Pedestrian Detection based on Visible Guided Occlusion Handling

Lien-Chieh Huang*, Ching-Te Chiu*, and Yung-Cheng Su*
* National Tsing Hua University, Hsinchu, Taiwan
E-mail: willy20104368@gmail.com, chiusms@cs.nthu.edu.tw, yungcheng.su@gapp.nthu.edu.tw

***Abstract*—Accurate and timely pedestrian detection is crucial in autonomous driving systems. There are a few major challenges including detecting pedestrians of varying sizes and distances within an image, occluded pedestrians or mutual occlusion, and high computational complexity. We propose a Visible Guided Pedestrian Detection (VGPD), in which the visible guided auxiliary head is only used during training, provides more supervision to reduce the ambiguity caused by occlusion, and helps eliminate false positives. The visible guided label assignment ensures accurate positive samples for heavily occluded pedestrians and improves classification. We utilize re-parameterization module that merges multiple computational layers into a single one at the inference stage, to enhance feature representation and reduce complexity. In addition, Adaptive Spatial Feature Fusion (ASFF) and Path Aggregate Network (PAN) effectively fuse multi-scale features. VGPD achieved state-of-the-art performance on CityPersons, reducing inference time by an average of 48%. Compared to our baseline, VGPD also reduced the log average miss rate (the lower the better) from 9.9 to 7.7.**

I. Introduction

For autonomous driving, pedestrian detectors aim to deliver rapid and performant perception, enabling precise real-time decision-making. Beyond the balance between speed and performance, the scale variance and occlusion of pedestrians in traffic make it more challenging.

Small-scale pedestrians only occupy fewer pixels, resulting in lower resolution and less distinctive features. In addition, occlusion can cause pedestrians to be partially or completely invisible in images, making it difficult for detectors to extract strong features for accurate recognition and localization of pedestrians. VLDP[1] leverages vision-language models to provide vision-language semantic self-supervision to address both small-scale and occlusion challenges. Moreover, [2] utilized head boxes to provide extra supervision. [3] proposed novel regression loss in occlusion scenes. [4] leverage the concept of density to make NMS more suitable for pedestrians in the crowd.

Efficient detection is important for pedestrian detectors, however, most state-of-the-art methods focus on improving performance without addressing the computational overhead. [5] proposes a novel approach for pedestrian detection that emphasizes speed through a method called Center and Scale Prediction (CSP). F2dNet[6] improved the inference time by replacing the regional proposal network with a focal detection network. LSFM[7] utilize MLP-Mixer and fully connected layers to improve both inference time and performance.

We propose an efficient pedestrian detector that leverages additional visible region annotations to provide more supervision and appropriate positive samples. The multi-scale detection architecture is based on [8]. The strong backbone consists of our RepC4 module and ELAN[9]. In addition, the proposed neck(feature fusion) combined by [10] and [11] to enrich semantic and spatial information of features.

Moreover, applying auxiliary heads, which can be removed at the inference stage, to introduce more supervision mitigates the ambiguity of occluded pedestrians which helps reduce duplicated predictions of the same pedestrian. Furthermore, we adjust the different visibility thresholds to assign more appropriate positive samples, improving classification performance and reducing false positives. We perform experiments using anchor-based and anchor-free detection heads. Further, we conduct the ablation study to determine the effectiveness of each component in the proposed VGPD.

II. Proposed Method

As illustrated in Fig. 1, we utilize re-parameterization[12] with ELAN[9] to build a stronger backbone. To obtain well-integrated features, we fuse multi-scale features via ASFF-PAN. We introduce visible guided auxiliary heads and label assignment, used only during training, to handle occlusion, reduce duplicate predictions, and mitigate wrong classifications from human-like objects.

A. RepC4ELAN-CA Backbone

The backbone is composed of RepC4ELAN-CA, as shown in Fig. 2 (a), which improves the feature representation and stability while training. RepC4ELAN-CA utilized our RepC4 as the computational units and applied lightweight channel attention, Squeeze-and-Excitation (SE) block[13] to informative representational features. The proposed RepC4 in Fig. 2 (b) provides more training-time nonlinearity and diverse gradient paths via four paths with different combinations of kernel sizes during training. In addition, we introduce a 3x1 vertical kernel to make it more robust to pedestrian left-right flip. The four paths are fused into a single path during inference.

