Pedestrian Detection Using YOLO with Improved Attention Module

Truong Quang Vinh
Faculty of Electrical and Electronics Engineering
Ho Chi Minh City University of Technology, Vietnam National University - Ho Chi Minh
Ho Chi Minh City, Vietnam
tqvinh@hcmut.edu.vn

Pham Hien Long
Faculty of Electrical and Electronics Engineering
Ho Chi Minh City University of Technology, Vietnam National University - Ho Chi Minh
Ho Chi Minh City, Vietnam
phlong.sdh19@hcmut.edu.vn

***Abstract*—Pedestrian detection is considered as an important component of intelligent traffic management systems. This paper presents an improved Yolov5 's framework for pedestrian detection. We propose two new attention modules for Yolov5 architecture to highlight significant information. First, the modified Efficient Channel Attention (M-ECA) was applied in the backbone of Yolo network to collect the useful features. Second, the modified Global Attention Mechanism (M-GAM) was inserted in the head of Yolo network to enhance feature representation. Thanks to two new improved attention modules, the proposed model achieves the high accuracy for overlapped pedestrian detection in crowded streets. The proposed model was trained with Penn-Fudan dataset. The experimental result shows that the proposed model has better detection accuracy and mAP, comparing to the original Yolov5.**

Keywords— Pedestrian Detection, Yolov5, Penn-Fudan, GAM, ECA

I. INTRODUCTION

Due to advancements in deep learning, convolutional neural network (CNN) imaging, and audio processing, there has been a growing interest in pedestrian detection in recent years. This interest stems from the fact that pedestrian detection has long been a challenging problem in the field of computer vision, prompting numerous studies [1]. In the past, several scholars have dedicated their efforts to addressing this issue[2-9]. Pedestrian detection is considered a crucial objective in object detection as it focuses on accurately predicting the bounding boxes of pedestrians in images. This particular area of research has gained significant attention within the computer vision community due to its growing significance in various applications, including self-driving cars, personnel re-identification, video surveillance, and robotics.

The evolution of pedestrian detection can be traced from handcrafted feature methods, such as method utilizes the feature in channel, to the advancement of deep learning techniques. Deep learning-based detectors can be categorized into two types: one-stage detectors and two-stage detectors. Two-stage detectors, such as Region-based Convolutional Neural Networks (RCNN), Fast RCNN, and Faster RCNN, have demonstrated superior accuracy [1]. Conversely, one-stage detectors like Yolo system, Single Shot Detection (SSD), and RetinaNet have excelled in terms of object detection speed [2]. Researchers continue to strive for pedestrian detectors that offer optimal performance, encompassing both accuracy and speed.

Yolo series was initialized in 2015, which was applied in object detection. Many studies had applied Yolo to detect pedestrians. One of the most difficult issues for pedestrian detection is to separate overlapping image regions of pedestrians. Hsu and Lin proposed adaptive fusion of multi-scale Yolo technique to overcome this issue [3]. Some researches focus on tracking of pedestrians by using both Yolov5 and Deep-Sort [4][5]. It is possible to improve the precision by modifying the prediction part of Yolo architecture. Peng et al. added Gaussian mixture model (GMM) detection together with Yolo to detect small pedestrian objects in the large images [6]. Yun and Kim added Long Term Short Memory (LTSM) module into Yolov5 network to achieve more accuracy while maintaining low computational cost [7]. Hsu and Lin proposed intelligent split algorithm make the model reduce the loss object proportion [8]. Another approach is to insert channel attention and spatial attention for recalibrating the weight adaptively and determining significant image regions [9].

The previous methods of Yolo based pedestrian detection have incorporated some techniques to improve the prediction process. This paper presents another approach to emphasize significant features on pedestrian detection using Yolov5 framework. We propose two improvements to ECA and GAM, which can extract both global information and local information at the same time. It helps Yolo model collect the useful features and enhance feature representation. The proposed model is trained with Penn-Fudan dataset. The Precision, F1-score, and mAP metrics are utilized to evaluate the performance of network.

The paper is structured as follows. Section II presents our proposed Yolov5 based method with two new M-ECA and M-GAM module. Section III presents the experimental results with specific training dataset. Finally, Section IV presents the conclusion and the future work.

