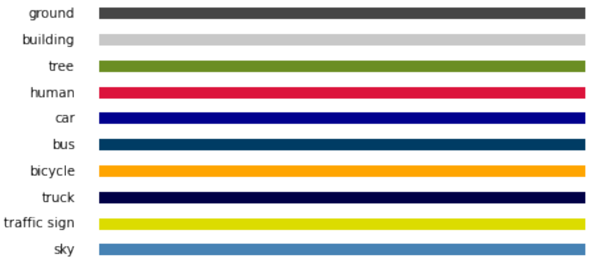

Figure 9. Different colors indicate the contextual classes segmented by our proposed VLS, based on the setting of CityScapes [1] with some modifications for highlighting the pedestrians.

more unusual than others. For instance, in the 3rd column, explicit contexts by our VLS carefully split the visible and invisible parts of the pedestrian occluded by the traffic signs by red and yellow colors. Similarly, the 4th column shows the pedestrian under the tree is divided by our VLS into red and blue colors for visible and occluded parts from the car. In the 1st column of Figure 8, two small pedestrians are missing, while our VLS highlights them via red color.

Besides, diversified human-like objects are confusing in the 1st column of Figure 7 and last 3 ones of Figure 8. Our proposed VLS marks them carefully with colors which represent non-human objects, e.g., yellow for "traffic sign".

Although the explicit contexts of semantic classes from our VLS are label-free and coarse-grained, not only human-like objects are classified, but also most ground truth pedestrians are correctly predicted. Moreover, under the fully-supervised detection training in Eq. 5 of the main paper, the Detection Head learns to detect with only useful clues.

With these explicit contexts, our proposed Prototypical Semantic Contrastive (PSC) learning supervises the detector to discriminate the pedestrians and contexts by contrastive self-supervision. Therefore, our proposed VLPD strikes the balance between keeping more salient regions of pedestrians as possible and avoiding the non-human objects.

In conclusion, our proposed VLPD tackles the challenges of both confusing non-human objects and unusual small-scale or occluded pedestrians which hinders the performance gains of previous methods, equipped with the powerful vision-language semantic self-supervisions of VLS and PSC without any extra laborious annotations.

6.2. Comparisons on CityPersons Benchmark

In this section, we provide qualitative analysis on another challenging benchmark CityPersons [9]. Note that its evaluation protocol is different from Caltech [3], where sitting persons, bicycle or motorcycle riders are not pedestrians.

