Journal of Electronic Research and Application, 2025, Volume 9, Issue 1
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

A Multi-Scale Attention-Based Pedestrian Detection Method for Roadways Using the YOLOv5 Framework

Ruihan Wang*, Boling Liu, Tingyu Liao

School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Corresponding author: Ruihan Wang, s221201032@stu.cqupt.edu.cn

Abstract: Due to multi-scale variations and occlusion problems, accurate traffic road pedestrian detection faces great challenges. This paper proposes an improved pedestrian detection method called Multi Scales Attention-YOLOv5x (MSA-YOLOv5x) based on the YOLOv5x framework. Firstly, by replacing the first convolutional operation of the backbone network with the Focus module, this method expands the number of image input channels to enhance feature expressiveness. Secondly, we construct C3_CBAM module instead of the original C3 module for better feature fusion. In this way, the learning process could achieve more multi-scale features and occluded pedestrian target features through channel attention and spatial attention. Additionally, a new feature pyramid detection layer and a new detection channel are embedded in the feature fusion part for enhancing multi-scale pedestrian detection accuracy. Compared with the baseline methods, experimental results on a public dataset demonstrate that the proposed method achieves optimal detection accuracy for traffic road pedestrian detection.

Keywords: YOLOv5; Pedestrian; Detection; Feature; Fusion

Online publication: February 13, 2025

1. Introduction

Pedestrian object detection plays a crucial role in the field of autonomous driving. It utilizes road images captured by vehicle-mounted camera equipment to identify pedestrians and accurately determine their locations by the use of computer vision techniques.

Traditional pedestrian detection methods, such as Histogram of Oriented Gradients (HOG) [1], Local Binary Pattern (LBP) [2], and Scale-Invariant Feature Transform (SIFT) [3], mainly rely on manually designed feature extraction methods. However, these methods always suffered from low accuracy performance and limited generalization ability in practice.

Object detection algorithms based on deep learning were first proposed by Girshick *et al.* in 2014, where the Region-CNN (RCNN) [4] was introduced. Subsequently, two-stage detectors such as Fast R-CNN [5], Faster R-CNN [6], and Mask R-CNN [7], as well as single-stage detectors like You Only Look Once (YOLO) [8] and Single Shot MultiBox Detector (SSD) [9], have emerged in the following ten years. Among these algorithms, YOLO performs significant performance in the field of pedestrian target detection due to its outstanding contributions to balance detection accuracy and processing speed. However, when facing multi-scale and occlusion scenarios, the YOLO method also suffers from issues such as increased leakage and false detection rates, resulting in poor detection results.

In real traffic road scenes, pedestrian object detection techniques are greatly affected by multi-scale and occlusion issues. Large-scale pedestrians often provide richer information, facilitating better detection. However, small-scale pedestrians usually reveal the following characteristics, such as reduced pixel sizes, blurred outlines, and appearances, resulting in limited information extraction effectiveness. Moreover, occlusion occurs when an object or a portion of an object obstructs another object, causing partial or complete invisibility in images or videos. Occlusion complicates the recognition of pedestrian parts, substantially increasing the difficulty of pedestrian detection.

To solve the above issues, this paper proposes MSA-YOLOv5x, an enhanced multi-scale attention network based on YOLOv5x, specifically designed to improve pedestrian detection accuracy in multi-scale and occluded scenarios.

2. Related work

2.1. Occlusion and multiscale studies in pedestrian detection

In recent years, pedestrian detection has made significant advancements in terms of detection accuracy and speed, owing to the continuous efforts and works of researchers. To address the challenge of occlusion, researchers have explored various techniques. For instance, Tian *et al.* [10] proposed DeepParts, which tackles occlusion by utilizing different partial detectors. By incorporating features from specific body parts, DeepParts aims to improve the detection of partially occluded pedestrians with higher robustness. Another technique that has been employed to address occlusion is center point detection. This approach transforms pedestrian detection into advanced semantic feature detection, which helps enhance accuracy in occlusion scenarios. One example of such a method is OAF-Net, proposed by Qiming *et al.* [11] in 2022. OAF-Net incorporates an occlusion-aware detection head, which consists of three independent centroid prediction branches along with scale and offset prediction branches.