II. PROPOSED PEDESTRIAN DETECTION METHOD

979-8-3503-3122-6/23/$31.00 ©2023 IEEE
DOI 10.1109/ACOMPA61072.2023.00024

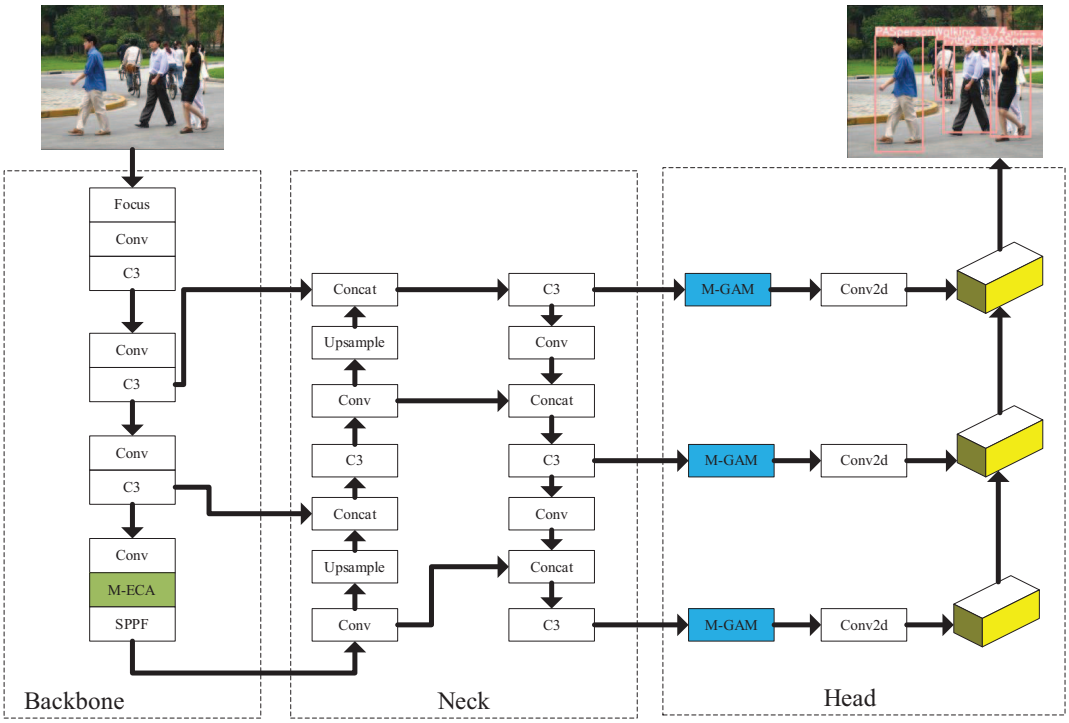


Fig. 1. Proposed Yolov5 Network Architecture with Modified GAM (M-GAM) and Modified ECA (M-ECA)

The proposed model architecture is based on Yolov5 with two new aspects as shown in Fig. 1. In our proposed model architecture, Yolov5 is chosen to improve. We have experimented with Yolov5, Yolov7, and Yolov8 on Penn-Fudan dataset. Among those models, Yolov5-version7 yielded the best results in terms of precision and recall, because this model has been improved with news segmentation models and classification models [10], [22]. Yolov5 has been updated to version 7 in November 2022, whereas the model update for Yolov7 is not continuous [11]. As for Yolov8, it is a relatively new version developed by Ultralytics [12].

Based on the Yolov5 model, we propose two new modules including the modified Efficient Channel Attention (M-ECA) and the modified Global Attention Mechanism (M-GAM). In the backbone part, C3 layer, which comprises three convolution layers, is used to extract the features from image frames. However, it may not distinguish significant features from the others. The M-ECA layer is proposed to replace C3 layer to focus on important features and ignore confusing information. In the head part, M-GAM layer was added before Conv2D layer to extract key information about the small targets or the overlapped targets in pedestrian detection. The detailed description of M-ECA layer and M-GAM layer is described as follows.

A. Modified Efficient Channel Attention (M-ECA)

The Efficient Channel Attention (ECA) [13] not only guides the neural network model where to focus but also enhances the representativeness of the network interests. The background environment to detect pedestrian is complex, sometime many pedestrians walk overlapped in the crowded street. In order to improve the accuracy, we need to collect the useful features. In this paper, we propose two modifications of ECA, and then append it to the backbone part, after C3 layer as shown in Fig. 2. Modified ECA shall enhance module by ignoring the fusion information and focusing on the significant feature.

