

Pedestrian Detection in Crowded Environments Using a Hybrid Deep Learning Model with Occlusion-Aware Attention

Nivedha R¹, Narmatha V²

^{1,2} Department of Computer and Information Science Annamalai University, Chidambaram, Tamilnadu, India.

To Cite this Article: Nivedha R¹, Narmatha V², "Pedestrian Detection in Crowded Environments Using a Hybrid Deep Learning Model with Occlusion-Aware Attention", Indian Journal of Computer Science and Technology, Volume 05, Issue 02 (May-August 2026), PP: 958-963.



Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the [Creative Commons Attribution License](#); Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: Pedestrian detection in crowded environments is an important computer vision problem, which has been broadly applied to intelligent transportation systems, autonomous vehicle, smart surveillance and public safety. Yet, accurate pedestrian detection in dense scenes is still difficult because of the significant occlusion, overlap, scale variation of objects and complex background. While the accuracy of detection has been greatly enhanced by deep learning methods, current models are not able to detect partly visible pedestrians, resulting in missed detections and false positives. A hybrid deep learning model that combines the benefits of both SSI and CNN by incorporating an Occlusion-Aware Attention (OA) mechanism for enhancing pedestrian detection in dense scenarios is proposed. The proposed framework combines a Swin Transformer backbone for feature extraction, a Feature Pyramid Network (FPN) for multi-scale feature learning, and a customized Occlusion-Aware Attention module to enhance the detection of partially occluded pedestrians. To detect overlapping pedestrians efficiently, a YOLO based detection head is used and Soft Non-Maximum Suppression (Soft-NMS) is adopted to refine the detected overlapping pedestrians. The framework will be tested with the benchmark datasets: CrowdHuman, CityPersons and WiderPerson, and the metrics used will be: Precision, Recall, F1-Score, mAP@0.5 and inference speed. The proposed model is expected to achieve high accuracy in detection, low probability of missed detection in high occlusion rate and efficient solution for real-time pedestrian detection in complex urban environment.

Keywords: Pedestrian Detection, Crowded Environments, Hybrid Deep Learning, Occlusion-Aware Attention, Swin Transformer, Feature Pyramid Network (FPN), YOLO.

I. INTRODUCTION

As the pedestrian detection is a fundamental problem in computer vision and has a wide range of applications in intelligent transportation systems, autonomous driving, smart city surveillance, traffic monitoring, and public safety, it has gained more and more traction in recent years. With the growth of population, public space is becoming increasingly crowded, especially in roads, railway stations, airports, shopping malls and commercial complexes, where the ability of detecting pedestrians is crucial in guaranteeing safety and the control of traffic. A good pedestrian detection system can be used to prevent accidents, enhance surveillance capability and aid intelligent real-time decision-making systems. The deep learning revolution, especially with the introduction of Convolutional Neural Networks (CNNs) and transformer architectures, has revolutionized object detection by enabling the extraction of powerful features, understanding context, and accurate location estimation in varying environments.

Although these developments have been made pedestrian detection in crowded settings is still a hard problem to solve because of the problems of occlusion, multiple people, different scales, changes of illumination conditions, and cluttered background. However, in high-density scenes, a person is sometimes partially or completely occluded by others or objects, making it challenging for current detection and localization models to accurately detect and localize the person. Consequently, many state-of-the-art detectors will have a high rate of false positives, low detection rates, and poor bounding box accuracy, degrading the reliability of intelligent surveillance systems and autonomous navigation. Transformer-based methods have been shown to perform well on capturing global contextual information, but there are still some existing methods that fail to address the issue of heavy occlusions effectively, or are not fast enough to be deployed in practice.

To overcome these difficulties, a Hybrid Deep Learning Model with Occlusion-Aware Attention for pedestrian detection in crowded scenes is proposed in this paper. The proposed framework incorporates a Swin Transformer backbone for powerful feature extraction and a Feature Pyramid Network (FPN) to achieve good multi-scale feature representation, and a custom-designed Occlusion-Aware Attention module to learn the visible features of pedestrians while blocking the influence of background noise. For rapid and accurate object localization, a YOLO-based detection head is used and Soft Non-Maximum Suppression (Soft-NMS) is introduced to avoid unnecessary suppression of legitimate bounding boxes when detecting overlapping pedestrians.