B. ASFF-PAN

Adaptive Spatial Feature Fusion (ASFF) and Path Aggregation Network (PAN) work together to enhance multi-scale feature representation for pedestrian detection. As illustrated in Fig. 3 (b), ASFF fuses multi-scale features from P3 to P5

979-8-3315-7206-8/25/$31.00

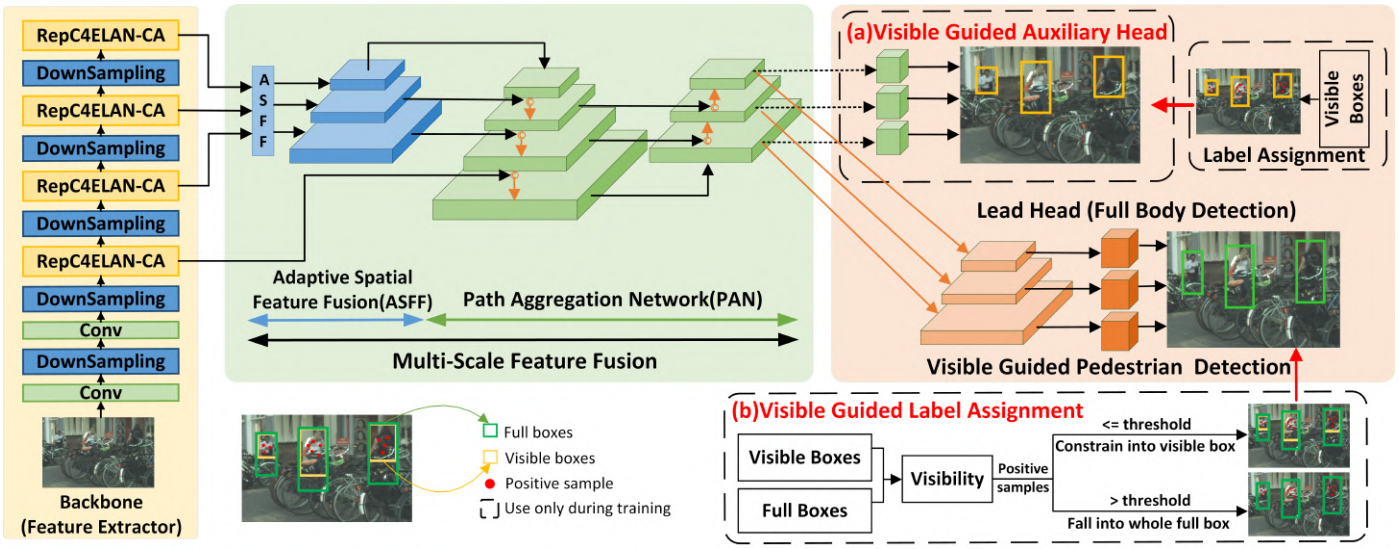
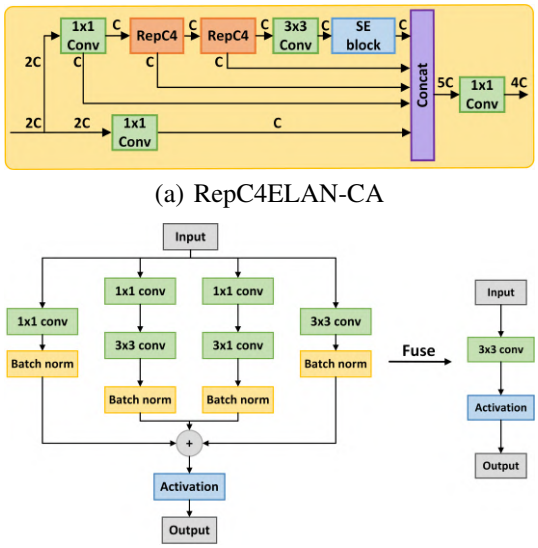