First, we choose the best value of kernel size k by experimenting with the range from 1 to 7. In the original ECA module, the kernel size k controls the extent of cross-channel interaction for MLP, given by the equation (1).

$$k = \varphi(C) = \left|\frac{\log_2 C}{b} + \frac{\gamma}{b}\right| \qquad (1)$$

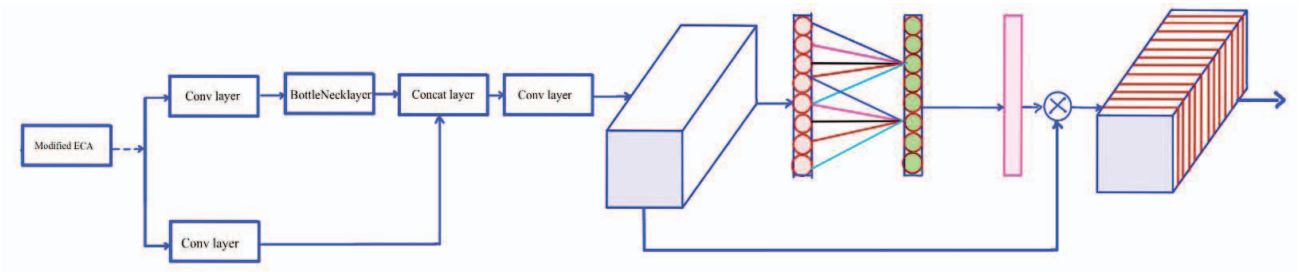


Fig. 2. Structure of modified ECA module connect with C3 module (M-ECA). Sigmoid was replaced with H-swish and value of ratio k is 3

C is the dimensions of the channel, and thus k is equal to 5, corresponding with the size of input frames in Yolov5. According to the result in Fig. 3, when k is equal to 3, the result yields the best performance with precision and recall around 0.8.

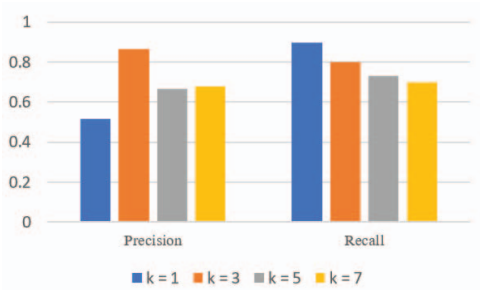


Fig. 3. Chart table of the result of the experiment to determine ratio k in ECA

Second, we replace Sigmoid activation with H-swish activation in the ECA module. If we use the Sigmoid function, as shown in Fig. 4, when the input value of the neurons network is too high, the output value from derivation is equal to 0. As a result, the gradient is eliminated, which may have a bad effect on the training result. H-swish can solve this problem. Based on derivation of H-swish shown in Fig. 5, when the input value of the neural network is too high, the corresponding gradient is not equal to 0, and then the process for updating the weight values is still meaning.

B. Modified Global Attention Mechanism (M-GAM)

In the Global Attention Mechanism (GAM) [14], the spatial attention and channel attention compute weights for distinct feature dimensions, effectively highlighting significant information in both aspects, thereby enhancing feature representation.

In this paper, we propose a modified GAM module not only to retain important information on both channel and spatial aspect, but also to enhance the cross-dimension interaction. In the original version of GAM, after spatial

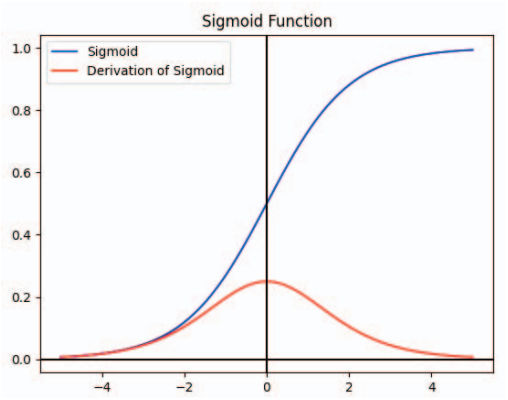


Fig. 4. Function and derivation of Sigmoid.