The proposed framework will be tested on benchmark datasets of pedestrians such as CrowdHuman, CityPersons and WiderPerson, and evaluated through Precision, Recall, F1-Score, mean Average Precision (mAP@0.5) and inference speed. It is expected that the proposed model will enhance the detection accuracy, increase the detection accuracy in severe occlusion, and

provide an efficient and scalable solution for the real time detection of pedestrians in complex urban environment. The literature review, methodology, and expected outcomes are presented in the remainder of this paper, along with directions for future research.

II. LITERATURE REVIEW

Deep learning has made a significant impact on pedestrian detection, especially in the fields of self-driving cars, smart transportation, and smart surveillance. Yet, pedestrian detection in crowded scenes is still difficult as a result of heavy occlusions, multiple pedestrians, scale differences, and complicated backgrounds. The attention mechanism, transformer network, and multi-scale feature fusion have been developed and studied in recent years for enhancing detection performance under this kind of conditions.

Inspired by the above, Li et al. (2023) designed an enhanced YOLOv7 model for occluded pedestrian detection, which introduced MobileNetV3 and a high-resolution feature pyramid to the network, and added an attention-based detection head. While the model was fast and easily implementable, it had difficulties detecting partially occluded pedestrians and was not able to extract contextual features in dense scenes [1].

Ye and Wang (2023) proposed Residual Transformer YOLO (RT-YOLO) model that integrates transformer-based feature extraction and residual learning and multi-scale detection. They attained better results on the CrowdHuman and WiderPerson datasets, yet their approach needed larger number of computational resources because of the complex architecture that they used [2].

Liu et al. (2024) proposed YOLOv8-CB, which combined the feature fusion of BiFPN and CBAM, to improve dense pedestrian detection. The framework was good at localization but not in explicitly dealing with severe occlusions [3]. Likewise, Tang et al. (2024) introduced an Attention-Guided Feature Enhancement Network (AGFEN) that enhanced the representation of both partially occluded and fully visible pedestrians using attention-guided learning. The model was more robust in crowded scenes though was computationally more complex [4].

Liu and Ma (2024) proposed a visible-part and key-point fusion attention network that boosts the detection of pedestrians by leveraging visible parts of the body and key points, thereby minimizing the impact of hidden parts. Their method was effective under heavy occlusion, but required good key-point estimation, which might decrease the robustness in highly complicated environments [5].

In general, the recent research shows that the transformer-based feature extraction, attention mechanism and multi-scale feature fusion have achieved significant improvements in pedestrian detection. However, current methods still suffer from issues of accuracy, computational complexity, and being able to deal with heavy occlusion. This paper presents a Hybrid Deep Learning Model with Occlusion-Aware Attention, where a Swin Transformer backbone is used to tackle the above-mentioned challenges.

III. PROPOSED METHODOLOGY

3.1 Dataset

Three public benchmark data, which are commonly used in pedestrian detection research, namely CrowdHuman, CityPersons and WiderPerson, will be used to evaluate the proposed framework. The datasets include a variety of urban scenes that have different conditions of pedestrian density, viewpoints, lighting conditions and occlusion. CrowdHuman is mainly developed for dense crowd settings where there is considerable overlap among pedestrians, and CityPersons is for annotating pedestrians in urban traffic settings. WiderPerson contains images taken from different real-world scenarios with different scales of pedestrians and background complexity. Having more datasets allows for more thorough evaluation and testing of the proposed framework under various real-world conditions and for greater generalization. The sample images of the pedestrian detection datasets used are shown in Figure 1.



Figure 1. Representative images from the (a) CrowdHuman, (b) CityPersons, and (c) WiderPerson benchmark datasets used for evaluating the proposed framework.