Recently, incorporating attention mechanisms and contextual information has been widely used to further promote the detection performance of occluded pedestrians. In 2019, Chi *et al.* proposed Context-Aware Feature Learning Networks (CAFL) [12]. CAFL utilizes pixel-level contextual embedding modules to integrate contextual information from multiple surrounding regions into the feature layer. Considering the context, CAFL enhances the discriminative ability of detectors and improves the robustness of occlusions. In 2021, Jin *et al.* proposed the Mask-Guided Attention Network (MGAN) [13].

Researchers have proposed various solutions for addressing the multi-scale problem of pedestrian detection. Some studies focused on enhancing multiscale pedestrian detection by incorporating contextual information and enabling the model to adapt to pedestrians at different scales more effectively. For instance, in 2019, Xie *et al.* introduced the inverse convolution and porous module into Faster R-CNN networks to enrich the feature map with semantic contextual information [14]. This augmentation resulted in a synthesized feature map that provided more detailed visual information and semantic contextual representation. Besides, the attention mechanism has been employed to identify the correlation between raw data and emphasize significant features. Integrating the attention mechanism into pedestrian detection facilitates the fusion of different features and improves the robustness of pedestrian detection across various scales. In 2020, Lin *et al.* proposed a granularity-aware deep feature learning method (CAGDFL) that utilizes a convolutional backbone to generate multiple feature maps representing pedestrian targets at different scales [15]. Subsequently, a scale-aware pedestrian attention module is employed to generate attention maps. Moreover, the fusion of multi-scale features has been utilized to enhance the robustness of pedestrian detection. In 2022, Chao *et al.* introduced RSSD based on the SSD algorithm [16]. This approach fuses feature maps of different scales during the feature fusion process, generating six prediction layers from varying depths. Additionally, residual blocks are incorporated into each prediction layer of the SSD to enhance prediction performance.

2.2. YOLOv5

YOLOv5 is one of the most popular deep learning-based object detection algorithms. Compared to its predecessor, YOLOv4, YOLOv5 introduces Adaptive Anchor Frames and Adaptive Picture Scaling. These improvements enable the model to adapt to objects of varying sizes and proportions, thereby improving its generalization ability. Additionally, YOLOv5 incorporates the Cross Stage Partial Network (CSPNet) structure into the Backbone and Neck components, effectively reducing the number of parameters and computation required, thereby enhancing the model's efficiency [17]. In terms of post-processing, YOLOv5 utilizes the weighted Non-Maximum Suppression (NMS) approach, which effectively handles overlapping targets and eliminates duplicate detection results, leading to improved model accuracy [18]. YOLOv5 consists of the input side, Backbone structure, Neck structure, and prediction side, as depicted in **Figure 1**.

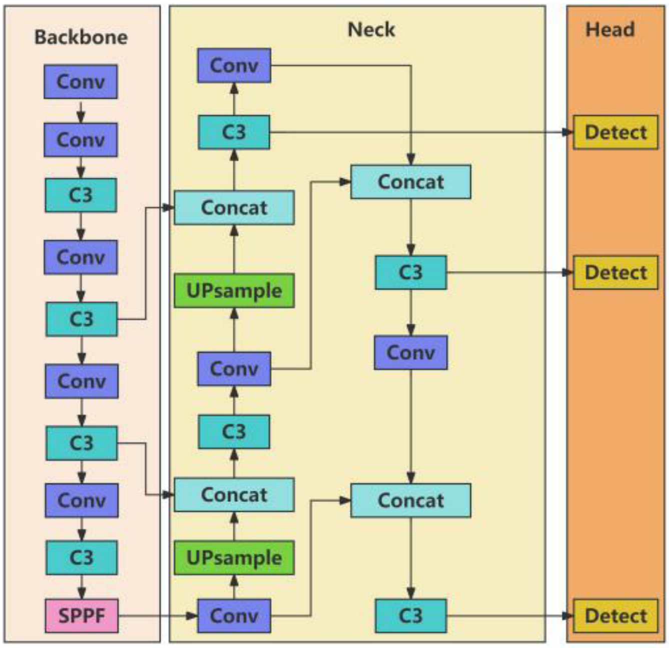

Figure 1. Yolov5x network structure

3. Method

This paper proposes a roadways pedestrian detection method with multi-scale attention based on the YOLOv5 Framework, abbreviated as MSA-YOLOv5x. The method aims to improve the accuracy of detecting partially occluded and differently-sized pedestrian targets. The structure of MSA-YOLOv5x is illustrated in **Figure 2**.