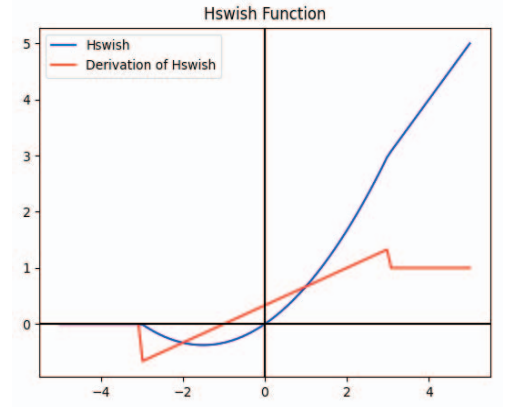


Fig. 5. Function and Derivation of H-swish.

attention and channel attention phase, the number of parameters is increased much. Therefore, we adopt the channel shuffle [15] group before providing the output F3, as shown in Fig. 6. For the number of groups channel shuffle layer, it is impacted with classification score. The model is trained with different numbers of groups g to determine the best value of g. Based on the result in Fig. 7, we choose the number of group channels in Shuffle Layer as 4, because it gains the best result with precision and recall.

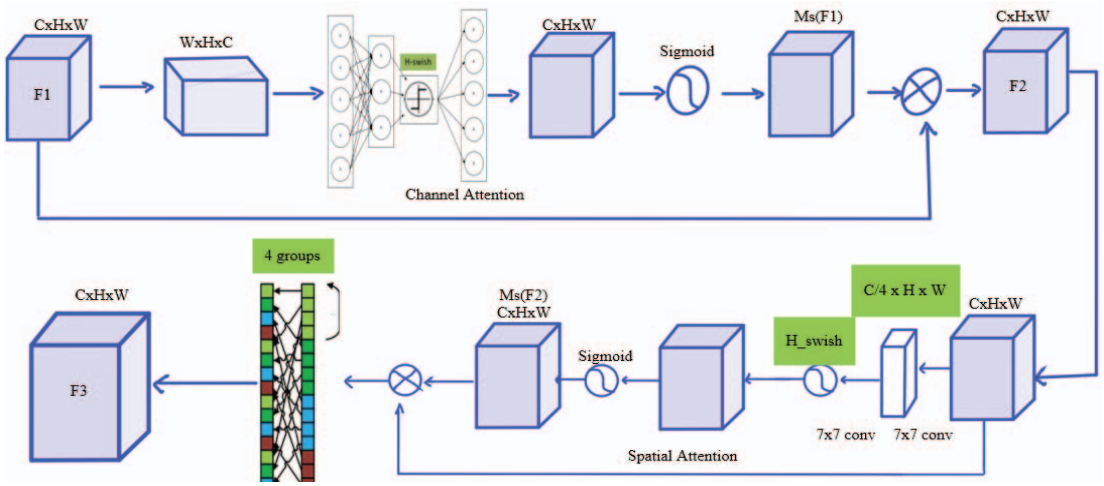


Fig. 6. Structure of Modified GAM (M-GAM) with adding shuffle layer, changing the ratio size of CNN, and replacing activate function

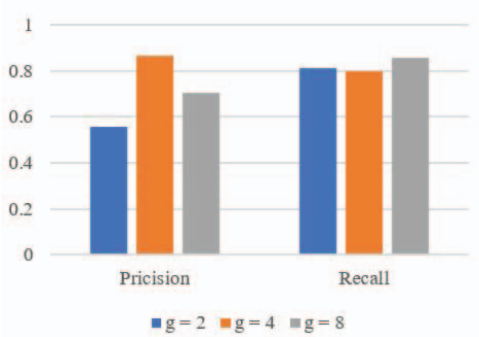


Fig. 7. Chart table of the number of groups *g* in Channel Shuffle relevant to P and R.

In the spatial attention phase of the GAM (Global Attention Module), there is a CNN layer with a kernel size ratio of 16. However, we will conduct experiments to select the best ratio for better pedestrian detection performance. Based on the result in Fig. 8, the ratio of 4 gains the best results with precision and recall.

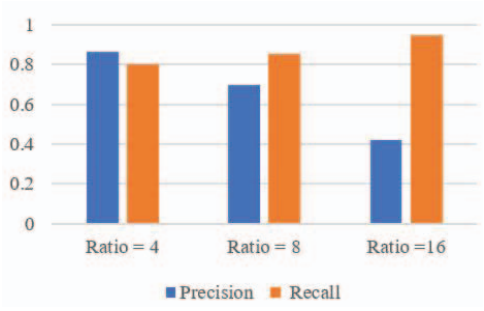


Fig. 8. Chart table of the experiment to determine Ratio in GAM.