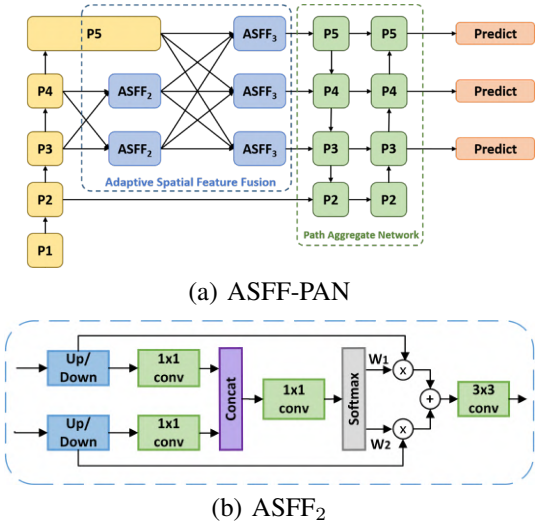
Fig. 1. Overall architecture of the proposed Visible Guided Pedestrian Detection(VGPD). (a) Visible guided auxiliary heads provide information about occlusion and are only used during training. (b) Visible guided label assignment based on visibility to assign appropriate positive samples.



(a) RepC4ELAN-CA

(b) RepC4 module. Left: training. Right: inference

Fig. 2. RepC4ELAN-CA and RepC4 module



(a) ASFF-PAN

(b) $ASFF_2$

Fig. 3. ASFF-PAN and $ASFF_2$ module

by computing adaptive weights, ensuring that the most relevant spatial information is preserved. Unlike traditional fixed fusion methods, ASFF dynamically adjusts feature integration by leveraging upsampling and downsampling, as shown in Fig. 3 (a), improving robustness in detecting pedestrians at varying scales. The $ASFF_3$ module is an extension of $ASFF_2$, designed to process three input feature layers, further improving multi-scale feature fusion and adaptability. PAN refines these features by introducing a top-down (descending) pathway, which propagates high-level semantic information to lower layers, and a bottom-up (ascending) pathway, which enhances the localization accuracy of low-level features. Additionally, high-resolution features (P2) from the backbone are incorporated to strengthen small-scale pedestrian detection by preserving finer details such as edges and textures. The integration of ASFF and PAN ensures precise object detection, particularly for small and occluded pedestrians, by combining fine-grained spatial information with enriched contextual representations.

C. Visible Guided Pedestrian Detection

The Visible Guided Pedestrian Detection (VGPD) method enhances pedestrian detection by leveraging auxiliary heads that provide additional supervision during training. As illustrated in Fig. 1, these auxiliary heads focus on detecting the visible regions of pedestrians, helping the model learn occlusion patterns more effectively. By introducing additional

supervision, the network can extract more discriminative features, allowing the lead heads to detect full-body pedestrians more accurately. However, to maintain efficiency, the auxiliary heads are removed at the inference stage (as depicted in Fig. 4), ensuring that they do not increase computational overhead while still improving detection performance.

During training, multi-scale features from the backbone are first fused using ASFF-PAN, which enhances both spatial and semantic information. The auxiliary heads then utilize visible boxes to guide the learning process, while the lead heads focus on full-body detections. This dual-head structure helps the model distinguish occluded pedestrians from background clutter. As shown in Fig. 4, without visible box guidance (Fig. 4(b)), lead heads learn ambiguous features from heavily occluded pedestrians, often leading to duplicate predictions for the same person. However, with visible box guidance (Fig. 4(d)), lead heads become aware of occlusion, utilizing features learned by the auxiliary heads (Fig. 4(c)) to reduce duplicate predictions.

To further refine pedestrian detection, we follow the loss settings from [8] for both anchor-based and anchor-free detectors. Additionally, we incorporate repulsion loss [3] to penalize overlapping predicted bounding boxes during training. This loss discourages excessive bounding box overlaps, helping the model better separate and distinguish closely spaced pedestrians. By integrating visible box supervision and repulsion loss, VGPD effectively improves occlusion handling without increasing inference time.