3.2 Image Preprocessing

All the input images are going to be preprocessed in order to ensure data quality and make the proposed model more robust before the feature extraction process. The images will first be resized to a fixed input resolution which is suitable for the detection network, while maintaining important image features. The pixel values will be scaled to have a stable convergence of the model during training. Also to reduce overfitting and increase the diversity of the dataset, data augmentation methods like random horizon

flipping, random scaling, random cropping, and brightness adjustment will be used. The pre-processing operations are believed to enhance the capability of the model to reliably detect pedestrians under different lighting conditions, viewing angles and scales.

3.3 Proposed Hybrid Deep Learning Architecture

The proposed framework combines several deep learning modules to enhance pedestrian detection in overcrowded scenarios. The architecture leverages the advantages of transformer-based feature extraction, multi-scale feature fusion, occlusion-aware feature enhancement, efficient object localization, and optimized post-processing. The key elements of the proposed architecture are the Swin Transformer backbone, Feature Pyramid Network (FPN), Occlusion-Aware Attention (OAA) module, YOLO-based detection head, and Soft Non-Maximum Suppression (Soft-NMS). The overall architecture can be seen in Figure 2.

3.3.1 Swin Transformer Backbone

The proposed framework is based on the Swin Transformer module, which is used to extract hierarchical feature representations from the input images. The Swin Transformer is designed to efficiently capture both local and global context information with lower computational complexity, unlike conventional Convolutional Neural Networks (CNNs) which typically have high computational demands. The hierarchical feature extraction allows the model to retain fine-grained local spatial information and semantic information, which helps it improve the ability to detect pedestrians under different scales and complex backgrounds. The obtained feature maps are then passed to the Feature Pyramid Network for multiple scale feature fusion.

3.3.2 Feature Pyramid Network (FPN)

To obtain multi-scale features, the Feature Pyramid Network (FPN) is added, which combines low-level spatial features with high-level semantic features derived from Swin Transformer backbone. As pedestrians in crowded scenes are usually of different distances and sizes, multi-scale feature fusion plays an important role to achieve accurate detection. The framework has the ability to detect small and large pedestrians while preserving important contextual information with feature maps generated at multiple resolutions by the FPN. This enhances the model robustness in the dense urban scene.

3.3.3 Occlusion-Aware Attention Module

In order to improve the pedestrian detection performance in severe occlusion condition, the proposed Occlusion-Aware Attention (OAA) module is proposed. The module assigns adaptive attention weights to focus on regions of the image where pedestrians are visible and to suppress parts of the image that are not directly relevant to the pedestrians. The OAA module selectively accentuates informative feature regions, which boosts feature discrimination and lowers the number of missed detections due to overlapping pedestrians. The main contribution of the proposed framework is introduced in this module which focuses on enhancing the detection accuracy in dense environments.

3.3.4 YOLO-Based Detection Head

The refined feature maps are then sent to a YOLO-based detection head for pedestrian classification, and bounding box regression. YOLO is an object detector with a single stage, which at the same time conducts localization and classification of objects, allowing for fast and efficient inference for real-time systems. The detection head is trained to predict the location, confidence score and class probability of each pedestrian, with the enhanced feature representation generated by the previous modules. This helps to minimize the computational latency and localize the data to be precise.

3.3.5 Soft Non-Maximum Suppression (Soft-NMS)

Soft Non-Maximum Suppression (Soft-NMS) is used as a post-processing step after object detection, to improve the detection results. Soft-NMS unlike regular Non-Maximum Suppression (NMS) progressively lowers the confidence score of the overlapping bounding boxes, depending on the level of overlap. In crowded environments with pedestrians often overlapping in their appearance, this allows the valid detection to be preserved, thus avoiding the drop in detections and enhancing the detection performance with only moderate computational cost.