Figure 10 mainly illustrates the different wrong detec-

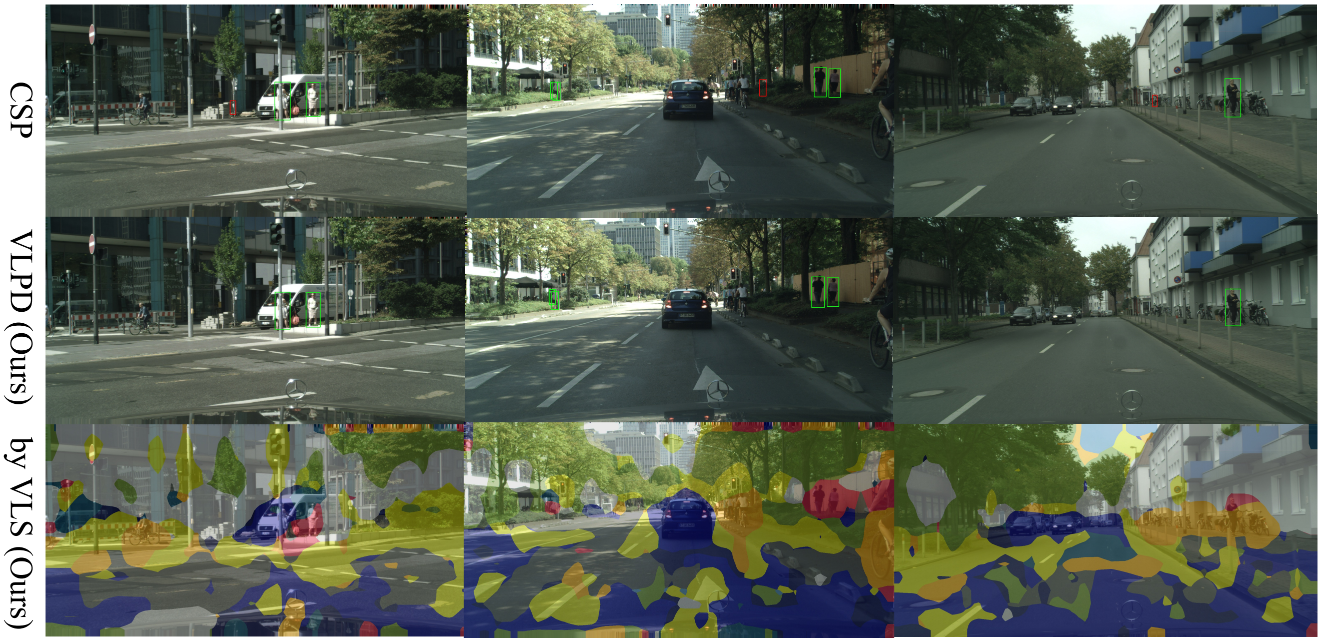

Figure 10. Qualitative analysis on CityPersons [9] between our proposed VLPD and the baseline CSP [4]. Green bounding boxes are correct detections, red are wrong detections, and dashed blue are missing detections. Semantic classes of explicit contexts are obtained from our proposed VLS and colored following [1], e.g., "human" is red, "traffic sign" is yellow and "car" is blue.

tions caused by non-human objects via the baseline CSP [4]. For example, some parts of buildings in the middle of images inside the 1st column are mistakenly detected as "pedestrian" in a red bounding box. Similarly, trees by the road and distant bicycles in the 2nd and 3rd columns of Figure 10 mislead the context-agnostic baseline. All of them are classified into non-human classes as white, green and orange colors by our proposed VLS, respectively.

In crowded scenes of the 1st and 2nd columns of Figure 12, the baseline CSP [4] mistakes some parts of pedestrians as the whole one or ignores some occluded pedestrians. In the 3rd column, CSP not only misses the heavily occluded pedestrian behind the buildings by the right side, but also detects a wrong pedestrian near the roadside.

Under these complex circumstances, our proposed VLPD utilizes explicit extra-annotation-free contexts via our VLS and self-supervised discriminative representations of pedestrians and non-human these contexts via our PSC. Then, the regions of pedestrians are especially concentrated and human-like objects are avoided. Consequently, detection results of our VLPD are more robust to small or occluded pedestrians and human-like objects.

Generally speaking, our proposed VLPD are evaluated by diversified challenging scenarios on both benchmarks Caltech [3] and CityPersons [9], including human-like confusing objects and hard small or occluded pedestrians, via **the first vision-language extra-annotation-free method for pedestrian detection to our best knowledge**.

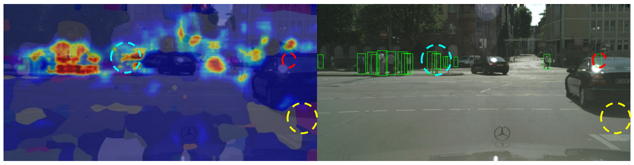

Figure 11. Visualization on how VLPD uses contexts of VLS (left) for detection (right). Yellow circle is ignored fake "human". Red is detected without VLS like CSP. Blue circle is rescued by VLS from CSP (Please refer to Col. 2 of Figure 12 and zoom in). VLPD also notices some human-like traffic signs and avoids them.

6.3. Visualization of Context Usage for Detection

As stated in the Figure 3 of the main paper, both basic detection capability of CSP and our CLIP-based VLS contribute the final detection. Hence, our VLPD learns from joint $\mathcal{L}_{Det}$ and $\mathcal{L}_{VLS}$ to choose helpful parts of VLS.

In Figure 11, pedestrian in red circle is detected like CSP without VLS clues, and fake red "human" regions in yellow circle are ignored by VLPD. Meanwhile, our VLPD keeps the capability from CSP and thus detects occluded pedestrian in red circle without VLS. Person with low-IoU VLS in Col. 3 of Figure 10 is similar as well. Furthermore, contexts via VLS empowers our VLPD to: 1) avoid human-like objects, e.g., red box in Col. 3 of Figure 12 is removed by orange "bicycle" regions; 2) highlight hard pedestrians like blue circle of Figure 11 from Figure 12.