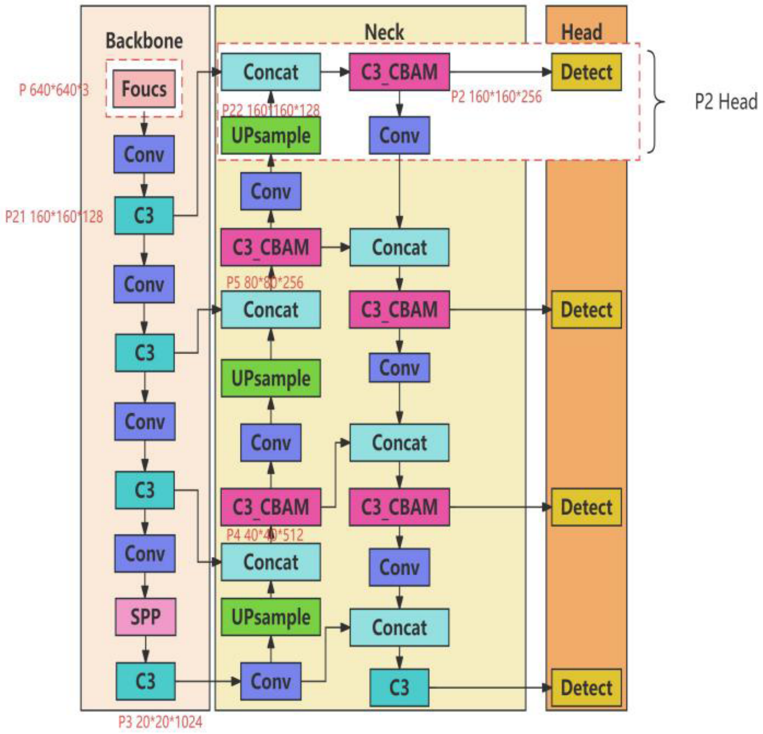

Figure 2. MSA-YOLOv5x network architecture

Within the neck network, the C3_CBAM module is constructed by incorporating the CBAM attention mechanism to replace the C3 module in the feature fusion part [19]. The feature maps are further processed by the successive up-sampling and C3_CBAM modules. They are then fused with the feature maps corresponding to the same scale in the backbone network to generate the output detection results. Additionally, to further enhance the multi-scale pedestrian detection accuracy, this method introduces a P2 feature pyramid detection layer in the feature fusion part and establishes a new detection channel in combination with the C3_CBAM module [20].

3.1. Image preprocessing

Compared to the traditional Convolution (Conv), we replace the first convolution operation of the backbone network with the Focus module. This replacement enhances the model's ability to capture local features. The Focus module performs a slicing operation on the input image within the backbone network.

3.2. C3_CBAM feature extraction module

We introduce the Convolutional Block Attention Module (CBAM) attention mechanism block to reconstruct C3_CBAM feature extraction module, enhancing the C3 structure of YOLOv5x. The C3_CBAM feature extraction module is depicted in **Figure 3(b)**. This module combines channel attention and spatial attention, significantly enhancing the method's perceptual ability. By adaptively learning these two attention types, the module enriches feature map representations, thereby improving the detection of small-scale pedestrians and partially occluded pedestrian targets. C3_CBAM introduces a CBAM attention mechanism module at the end of each C3 structure in the original YOLOv5x, with the same number of channels as the C3 structure. This new module enables the method to adaptively learn channel attention and spatial attention. It enhances the feature map representation while preserving the original feature map channels. The specific operation of C3_CBAM is as follows: For the input feature map of the C3 module, H and W denotes as with dimensions and C indicates the number of channels. It undergoes two parallel branches. In the first branch, a 3 × 3 convolution kernel is applied to generate a shallower feature map . In the second branch, the input shallow feature map is divided into two sub-branches within a BottleNeck module [21]. One sub-branch remains unchanged, while the other sub-branch is further processed by two 3 × 3.