(a) Ground Truth (b) W/o VG auxiliary heads

(c) VG auxiliary heads (d) VG lead heads

Fig. 4. The visualization of the results with and without Visible box Guided auxiliary heads. (a) The red box denotes the visible box and the green box indicates the full box. (b)(c)(d) The green box represents prediction

D. Visible Guided Label Assignment

The proposed visible guided label assignment is based on [14], which is used in a decoupled head to align different tasks, such as regression of bounding box and classification, by calculating an align metric, which can be formulated as:

$$t = s^{\alpha} \times u^{\beta},\ u = \text{IoU}(b^{gt}, b^{p}) \tag{1}$$

, where $\alpha = 1$, $\beta = 6$, s denotes as classification scores, and u represents the IoU value. The alignment metric is calculated based on the IoU between predictions and ground-truth boxes, as well as classification results.

To avoid the positive samples falling onto other objects or background regions in crowded scenes, we compute the alignment metric by full or visible boxes depending on visibility, which is defined as the ratio of the visible region of a pedestrian within the full box. As shown in Eq.1 and 2, if the visibility is below the threshold, indicating that the pedestrian is heavily occluded, we use the visible boxes to compute the IoU with the predictions, otherwise, we use the full boxes. b^{gt} and b^{p} are denoted as ground truth boxes and predicted boxes. Visible boxes are used to calculate the IoU and constrain the positive sample regions, not to train lead heads.

$$b^{gt} = \begin{cases} \text{visible box}, & \text{if visibility} \leq \text{threshold}; \\ \text{full box}, & \text{if visibility} > \text{threshold}. \end{cases} \tag{2}$$

This approach constrains positive samples into the visible region of the pedestrian and reduces the number of positive samples from heavily occluded pedestrians. The boundaries between pedestrians and other objects become clearer, helping to distinguish pedestrians from human-like objects, and reducing the number of false positives. Furthermore, the reduced number of positive samples for heavily occluded pedestrians allows the model to focus more on non-occluded and slightly occluded pedestrians. The detailed setting of the visibility threshold will be discussed in the ablation study.

E. Detection Head

The detection head integrates anchor-based and anchor-free approaches for pedestrian detection, as shown in Fig. 5.

The anchor-based detection head applies a 3×3 and 1×1 convolution to refine predefined anchor boxes. The anchor-free detection head directly predicts object locations, using separate branches for localization and classification to enhance flexibility.

Adapted from [8] and [15], both detection heads are used in NMS-based and NMS-free VGPD models, with auxiliary heads removed during inference for efficiency.

III. EXPERIMENTS

A. Experimental Setup

Dataset: Since our proposed VGPD leverages visible box annotations for improved performance, we select the CityPersons dataset [16] and the Caltech Pedestrian dataset [17], both of which provide additional visible region bounding boxes.

Evaluation Setting: MR^{-2} is calculated by averaging miss rates at nine different false positives per image (FPPI) rates, logarithmically spaced between 10^{-2} and 10^{0}. A lower MR^{-2} value indicates better detection performance. Depending on [17], pedestrian detection can be divided into subsets based

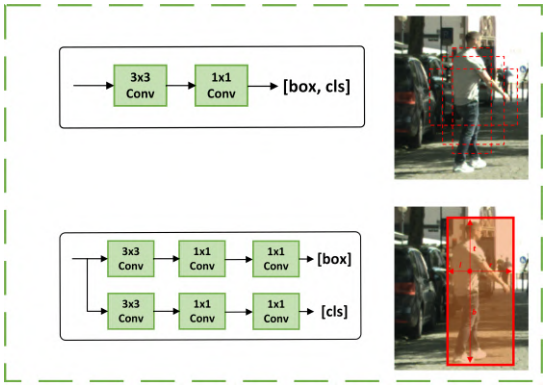


Fig. 5. Anchor-based (top) and anchor-free (bottom) detection heads. The former refines anchor boxes, while the latter predicts object centers and boundaries.

on the visibility ratio and the pixel height of the bounding box, namely reasonable, small, and heavy.