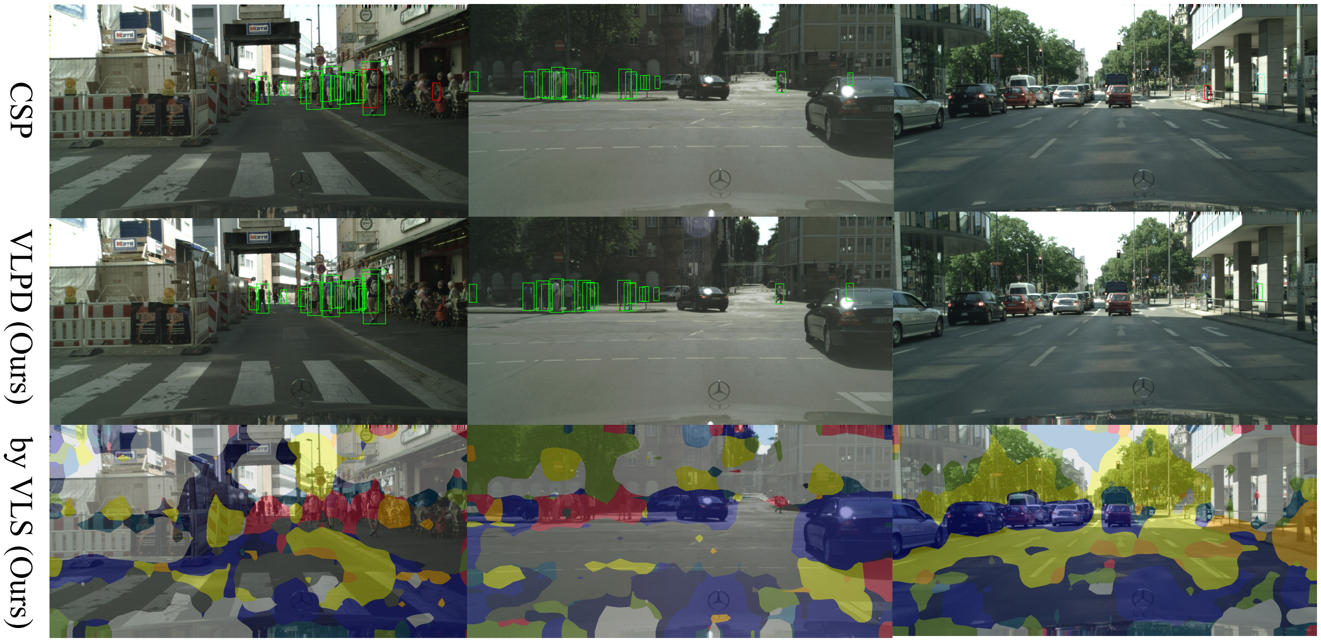

Figure 12. Qualitative analysis on CityPersons [9] between our proposed VLPD and the baseline CSP [4]. Green bounding boxes are correct detections, red are wrong detections, and dashed blue are missing detections. Semantic classes of explicit contexts are obtained from our proposed VLS and colored following [1], e.g., "human" is red, "traffic sign" is yellow and "car" is blue.

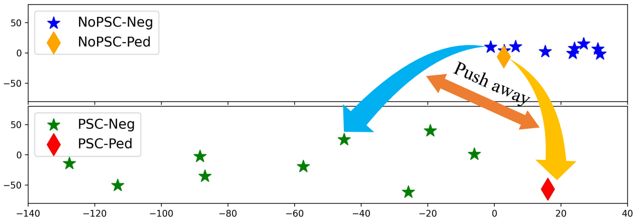

Figure 13. t-SNE visualization of prototypes by PSC (=VLPD) or not (CSP w/ CLIP+VLS) on a random image from CityPersons.

6.4. t-SNE Visualization of Prototypes

As shown in Figure 13, a push-away of positive prototypes for pedestrians from various negative ones is observed after our PSC is applied. We select a random image from CityPersons [9] dataset, and then obtain the "Detection Feature" E and $\hat{S}$ from VLPD to calculate the prototypes.

7. Quantitative Analysis

In this section, further quantitative analysis are provided to evaluate the two key components of our proposed **V**ision-**L**anguage semantic self-supervision for **P**edestrian **D**etection (**VLPD**): Vision-Language Semantic (VLS) segmentation and Prototypical Semantic Contrastive (PSC) learning, including network achitectures, statistics, hyperparameters, and learning schemes.

Table 8. Different network architectures about $\bar{S}$ of VLS.

Method	Reasonable	Small	HO
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**
+VLS w/o C & U	9.72	12.67	36.35
+VLS w/o C & U+PSC	10.16	12.30	38.31

7.1. Different Network Architectures of VLS

As illustrated by the Figure 3 of our main paper, we follow the network architecture design of DenseCLIP [6], where the predicted $\bar{S}$ via VLS are up-sampled ("U") into $\dot{S}$, concatenated ("C") with Detection Feature E, and then fed into Detection Head for an explicit contextual reference.