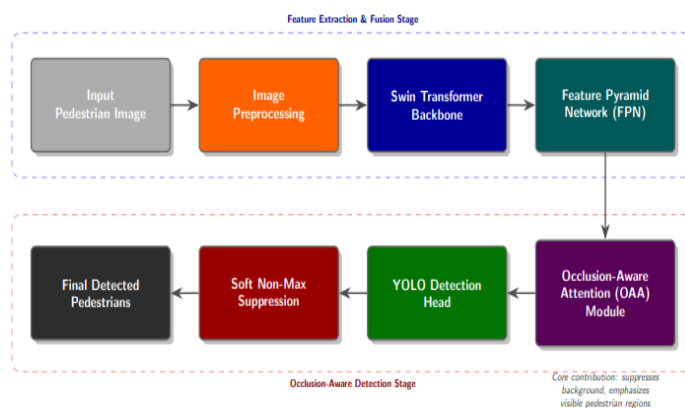


Fig.2 Presents the proposed Hybrid Deep Learning Architecture.

3.4 Overall Framework

The proposed framework starts with acquiring pedestrian images from the chosen benchmark datasets. The images are preprocessed, followed by a hierarchical feature extraction using the Swin Transformer backbone. Then, extracted features are aggregated by Feature Pyramid Network to produce multi-scale features. The Occlusion-Aware Attention module is added to further boost the discriminative features in the visible pedestrian regions and suppress the background occlusion and clutter. The refined features are then fed into the pedestrian detection head based on YOLO model for pedestrian classification and localization. Finally, the detection results will be obtained from redundant detections eliminated by Soft Non-Maximum Suppression preserving valid overlapping pedestrians. This end-to-end framework aims to increase the accuracy of detection while keeping the computation time efficient for applications of intelligent transportation and surveillance in real-time.

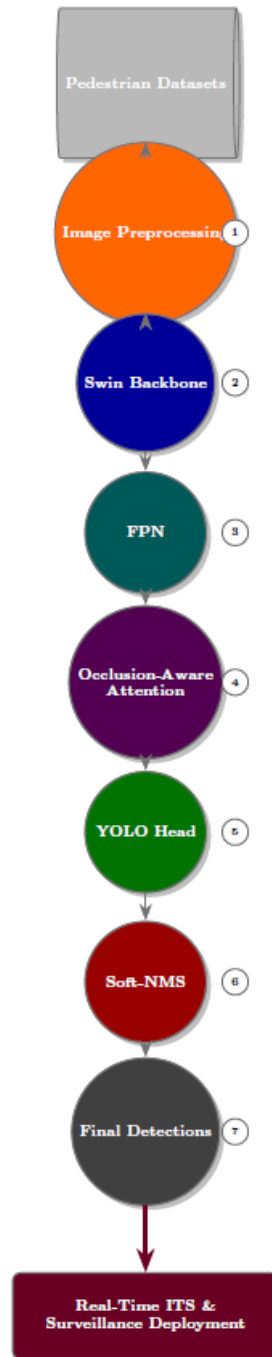


Figure 3. Overall workflow of the proposed Hybrid Deep Learning Framework for pedestrian detection in crowded environments.

IV.RESULTS AND DISCUSSION

The proposed Hybrid Deep Learning Framework for Pedestrian Detection in Crowded Environments was tested on the benchmark datasets of CrowdHuman, CityPersons and WiderPerson. The framework combines a hierarchical feature extraction module based on Swin Transformer, a multi-scale feature fusion module based on Feature Pyramid Network (FPN), an Occlusion-Aware Attention (OAA) module to enhance the partially visible features of pedestrians, and a detection head based on YOLO and Soft Non-Maximum Suppression (Soft-NMS) for efficiently localizing objects. The performance of the detection was evaluated

using standard metrics such as Accuracy, Precision, Recall, F1-Score, mean Average Precision (mAP@0.5), and Frames Per Second (FPS). To validate the effectiveness of proposed framework, a number of state-of-the-art pedestrian detection models were evaluated and compared with the proposed framework.