TABLE I
COMPARISON OF PEDESTRIAN DETECTION METHODS ON CITY PERSONS AND CALTECH DATASETS. AB AND AF INDICATE THE USE OF AN ANCHOR-BASED OR ANCHOR-FREE DETECTION HEAD. OUR CITYPERSONS MODELS ARE TRAINED FROM SCRATCH WITH 20% THRESHOLD BUT CALTECH MODELS WITH 40% THRESHOLD USED IMAGENET PRETRAINED BACKBONE FOR BETTER RESULTS. LOWER MR^{-2} MEANS BETTER.

Method	Reasonable	Small	Heavy	Inference
CityPersons[16]				
Pedestron[18]	11.2	14.0	37.0	0.73s
CSP[5], [18]	11.0	16.0	49.3	0.33s
PRNet[19]	10.8	–	42.0	0.22s
APD[4]	8.8	–	46.6	0.16s
F2DNet[6]	8.7	11.3	32.6	0.44s
LSFM P[7]	8.7	**8.7**	32.4	0.13s
LSFM[7]	8.5	8.8	31.9	0.18s
MTOM[20]	8.3	15.56	**27.07**	-
VGPD AB(ours)	**7.7**	9.87	32.0	**0.10s**
VGPD AF(ours)	7.8	10.1	32.0	0.12s
Caltech [10]				
Pedestron[18]	6.2	7.4	55.3	0.20s
ALFNet[10]	6.1	7.9	51.0	0.05s
AR-Ped[21]	4.4	–	48.8	0.09s
F2DNet[6]	2.2	**2.5**	38.7	0.14s
LSFM P[7]	3.9	4.2	37.6	0.03s
LSFM[7]	3.1	3.4	35.8	0.09s
MTOM[20]	**2.0**	2.8	38.6	-
VGPD AB(ours)	4.5	5.7	28.7	**0.027s**
VGPD AF(ours)	3.4	4.4	**27.7**	0.033s

B. Experimental Results

Comparison with other SOTA: Tab. I shows that VGPD achieves the shortest inference time and outperforms SOTA methods on CityPersons and Caltech. Fig. 6 compares inference times, while Fig. 7 highlights VGPD's superior performance under occlusion. Tab. II evaluates VGPD components, showing that RepC4ELAN-CA improves feature extraction, the Visible Guided Auxiliary Head enhances occlusion handling, mix-up boosts generalization, and repulsion loss reduces

TABLE II
ABLATION STUDY OF EACH COMPONENT. THE FIRST AND SECOND ROWS ARE THE RESULTS OF OUR BASELINE[8]. REP DENOTES REPC4ELAN-CA. MIX-UP REPRESENTS THE RATIO OF MIXUP DATA AUGMENTATION. MASK MEANS PEOPLE GROUP LABELS ARE MASKED IN TRAINING IMAGES TO AVOID CONFUSION BETWEEN PEOPLE GROUP AND SINGLE PERSON.

Components						CityPersons (MR^{-2})		
Rep	VG Aux	Mix-up	Repulsion	Feature Fusion	Mask	Reasonable	Small	Heavy
×	×	0.15	×	PAN	×	9.9	12.2	36.2
×	×	0.15	×	PAN	✓	9.5	11.6	35.2
✓	×	0.15	×	PAN	×	9.4	12.1	33.2
✓	×	0.15	×	ASFF	×	9.1	11.5	35.4
✓	✓	0.15	×	ASFF	×	9.1	11.4	33.8
✓	✓	0.15	✓	ASFF	×	8.4	10.8	33.5
✓	✓	**0.30**	✓	ASFF	×	8.6	10.1	32.1
✓	✓	**0.30**	✓	**ASFF-PAN**	×	8.4	9.88	32.1
✓	✓	**0.30**	✓	**ASFF-PAN**	✓	**7.7**	**9.87**	**32.0**

TABLE III
ABLATION STUDY OF VISIBLE GUIDED LABEL ASSIGNMENT WITH DIFFERENT THRESHOLDS ON CALTECH.