Although such a design is not individually evaluated via ablation study in [6], we make thoroughly quantitative analysis in Table 8. Without the reference of $\dot{S}$ as "+VLS w/o C & U", Detection Feature E is the only input of Detection Head, thus the pixel-wise explicit classes from $\dot{S}$ are unavailable. Detection Head is only aware that some pixels of E are concentrated by contrastive learning of PSC, but their explicit classes (human or non-human objects, and more likely to be human-like objects or not) are unknown.

Hence, the $\dot{S}$ is evaluated to be significant for both our

Table 9. Different selection of λ_2 for PSC on Caltech dataset.

Method	Reasonable	All	Heavy
(CSP w/ CLIP) + VLS	2.80	54.34	41.26
+ PSC (10^{-3})	3.27	54.47	41.70
+ PSC (10^{-4}, =VLPD)	**2.27**	**52.37**	**37.12**
+ PSC (10^{-5})	3.17	53.43	40.75

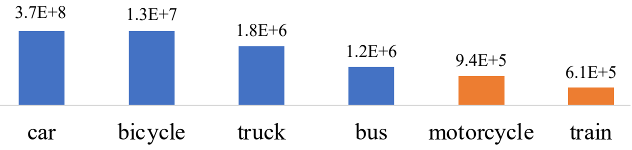

Figure 14. Frequencies among vehicle classes via not only counting pixels but also re-weighting by image-wise occurrence times.

proposed VLS and PSC to cooperate with each other in our proposed VLPD for the best performance.

7.2. Statistics for Compacted Class Policy of VLS

Due to the excessive variance inside the "vehicle" class, performance is decreased as "Full Compacted" in Table 3 of the main paper. Therefore, frequencies are obtained based on pixel-wise annotations of CityScapes for the images shared by CityPersons [9]. The frequency F_c is computed via not only pixel-wise counting K_c but also reweighting by image-wise occurrence times T_c, where each class $c \in C$.

Therefore, as shown in Figure 14, we observe a significant gap between the head and tail classes. In order to strike a balance between maintaining the variance and eliminating the low-frequent tail classes, we choose "motorcycle" and "train" as omitted classes by a threshold $F_c^* = 10^6$.

Intuitively, once the instances of these classes occur inside the input images, they can also be predicted as other classes with similar appearance like "bicycle", "bus" or "truck" via the cross-modal mapping in our VLS. Hence, this refined policy of VLS lead to better performances in Table 2 and 3 of the main paper.

7.3. Model Capacity Compared with Baseline

Since most state-of-the-art methods do not publish their results, only our VLPD and baseline CSP can be compared: 46.3M vs 40.0M, 180.3G MACs vs 173.5G MACs. Extra 6M parameters are to predict contextual score maps of VLS. Please note that the linguistic vectors in Figure 3 of the main paper are fixed, so text encoder is not for inference.

7.4. Loss Weight Selection of PSC

Due to the characteristics of contrastive learning, $\mathcal{L}_{PSC}$ of PSC learns the features merely under a weak constraint that pixel-wise features of pedestrians are pulled close to the positive prototype and pushed away from the negative ones. Differently, $\mathcal{L}_{VLS}$ still has pseudo labels as a stronger constraint, and $\mathcal{L}_{Det}$ is fully-supervised and thus similar.

Table 10. Different selection of λ_2 for PSC on CityPersons dataset.

Method	Reasonable	Small	HO
(CSP w/ CLIP) + VLS	9.70	12.57	36.50
+ PSC (10^{-2})	10.10	12.88	35.95
+ PSC (10^{-3}, =VLPD)	**9.41**	**10.93**	**34.88**
+ PSC (10^{-4})	9.79	13.40	36.09

Table 11. Different contrastive learning for our proposed PSC.

Method	Reasonable	Small	HO
VLPD (w/ PSC, ours)	**9.41**	**10.93**	**34.88**
P^+ $(G) \to P^{\text{h}+}$ (VLS)	10.07	12.28	38.23
Each $E_j \to P^+$	10.57	12.87	39.81

Consequently, how far to locate the pedestrian features inside their feature space might be either over-fitted or under-fitted, if the loss weight λ_2 of $\mathcal{L}_{PSC}$ in Eq. 5 from the main paper are defined improperly. Here, experiments of the loss weight λ_2 are conducted on both benchmarks, i.e., Caltech [3] and CityPersons [9], respectively.

In Table 9, both the $\lambda_2 = 10^{-3}$ and 10^{-5} worsen the performance of all the subsets on Caltech. Although other two λ_2 improve the HO on CityPersons in Table 10, the best results on all the subsets are still achieved by $\lambda_2 = 10^{-3}$. Therefore, we chose $\lambda_2 = 10^{-4}$ for Caltech and 10^{-3} for CityPersons, respectively.