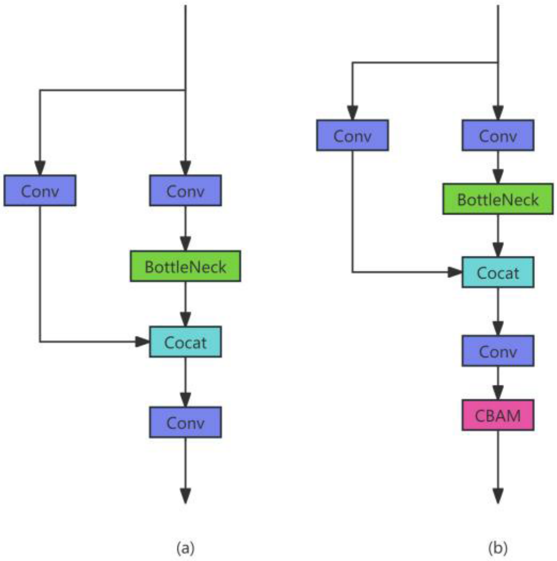

Figure 3. C3_CBAM module: (a) the original C3 module; (b) the improved C3 module with the introduction of CBAM convolution kernels

The resulting two sub-branches within the BottleNeck module are then combined using residual connections to produce a deeper feature map $F_{12} \in R^{H*W*C/2}$. The shallow feature map and the deeper feature map are concatenated to obtain the feature map F_2, which can be calculated using the following formula:

$$F_2 = \text{Concat}\,(F_{11}, F_{12}) \in R^{H*W*C} \quad (1)$$

The F_2 feature map integrates the information from the shallower and deeper feature maps, facilitating cross-layer feature fusion. This fusion enables the model to gain a better understanding of targets at various scales, thereby enhancing the accuracy of detection. Subsequently, the F_2 feature map is passed through the CBAM attention block. The CBAM attention block consists of two components: channel attention and spatial attention. First, the channel attention mechanism calculates the importance of each channel. It achieves this by employing global average pooling and fully connected layers to compute the weights for each channel. These weights indicate the significance of each channel in extracting valuable features. The weights are then multiplied with the corresponding feature maps of each channel, resulting in a one-dimensional channel attention feature map that contains channel attention information. The formula for the one-dimensional channel attention feature map, denoted as F_C, is as follows:

$$F_C = \sigma\ (MLP\ (Avg\ (F_2)) + MLP\ (Max\ (F_2))) \otimes F_2 \qquad (2)$$

The one-dimensional channel attention feature map, *Fc*, obtained from the channel attention block, is then passed into the spatial attention block to calculate the spatial attention. The spatial attention mechanism aims to learn the significance of each spatial location in the feature map. It utilizes channel max pooling and fully connected layers to compute the weights for each spatial location. These weights indicate the importance of each location in extracting valuable features. Subsequently, the weights are multiplied by the corresponding locations of the feature map, resulting in a 2D spatial attention feature map, *Fs*, that incorporates both channel attention and spatial attention information. The formula for the 2D spatial attention feature map, *Fs*, is presented below:

$$F_s = \sigma\ (f^{7*7}\ ([Avg\ (F_C)];\ [Max\ (F_2)])) \otimes F_C \qquad (3)$$

3.3. P2 feature pyramid detection layer

To detect pedestrians at multiple scales in traffic road scenes, a new *P2* feature pyramid detection layer and a new detection channel are embedded in the feature fusion part of YOLOv5x for enhancing multi-scale pedestrian detection accuracy. The structure of this module is illustrated in **Figure 2**.

Down-sampling the original image $P \in R^{640*640*3}$ by the use of both the Focus module and Conv, we can get a new feature map after passing through the first C3 module. This feature map is denoted as $P21 \in R^{160*160*128}$. Subsequently, *P3* is obtained through successive down-sampling of the feature map $P21$ using Conv in the backbone network. This process is facilitated by employing the Spatial Pyramid Pooling (SPP) feature pyramid [22]. At this stage, the feature map is referred to as $P3 \in R^{20*20*1024}$.