Visible Guided Label Assignment	Caltech (MR^{-2})		
	Reasonable	Small	Heavy
×	4.2	5.4	27.9
20%	4.5	5.5	25.4
40%	3.4	4.4	27.7
65%	4.0	5.1	29.5
100%	3.6	4.2	29.7

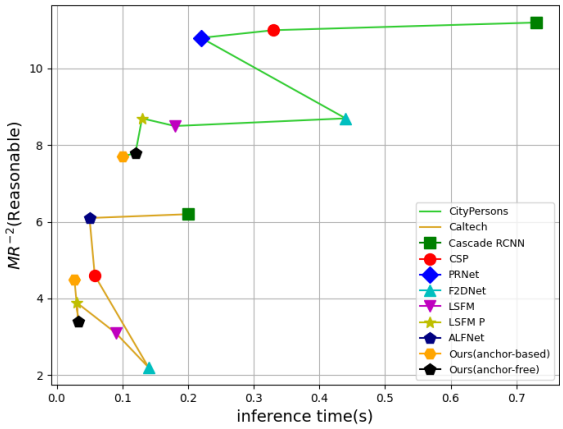
overlapping predictions. Feature fusion results indicate that PAN improves occlusion robustness, ASFF enhances multi-scale detection, and ASFF-PAN achieves the best balance.

Ablation Study: Tab. II shows the ablation study that explains the effectiveness of each component in VGPD. As shown in Tab. III, a higher threshold makes it more challenging to accurately regress occluded pedestrians because the anchor center may be far from the box centroid. Additionally, the reduced number of positive samples, resulting from the smaller visible box, can lead to an increase in missed detections. As the threshold decreases, positive samples for slightly occluded pedestrians can fall on full body regions rather than being limited to the visible region, making regression easier. Selecting an appropriate threshold can ensure accurate classification in cases of extreme occlusion while avoiding missed detections in slight occlusion.

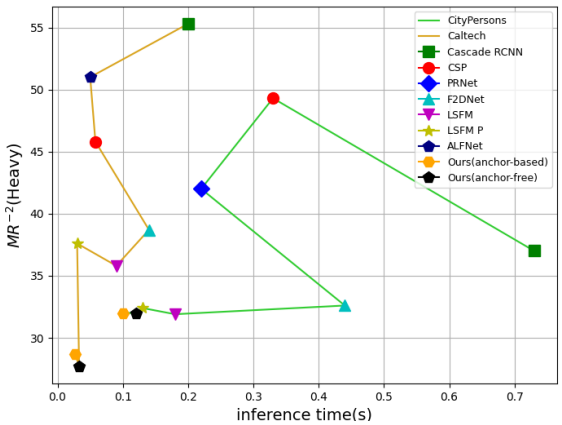
We evaluate PAN, ASFF, and ASFF-PAN on the CityPersons dataset. Table IV shows that PAN improves occlusion handling (33.2% vs. 35.4%), while ASFF is better for small pedestrians (11.5% vs. 12.1%). PAN enhances contextual information, whereas ASFF improves small-object detection. Combining both, ASFF-PAN achieves balanced detection across scales and occlusions.

TABLE IV
EVALUATION OF FEATURE FUSION TECHNIQUES ON THE CITYPERSONS DATASET.