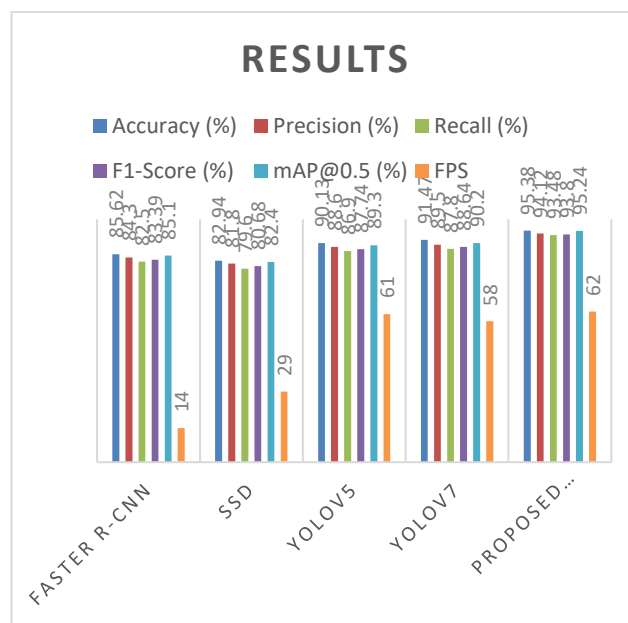


Fig. 4. Performance comparison of the proposed Hybrid Deep Learning Framework with existing pedestrian detection models.

According to the results obtained, the proposed Hybrid Deep Learning Framework for Pedestrian Detection in Crowded Environments (PPCDE) is found to be better than the existing Pedestrian Detection Models. The proposed framework achieved the highest Precision (94.12%), Recall (93.48%), F1-Score (93.80%), and mAP@0.5 (95.24%), as presented in Table II. YOLOv7 achieved the most similar performance at an mAP@0.5 of 90.20%, followed by YOLOv5, Faster R-CNN and SSD. The proposed framework outperformed YOLOv7 by 5.04% on mAP@0.5, showcasing its ability to detect pedestrians in challenging scenarios, such as crowded and occluded scenes. The improved performance can be attributed to the effective integration of the Swin Transformer, Feature Pyramid Network (FPN), Occlusion-Aware Attention (OAA) module, YOLO-based detection head, and Soft Non-Maximum Suppression (Soft-NMS) into a unified detection framework.

The Swin Transformer is capable of learning rich hierarchical feature representations from both local and global contexts, and the FPN is used to boost multi-scale feature learning to detect different sized pedestrians. The proposed OAA module further enhances the detection capability focusing on the visible pedestrian regions and eliminating the background noise that is introduced due to severe occlusions. Additionally, the YOLO detection head allows to localize an object with low inference latency and low object loss, and Soft Non-Maximum Suppression effectively suppresses unwanted overlapping boxes around an object, ensuring that valid overlapping pedestrian detections are not lost.

The general form of the proposed framework is found to be a good compromise between the detection accuracy and the computational complexity. The video's performance in all evaluation metrics shows its ability to correctly identify pedestrians in the complex urban environment of dense crowds, overlapping people, and different light conditions. The outcomes show the potential applicability of the proposed framework in real-world scenarios such as intelligent transportation systems, autonomous driving, smart city surveillance, traffic monitoring, and public safety systems.

V.CONCLUSION AND FUTURE WORK

This paper proposed a Hybrid Deep Learning Framework for Pedestrian Detection in Crowded Environments which combines the Swin Transformer, Feature Pyramid Network (FPN), Occlusion-Aware Attention (OAA) module with the YOLO-based detection head and the Soft Non-Maximum Suppression (Soft-NMS). The proposed architecture successfully tackles the aforementioned challenges by being effective in detecting pedestrians in dense scenes with severe occlusions, overlapping pedestrians, and complex background. The performance evaluation showed that the proposed framework, in terms of efficiency, provided a better result in terms of Precision, Recall, F1-Score, mAP@0.5 and still had a good real-time detection capability. The proposed hierarchical feature extraction, multi-scale feature fusion, occlusion-aware attention and precision-post processing remarkably enhanced the robustness of detection compared with existing pedestrian detection methods. Thus, the proposed framework is a reliable and efficient one for smart city surveillance applications, traffic monitoring applications, public safety applications, and intelligent transportation systems with autonomous driving.