The *P3* feature map is fed into the neck feature pyramid, where it undergoes up-sampling and is then combined with feature maps of the same scale from the backbone network. This fusion process results in the generation of a new feature map. This feature map is denoted as $P4 \in R^{40*40*512}$.

The *P4* feature map is up-sampled and further fused with feature maps of the same scale from the backbone network, resulting in the generation of a new feature map $P5 \in R^{80*80*256}$. In this paper, we introduce a $P2$ feature pyramid layer, as an addition to the feature fusion structure of the neck network. This layer is formed by concatenating the feature map $P2 \in R^{160*160*128}$, obtained from up-sampling $P5$, with $P21$ from the backbone network. The calculation formula of the $P2$ feature map is as follows:

$$P2 = Concat\ (P21, P22) \in R^{160*160*256} \qquad (3)$$

4. Experimentation

4.1. Dataset and experimental setup

The CityPersons dataset comprises over 3,000 high-resolution real-world images, which depict diverse urban scenarios such as roads, sidewalks, intersections, and more [23]. It provides a comprehensive representation of urban environments, capturing various elements like different weather conditions (sunny, cloudy, rainy, etc.), periods (daytime, nighttime), and population densities (busy city centers, suburbs, etc.). In this paper, the experimental evaluation metrics include mAP and GFLOPS.

4.2. Experiment 1: ablation study

In this paper, we introduce the CBAM attention mechanism to enhance the feature extraction structure of the

YOLOv5x backbone network. Additionally, we incorporate an extra *P2* feature pyramid detection layer to reduce the false positives and false negatives of pedestrian targets at various scales. To evaluate the effectiveness of these improvements, we conduct three sets of ablation experiments using the CityPersons dataset. The experimental results are presented in **Table 1**.

4.3. Experiment 2: comparative analysis

To validate the detection performance of the improved method, this paper compares and analyzes MSA-YOLOv5x with the baseline methods in terms of mAP, GFLOPS, and the number of parameters on the CityPersons test set. **Table 2** presents the baseline methods utilized in the experiments, comprising the two-stage target detection method Faster R-CNN [6], the one-stage target detection method SSD [9], YOLOv3 [21], YOLOv4 [10], YOLOv5x, and YOLOv8x. These methods are evaluated for their detection accuracy on the CityPersons dataset specifically for pedestrian detection.

Table 1. Results of ablation experiments on the CityPersons dataset

Methods	Params	GFLOPS	mAP50 (%)	mAP50:95 (%)
YOLOv5X	86.21M	203.9	59.7	35.5
YOLOv5X+P2	90.88M	313.2	64.2	39.5
MSA-YOLOv5x	91.22M	314.1	65.3	40.4

Table 2. Comparative results of various algorithms on the CityPersons dataset

Methods	Params	GFLOPS	mAP50 (%)	mAP50:95 (%)
Faster-RCNN	41.5M	207.1	43.9	19.0
SSD	34.3M	386.2	31.4	11.3
YOLOv3	103.69M	283	64.8	40.5
YOLOv4-CSP	52.52M	119.8	64.6	40.4
YOLOv5x	86.21M	203.9	59.7	35.5
YOLOv8x	68.15M	258.1	62.8	40.8
MSA-YOLOv5x(ours)	91.22M	314.1	65.3	40.4

5. Discussion

The MSA-YOLOv5x method demonstrated significant improvements in pedestrian detection, especially for occluded and multi-scale pedestrians, thanks to the integration of the CBAM attention mechanism and the P2 feature pyramid detection layer. The CBAM module helped the model focus on the most relevant features by applying channel and spatial attention, while the P2 layer enhanced multi-scale detection. When compared to other models like Faster R-CNN and SSD, MSA-YOLOv5x outperformed them in both detection accuracy (mAP) and computational efficiency (GFLOPS). The results suggest that the model is well-suited for real-time pedestrian detection in complex urban settings, such as in smart city applications or autonomous vehicles.