Finally, the activation function ReLU is replaced with H-swish in MLP from channel attention and CNN layer of spatial attention. This modification fixes the disadvantages of the ReLU function. In the ReLU function, when the input value is less than 0, the gradient was 0, as shown in Fig. 9. The activate function shall make the output value be 0, which can impact the result of training. H-swish function can solve this issue. Based on Fig. 5, the gradients of H-swish are a non-zero values within [-3 0]. Therefore, the output value also has a non-zero result with input values in the range of [-3 0].

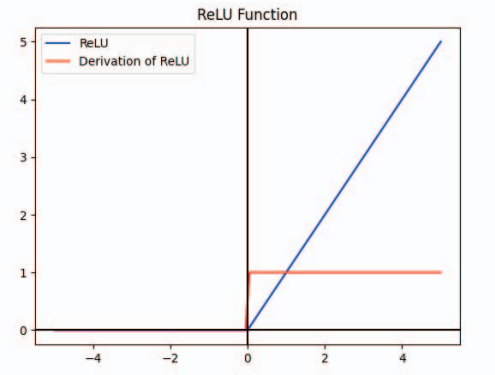


Fig. 9. Function and derivation of ReLU

III. Experimental Results

A. Dataset

In our experiment, we used dataset Penn-Fudan [16], which included 170 images with 345 labeled pedestrians

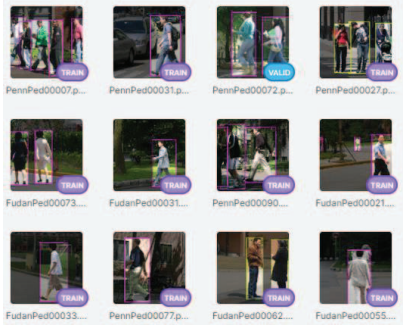


Fig. 10. Dataset Penn-Fudan [16]

The labeled pedestrians in this database have heights ranging from 180 to 390 pixels, and all of them are depicted in an upright position. There are 120 images which used for training and 50 images which used for validation. Many research papers about pedestrian detection and tracking use this dataset [17], [18], [19].

B. Experimental Results

We trained our proposed model by Google Collaboratory with 420 epochs, learning rate 0.01, batch size 16. We compared the performance of the proposed model with two other methods including original Yolov5s-version7, original Yolov5s-version6.2, and Gou's method [19]. The average training time to complete was around 1-2 hours. The evaluation metrics in our experiment are Precision, Recall, Mean Average Precision, and F1-score which were defined in [19]. The experimental result is shown in Table I.

TABLE I. Comparison Results

Algorithm	Precision	Recall	map@0.5	map@0.95	F1-Score
Yolov5s-version6.2 [10]	0.68	0.69	0.72	0.52	0.68
Yolov5s-version7 [10]	0.821	0.815	0.863	0.563	0.82
Gou's method [19]	0.85	0.72	0.82	0.54	0.78
Proposed method	0.865	0.8	0.874	0.616	0.83

According to the result in Table I, the precision of proposed method higher 4.4 % than original Yolo5s- version7, 1.5% than Gou's method [19], and 17% than original Yolo5s-version6.2. The recall lower than original Yolo5s-version7 1.5% but higher than Gou's 8% and 11% than original Yolo5s-version6.2. Our proposed method yields the highest F1-score, map@0.5, and map@0.95. Thanks to the modified ECA and modified GAM were applied in the backbone and the head of Yolo network, the proposed method outperforms the other method.

In order to evaluate the convergence rate of the proposed method, we compare the loss functions of the training processes of the proposed method and original Yolov5s-version7. There are two different types of loss functions including box loss and object loss. The box loss represents the

quality of the algorithm whether the algorithm can locate the center of an object. This box loss shows how well the predicted bounding covers the object. The object loss is used to measure the probability that an object exists in a proposed region. According to comparison result as shown in Fig.11, the box loss is almost the same between proposed method (Fig.11c) and original Yolov5s-version7 (Fig.11a). However, the object loss of the proposed method (Fig.11d) is more stable than the one of the original Yolov5s-version7 (Fig.11b).