Feature Fusion	CityPersons (MR^{-2})		
	Reasonable	Small	Heavy
PAN	9.4%	12.1%	**33.2%**
ASFF	**9.1%**	**11.5%**	35.4%

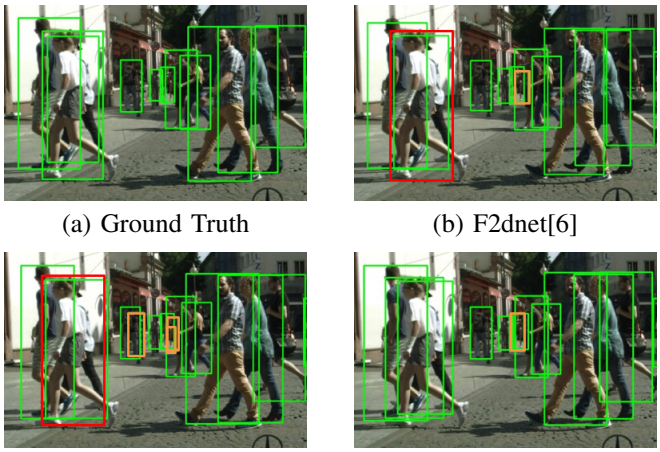


(a) Reasonable



(b) Occlusion

Fig. 6. The performance of different pedestrian detectors over inference time(s). Both figures contain two different pedestrian datasets namely, CityPersons(Green) and Caltech Pedestrian(yellow). Y-axis of both (a) and (b) are % based.



(a) Ground Truth (b) F2dnet[6]

(c) Pedestron[18] (d) VGPD(ours)

Fig. 7. Qualitative comparison of VGPD with other SOTAs. Green boxes indicate correct predictions, red boxes represent missed predictions, and orange boxes denote incorrect predictions.

IV. CONCLUSION

In this work, we propose Visible Guided Pedestrian Detection(VGPD), which leverages the additional information from the visible boxes without increasing computational complexity. A visible guided auxiliary head provides extra supervision via visible boxes that are particularly beneficial in crowded scenes. Moreover, the label assignment based on pedestrian visibility further improves performance by offering more appropriate positive samples to balance false positives and missed detections. Experimentally, we demonstrate that VGPD outperforms other pedestrian detectors while achieving the lowest inference time.

REFERENCES

[1] M. Liu, J. Jiang, C. Zhu, and X.-C. Yin, "Vlpd: Context-aware pedestrian detection via vision-language semantic self-supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6662–6671.

[2] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Pedhunter: Occlusion robust pedestrian detector in crowded scenes," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 10 639–10 646.

[3] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7774–7783.

[4] J. Zhang, L. Lin, J. Zhu, *et al.*, "Attribute-aware pedestrian detection in a crowd," *IEEE Transactions on Multimedia*, vol. 23, pp. 3085–3097, 2020.

[5] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5187–5196.

[6] A. H. Khan, M. Munir, L. van Elst, and A. Dengel, "F2dnet: Fast focal detection network for pedestrian detection," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, 2022, pp. 4658–4664.

[7] A. H. Khan, M. S. Nawaz, and A. Dengel, "Localized semantic feature mixers for efficient pedestrian detection in autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5476–5485.

[8] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7464–7475.

[9] C.-Y. Wang, H.-Y. M. Liao, and I.-H. Yeh, "Designing network design strategies through gradient path analysis," *arXiv preprint arXiv:2211.04800*, 2022.
[10] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," *arXiv preprint arXiv:1911.09516*, 2019.
[11] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.
[12] X. Ding, X. Zhang, J. Han, and G. Ding, "Diverse branch block: Building a convolution as an inception-like unit," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 886–10 895.
[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[14] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 2021, pp. 3490–3499.
[15] A. Wang, H. Chen, L. Liu, *et al.*, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.
[16] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *CVPR*, 2017.
[17] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, pp. 304–311.
[18] X. Li, W. Wang, L. Wu, *et al.*, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 002–21 012, 2020.
[19] X. Song, K. Zhao, W.-S. Chu, H. Zhang, and J. Guo, "Progressive refinement network for occluded pedestrian detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, Springer, 2020, pp. 32–48.
[20] K. A. Shastry, K. R. S. Teja, A. Nigam, and C. Arora, "Favoring one among equals - not a good idea: Many-to-one matching for robust transformer based pedestrian detection," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 748–757. DOI: 10.1109/WACV57701.2024.00081.
[21] G. Brazil and X. Liu, "Pedestrian detection with autoregressive network phases," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2019, pp. 7231–7240.