6. Conclusion

This study introduced MSA-YOLOv5x, a novel pedestrian detection model that enhances YOLOv5x with the CBAM attention mechanism and a P2 feature pyramid detection layer. Experimental results on the CityPersons dataset showed that MSA-YOLOv5x improves detection accuracy, particularly in challenging environments with occlusions and varying pedestrian sizes. Compared to baseline models, MSA-YOLOv5x achieved higher performance in both mAP and GFLOPS, making it a promising approach for real-time pedestrian detection in urban scenarios. Future work can focus on further optimizations and testing in diverse environments.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Dalal N, Triggs B, 2005, Histograms of Oriented Gradients for Human Detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1. IEEE, 2005: 886–893.

[2] Ahonen T, Hadid A, Pietik¨ainen M, 2004, Face Recognition with Local Binary Patterns. Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Proceedings, Part I 8, Springer, 2004: 469–481.

[3] Lowe DG, 2004, Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60: 91–110.

[4] Girshick R, Donahue J, Darrell T, et al., 2014, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580–587.

[5] Girshick R, 2015, Fast r-cnn. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440–1448.

[6] Ren S, He K, Girshick R, et al., 2015, Faster r-cnn: Towards Realtime Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems, 28: 2015.

[7] He K, Gkioxari G, Doll´ar P, et al., 2017, Mask r-cnn. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961–2969.

[8] Redmon J, Divvala S, Girshick R, et al., 2016, You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779–788.

[9] Liu W, Anguelov D, Erhan D, et al., 2016, SSD: Single Shot Multibox Detector. Computer Vision–ECCV 2016: 14th European Conference Proceedings, Part I 14. Springer, 2016: 21–37.

[10] Tian Y, Luo P, Wang X, et al., 2015, Deep Learning Strong Parts for Pedestrian Detection. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1904–1912.

[11] Li Q, Su Y, Gao Y, et al., 2022, Oaf-Net: An Occlusion-Aware Anchor-Free Network for Pedestrian Detection in a Crowd. IEEE Transactions on Intelligent Transportation Systems, 23(11): 21291–21300.

[12] Fei C, Liu B, Chen Z, et al., 2019, Learning Pixel-Level and Instance-Level Context-Aware Features for Pedestrian Detection in Crowds. IEEE Access, 7: 94944–94953.

[13] Xie J, Pang Y, Khan MH, et al., 2020, Mask-Guided Attention Network and Occlusion-Sensitive Hard Example Mining for Occluded Pedestrian Detection. IEEE Transactions on Image Processing, 30: 3872–3884.

[14] Xie H, Chen Y, Shin H, 2019, Context-Aware Pedestrian Detection Especially for Small-Sized Instances with Deconvolution Integrated Faster Rcnn (dif r-cnn). Applied Intelligence, 49: 1200–1211.

[15] Lin C, Lu J, Wang G, et al., 2018, Graininess-Aware Deep Feature Learning for Pedestrian Detection. Proceedings of the European conference on Computer Vision (ECCV), 2018: 732–747.

[16] Yan C, Zhang H, Li X, et al., 2022, R-ssd: Refined Single Shot Multibox Detector for Pedestrian Detection. Applied Intelligence, 52(9): 10430–10447.

[17] Wang CY, Liao HYM, Wu YH, et al., 2020, Cspnet: A New Backbone that can Enhance Learning Capability of Cnn. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 390–391.

[18] Felzenszwalb PF, Girshick RB, McAllester D, et al., 2009, Object Detection with Discriminatively Trained Part-Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9): 1627–1645.

[19] Woo S, Park J, Lee JY, et al., 2018, Cbam: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3–19.

[20] Redmon J, Farhadi A, 2018, Yolov3: An Incremental Improvement. arXiv Preprint, arXiv:1804.02767.

[21] He K, Zhang X, Ren S, et al., 2016, Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.

[22] He K, Zhang X, Ren S, et al., 2015, Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904–1916.

[23] Zhang S, Benenson R, Schiele B, 2017, Citypersons: A Diverse Dataset for Pedestrian Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3213–3221.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.