(a) (b) (c) (d)

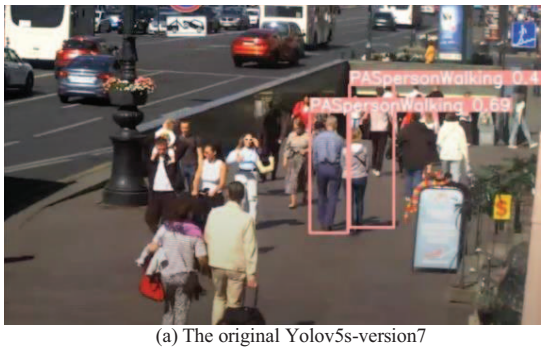
Fig. 11. Comparison of the loss function result with the original Yolov5s-version7 and the proposed method. Picture (a) and (b) obtain the box loss function and object loss function of original Yolov5s-version7. (c) and (d) are shown the result of proposed method.

The proposed method is deployed with pictures which was captured by the public camera [20], [21] with the IOU 0.4 and Conf 0.45. Fig.12, Fig.13 and Fig.14 show the comparison between the result of original Yolov5s-version7 and proposed method. The image regions of pedestrians in the frames are overlapped in crowded street. These overlapped regions of pedestrian consist body parts and luggage from other pedestrians. Therefore, such kind of images are difficulty in pedestrian's detection. The result indicates that almost the pedestrians in images cannot be detected by original Yolov5s-version7. The miss rate in Fig.12(a), Fig.13(a) and Fig.14(a) is 92%, 90%, and 65% respectively. The miss rate of the proposed method in Fig.12(b), Fig.13(b) and Fig.14(b) is 62.5%, 57%, and 50% respectively. As a result, the proposed method can detect the pedestrians inside the overlapped regions of the images. The resources of our experiments are uploaded at [22].

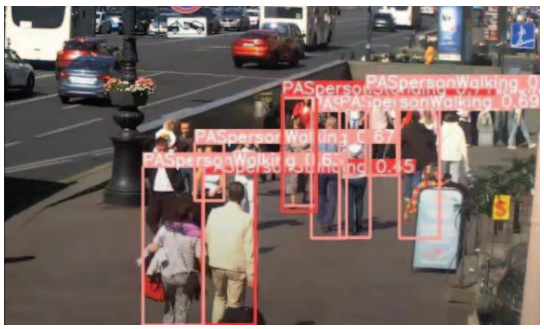
(a) The original Yolov5s-version7

(b) The proposed method

Fig. 12. Comparison of the pedestrian detection result for the crowded street with high several scattered pedestrians



(a) The original Yolov5s-version7



(b) The proposed method

Fig. 13. Comparison of the pedestrian detection result for the crowded street with high density of pedestrians.

(a) The original Yolov5s-version7

(b) The proposed method

Fig. 14. Comparison of the pedestrian detection result for the intersection with high density of pedestrian.

IV. Conclusion

This study presented the improved model Yolov5 for pedestrian detection, which incorporates two new modules including the modified Global Attention Mechanism and the modified Efficient Channel Attention module. The proposed model yields better results than the original model in detecting overlapped pedestrians in crowded streets, although there is still a limitation in detecting a pedestrian far way from the camera with a very small anchor box. The proposed model can be applied to autonomous driving and traffic management systems. In the future, we will integrate the proposed model to the intelligent traffic ligh control and implement the system on GPU.

ACKNOWLEDGMENT

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

REFERENCES

[1] Cao, Jiale, et al. “From Handcrafted to Deep Features for Pedestrian Detection: A Survey.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 9, Institute of Electrical and Electronics Engineers (IEEE), Sept. 2022, pp. 4913–34. *Crossref*, doi:10.1109/tpami.2021.3076733.

[2] E. R. Vikram Reddy and S. Thale, "Pedestrian Detection Using YOLOv5 For Autonomous Driving Applications," 2021 IEEE Transportation Electrification Conference (ITEC-India), New Delhi, India, 2021, pp. 1-5, doi: 10.1109/ITEC-India53713.2021.9932534.