7.5. Different Contrastive Learning of PSC

In addition to the Table 4 of our main paper, we further investigate the design of contrastive learning of our proposed PSC, including different positive prototypes and prototype-to-prototype learning.

To evaluate different positive prototypes, 2D Gaussians of pedestrian positions G is replaced by the score map $\hat{S}^{\text{h}}$ of contextual class "Human" denoted as "h". Since this class is merely used to avoid that the pixels of pedestrians are mistaken for other classes, $\hat{S}^{\text{h}}$ is obtained via more coarse-grained self-supervision of VLS like Figure 7, 8, 10 and 12, where the positive prototype is denoted as $P^{\text{h}+}$. In Table 11, "P^+ $(G) \to P^{\text{h}+}$ (VLS)" leads to performance loss by $\hat{S}^{\text{h}}$ which misleads the feature learning of pedestrians.

Furthermore, we evaluate a full-prototype version of contrastive learning, which only prototype P^+ is supervised like [8, 11], rather than pixel-wise features E_j. Con-

Table 12. Different learning schemes of the visual encoder, i.e., self-supervision of CLIP [5] or re-training of ImageNet (IN) [2].

Method	Reasonable	Small	HO
VLPD	**9.41**	**10.93**	**34.88**
VLPD w/ re-train IN [2]	11.30	15.67	42.47

sequently, the loss function $\mathcal{L}_{PSC}$ is modified as:

$$\mathcal{L}_{PSC} = -\log \frac{\exp(\mathbf{1}/\tau)}{\exp(\mathbf{1}/\tau) + \sum_{c,b} \exp(P^+ \cdot P_b^{c-}/\tau)}, \quad (6)$$

where $P^+ \cdot P^+ = \mathbf{1}$ and self-normalization, e.g, $P^+/\|P^+\|$, is omitted. As illustrated in Table 11, Eq. 6 with only P^+ cannot directly keep the variance among the pixel-wise E_j, which also causes the declined performances.

7.6. Different Learning Schemes of Visual Encoder

Following [6], we initialize the visual encoder from CLIP [5] to keep the cross-modal mapping from vision-language pretraining. Such a learning scheme is self-supervision, because the CLIP visual encoder is used for both training and labeling without any extra labels.

Instead, re-training the encoder based on ImageNet [2] via the CLIP one is evaluated to be sub-optimal by different results in Table 1 and 5 from [6]. Furthermore, the experiments are also conducted on our proposed VLPD.

As illustrated in Table 12, since the ImageNet visual encoder "w/ re-train IN" has not learned any cross-modal mapping before, it is difficult to use down-stream re-training on pedestrian detection tasks by our proposed VLS to obtain an equivalently powerful capability of cross-modal mapping from the up-stream vision-language pretraining.

8. Discussions on Limitations

Our proposed VLPD is designed to enhance urban pedestrian detection by the awareness of contexts based on self-supervision, which is featured with both domain-specific characteristics and coarse-grained pseudo labels. Hence, there are some limitations in its application.

On the one hand, although the semantic classes of contextual objects are mostly shared between both benchmarks for urban scenes, i.e., Caltech [3] and CityPersons [9], the vocabulary of contextual classes limits the usages for more generalized purposes. For instance, crowded or wide-scene benchmarks CrowdHuman [7] or WiderPerson [10] comprise more out-door and in-door circumstances, which require a larger, more open and adaptive vocabulary than our proposed one for merely urban contexts.

On the other hand, the self-supervised semantic segmentation for contextual classes via our proposed VLS is coarse-grained, as is shown in the qualitative analysis before. Although it decreases the heavy burden of laborious manual annotations by cross-modal mapping from vision-language CLIP [5] model, the noises in the pseudo labels still affect the accuracy of context modeling. In our opinions, a few shots of manual annotations (e.g., 10∼50 annotated images) as semi-supervision might help to solve this problem. Hence, coarse-grained pseudo labels also limit the performance of our proposed method.

In conclusion, domain characteristics of urban pedestrian detection and coarse-grained pseudo labels are the major limitations of our proposed method, and they also provide more potentials for the future works.

References

[1] 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. 1, 2, 3, 4

[2] 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. 6

[3] 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. 1, 2, 3, 5, 6

[4] 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] 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. 6

[6] 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. 4, 6

[7] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 6

[8] 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. 5

[9] 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. 1, 2, 3, 4, 5, 6

[10] Shifeng Zhang, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z Li, and Guodong Guo. Widerperson: A diverse dataset for dense pedestrian detection in the wild. *IEEE Transactions on Multimedia*, 22(2):380–393, 2019. 6

[11] 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. 5