The proposed framework will be further strengthened and optimized in the future, combining lightweight backbone networks and deploying them on edge devices with limited computing power. Moreover, video-based pedestrian tracking will be considered for adaptive temporal feature learning to promote consistency in the pedestrian detection process over time. The framework can be easily extended to work under bad weather, low light and night-time surveillance conditions, and utilize multimodal sensor fusion of RGB, thermal and LiDAR data. Moreover, further enhancement of the framework for real-time

deployment on embedded AI platforms and the exploration of transformer-based self-supervised learning methods could further boost the accuracy of detection, computational efficiency, and scalability of large-scale intelligent surveillance systems.

REFERENCES

1. Li et al. (2023) – Improved YOLOv7 for Occluded Pedestrian Detection <https://www.mdpi.com/1424-8220/23/13/5912>
2. Ye and Wang (2023) – RT-YOLO: Residual Transformer YOLO for Detecting Multi-Scale Crowded Pedestrian <https://www.mdpi.com/2076-3417/13/21/12032>
3. Y. Zhang, H. He, J. Li, Y. Li, J. See, and W. Lin, “Variational Pedestrian Detection,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 11622–11631, 2021. https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_Variational_Pedestrian_Detection_CVPR_2021_paper.html
4. M. Lin, C. Li, X. Bu, M. Sun, C. Lin, J. Yan, W. Ouyang, and Z. Deng, “DETR for Crowd Pedestrian Detection,” *arXiv preprint arXiv:2012.06785*, 2020 <https://arxiv.org/abs/2012.06785>
5. Liu and Ma (2024) – Key Points and Visible Part Fusion Attention Network <https://journal.hep.com.cn/ol/EN/10.1007/s11801-024-4053-x>
6. Z. Liu et al., “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021. https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html
7. T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, 2017. https://openaccess.thecvf.com/content_cvpr_2017/html/Lin_Feature_Pyramid_Networks_CVPR_2017_paper.html
8. N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, “Soft-NMS—Improving Object Detection With One Line of Code,” *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 5562–5570, 2017. https://openaccess.thecvf.com/content_iccv_2017/html/Bodla_Soft-NMS_--_Improving_ICCV_2017_paper.html
9. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016. https://openaccess.thecvf.com/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html
10. J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 7263–7271, 2017. https://openaccess.thecvf.com/content_cvpr_2017/html/Redmon_YOLO9000_Better_Faster_CVPR_2017_paper.html
11. S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017. <https://ieeexplore.ieee.org/document/7485869>
12. [W. Liu et al., “SSD: Single Shot MultiBox Detector,” *Proc. European Conf. Computer Vision (ECCV)*, pp. 21–37, 2016. https://link.springer.com/chapter/10.1007/978-3-319-46448-0_2
13. 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,” *arXiv preprint arXiv: 2207.02696*, 2022. <https://arxiv.org/abs/2207.02696>
14. S. Zhang, R. Benenson, and B. Schiele, “CityPersons: A Diverse Dataset for Pedestrian Detection,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3221, 2017. https://openaccess.thecvf.com/content_cvpr_2017/html/Zhang_CityPersons_A_Diverse_CVPR_2017_paper.html
15. S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, “Occlusion-Aware R-CNN: Detecting Pedestrians in a Crowd,” *Proc. European Conf. Computer Vision (ECCV)*, pp. 657–674, 2018. https://openaccess.thecvf.com/content_ECCV_2018/html/Shifeng_Zhang_Occlusion-Aware_R-CNN_Detecting_ECCV_2018_paper.html
16. S. Qian, L. Chen, M. Li, and X. Wang, “WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild,” *IEEE Transactions on Multimedia*, 2021. <https://github.com/PeizeSun/WiderPerson>
17. W. Yan, Y. Chen, X. Shen, and Z. Wang, “Multi-Scale Feature Pyramid Network: A Heavily Occluded Pedestrian Detection Network Based on ResNet,” *Sensors*, vol. 21, no. 6, 2021. <https://www.mdpi.com/1424-8220/21/6/1963>