[3] W. -Y. Hsu and W. -Y. Lin, "Adaptive Fusion of Multi-Scale YOLO for Pedestrian Detection," in IEEE Access, vol. 9, pp. 110063-110073, 2021, doi: 10.1109/ACCESS.2021.3102600

[4] Y. Gai, W. He and Z. Zhou, "Pedestrian Target Tracking Based On DeepSORT With YOLOv5," *2021 2nd International Conference on Computer Engineering and Intelligent Control (ICCEIC)*, Chongqing, China, 2021, pp. 1-5, doi: 10.1109/ICCEIC54227.2021.00008

[5] Y. Wang and H. Yang, "Multi-target Pedestrian Tracking Based on YOLOv5 and DeepSORT," *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, Dalian, China, 2022, pp. 508-514, doi: 10.1109/IPEC54454.2022.9777554M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[6] Q. Peng *et al*., "Pedestrian Detection for Transformer Substation Based on Gaussian Mixture Model and YOLO," *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, China, 2016, pp. 562-565, doi: 10.1109/IHMSC.2016.130.

[7] S. Yun and S. Kim, "Recurrent YOLO and LSTM-based IR single pedestrian tracking," *2019 19th International Conference on Control, Automation and Systems (ICCAS)*, Jeju, Korea (South), 2019, pp. 94-96, doi: 10.23919/ICCAS47443.2019.8971679.

[8] W. -Y. Hsu and W. -Y. Lin, "Ratio-and-Scale-Aware YOLO for Pedestrian Detection," in *IEEE Transactions on Image Processing*, vol. 30, pp. 934-947, 2021, doi: 10.1109/TIP.2020.3039574

[9] Jung, H.-K.; Choi, G.-S. Improved Yolov5: Efficient object detection using drone images under various conditions. Applied Sciences 2022, 12, 7255.

[10] Ultralytics. “Releases · Ultralytics/Yolov5 · GitHub.” GitHub, https://github.com/ultralytics/yolov5/releases. Accessed 29 July 2023.

[11] Wang, Chien-Yao, Alexey, Bochkovskiy, and Hong-Yuan Mark, Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors".arXiv preprint arXiv:2207.02696 (2022).

[12] Ultralytics. “Comprehensive Guide to Ultralytics YOLOv5 - Ultralytics YOLOv8 Docs.” Home - Ultralytics YOLOv8 Docs, https://docs.ultralytics.com/yolov5/. Accessed 15 July 2023.

[13] Wang, Qilong, Banggu, Wu, Pengfei, Zhu, Peihua, Li, Wangmeng, Zuo, and Qinghua, Hu. "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks." . In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.2020.

[14] Liu, Yichao, Zongru Shao and Nico Hoffmann. *“Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions.” ArXiv abs/2112.05561 (2021): n. pag*

[15] Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2018**.

[16] Applied Science, Department of the School of Engineering and. “Pedestrian Detection Database.” *Computer and Information Science | A Department of the School of Engineering and Applied Science*, https://www.cis.upenn.edu/~jshi/ped_html/. Accessed 26 June 2023

[17] P Y. Bo and C. C. Fowlkes, "Shape-based pedestrian parsing," CVPR 2011, Colorado Springs, CO, USA, 2011, pp. 2265-2272, doi: 10.1109/CVPR.2011.5995609.

[18] Junaid, Mohammad, et al. “Evaluation of Non-Classical Decision-Making Methods in Self Driving Cars: Pedestrian Detection Testing on Cluster of Images with Different Luminance Conditions.” Energies, no. 21, MDPI AG, Nov. 2021, p. 7172. Crossref, doi:10.3390/en14217172.

[19] W. Guo, N. Shen and T. Zhang, "Overlapped Pedestrian Detection Based on YOLOv5 in Crowded Scenes," 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA), Changchun, China, 2022, pp. 412-416, doi: 10.1109/CVIDLICCEA56201.2022.9825055.

[20] Russia, Mobotix Webcams. LIVE Nevskiy Avenue St. Petersburg Russia, Gostiny Dvor. Невский Пр. Санкт-Петербург, Гостиный Двор. YouTube, 26 Jan. 2021, https://www.youtube.com/watch?v=h1wly909BYw.

[21] Tokyo Shibuya Scramble Crossing - 4 Shocking Events Caught On Street Cam. YouTube, 2 Jan. 2021, https://www.youtube.com/watch?v=oLkQx2tYCpQ.

[22] “GitHub - MAK1647/ACOMPA2023-Submission-1922.” GitHub, https://github.com/MAK1647/ACOMPA2023-Submission-1922. Accessed 24 Sept. 2023.