International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhanced Yolov5 Deep Learning Technique for Multi-Object Detection in Autonomous Vehicle in Extreme Weather Condition

Ranjitha P[1], Saira Banu Atham[2]

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: Autonomous vehicle requires more accurate object detection and object classification for the real-world environment. Detection and classification of objects in the stable environment using deep learning framework yields the very efficient result. But in real situation due to rapid changes in the real environment object detection in harsh condition fails due to the various challenges in it. In the proposed work considered two types of dataset image [KITTI] and video [BDD] with nine various environment condition -normal, night, rain, rain with night, low light, high illumination, cluttered environment, FOG, high speed. Applied enhanced robust deep learning algorithm YOLOV5 to all the above conditions and results are compared with the accuracy with the previous work. From the enhanced Yolov5 resulted in 98.7%, 76%, 76%,75%, 71% in normal, rainy, haze, high illumination, night weather condition respective. Proposed model is giving significant improvement in accuracy, precision, recall than the previous work in the extreme weather condition.

Keywords: *CNN, R-CNN, YOLOV5, Single shot object detection, Multi object detection*

1. Introduction

Autonomous vehicle is the type of vehicle where it captures its surrounding environment then process it to take its own decision and aims to drive safely in the road without any human input .Autonomous vehicle efficiency depends on the result of two major components , first one is hardware component where consists of various kinds of sensors to capture its surrounding environment and next component is computing software where it need to process the video/image captured from the different advance sensors. In processing the capture video, the main task is object detection and classification. Since Autonomous vehicle is running in the real environment it needs to detect and classify the multi object in the road like various vehicle (Bus, Car, Bicycle, Van), traffic sign, lane detection, pedestrian and others objects on road.

For object detection there are n number of algorithms available in machine learning and deep learning such as CNN, R-CNN, Faster R-CNN, Yolo etc[1]. Where all these algorithms work well and yield the best efficiency. With the help of advanced sensor and high computing power of software technology the Autonomous vehicle has achieved the best results towards the multi object detection and classification but it is still lacking in terms of extreme environmental conditions such as low illumination, high illumination, sunny, rain, night vision, scattered environment [2]. There are various challenges. There are n objects in the frame, but the model will detect only a few. When a multi-object is in the frame, the model fails to see the entire Object in the given frame. In the night vision and low light frames due to low pixel value results in very less object detection rate. In sunny or high illumination frames due to high pixel and occlusion results in very less average object detection rate [3]. The purpose of this research is to investigate the efficiency of object detection and classification of image /Video of autonomous vehicle in the different extreme condition as seen above and improving the prediction rate by using enhanced Yolov5 deep learning model.

The performance of several object identification algorithms, this paper compared different types of detectors, comprising multiple-stage detectors with single-stage detectors. Also, YOLOv5 is employed to perform single shot object detectors. Real-world datasets that are accessible to the public are used to test the object detection algorithms. Accuracy is used to compare how well detection algorithm's function objects. When compared to the other convolution neural networks, the enhanced YOLOv5 approach is used. The work is organized as: section 2 gives wider analysis of various prevailing approaches. The methodology is demonstrated in section 3 with outcomes in section 4. The conclusion is provided in section 5.

[1] *Assistant Professor, School of Computer Science and Engineering & Information Science, Department of Computer Science & Engineering, Presidency University Bangalore, India-*
ORCID ID: 0000-0002-1046-2319
[2] *Professor, School of Computer Science and Engineering & Information Science, Department of Computer Science & Engineering, Presidency University Bangalore, India-*
ORCID ID: 0000-0002-4469-5153
** Corresponding Author Email: ranjitha.p@presidencyunivercity.in*

1.1. Abbreviations and Acronyms

Abbreviation and Acronyms	*Meaning*
Yolov5	You only look once version 5
CNN	Convolutional neural network

2. Literature Survey

Since Autonomous vehicles need to run on the road it needs to detect and classify objects in the rapidly changing weather condition like heavy fog, sleeting rain, snowstorms, dusty blasts and low light conditions. But to perform this first task here is to detect the various weather condition. The deep learning-based framework is used to classify the adverse and the normal weather condition. To do this task three algorithms are used in CNN: Squeeze net, RestNet2020, MCWRD2018 which resulted in 98.48% of accuracy, 98.51% of precision, 98.41% of sensitivity [4]. [5] Taken clearly annotated visual data and rainy annotated visual data were applied with domain adaption, R-CNN and yolo on BDD100K data. Using de-raining algorithm degrade the average precision performance when tested on scene distorted by natural rain, improvements can be achieved while employing image to image translation and domain adaption as mitigating techniques. Yolo has achieved 37% average precision rate. [6] object detection in the harsh environment condition like snow, rain, fog is crucial for the autonomous vehicle. Used Dense Net for Adversarial defense module (ADM) on COCO2015 and BDD100K dataset resulted in 43.7% mean average precision for CoCo2015 and 39.0% for the BDD100K dataset.

[7] Localizing the various objects like pedestrians, different vehicles, traffic signals and barriers in the road in the real time is the open challenges for the autonomous vehicles. In that there are two types of networks i.e. Two stage detection (R-CNN, Fast R-CNN, Faster R-CNN) and single stage detectors (Yolo, SSD, Retina net). To types of data considered here first 2D object detector and 3D object detectors (Monocular image based, Point cloud based, Point nets based). [8] To detect Object in the normal condition various algorithms used such as VGG16, Fast R-CNN, Simple net, yolo were resulted in 47s, 0.35s, 0.09s, 0.022s time to predict the single image respectively.

3.Methodology

The methodology is elaborated in this section where the performance of the anticipated model is compared with other approaches and the significance is demonstrated in section 4.

3.1 Dataset

This work considers data collected from two types of datasets: KITTI (Image data set) and the BDD dataset [9]. In the KITTI dataset, images are divided into two categories, i.e., Low light and standard condition images. In the BDD dataset, a total of 1000 videos were present, and each video ranges from 30 to 50 and contains various environmental conditions. From that, 9 video categories are divided based on different environmental conditions, shown in Table 1. Figure 1 illustrates how the suggested paradigm functions overall. The first step is to collect data from BDD and KITTI where it is divided into nine different environmental condition videos. Later, pre-processing techniques (noise removal, removing the corrupted part of the video) are applied to the videos. After the pre-processing technique, the model is tested in the trained enhanced yolo5 model, and the accuracy is noted for each environmental condition video.

Table 1: Displays the different BDD dataset classifications based on other environmental circumstances.

SL No	*Various Condition*
1	Normal
2	Night
3	Rain
4	Rain + night
5	Low light
6	High illumination
7	Cluttered Environment
8	Snow + night
9	Rain + night +snow

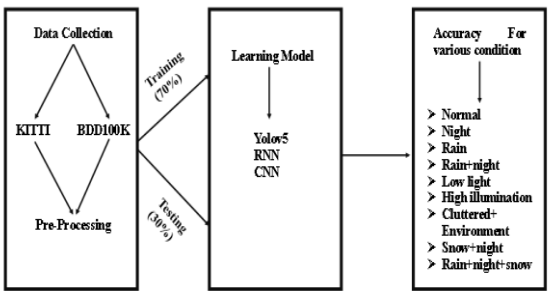


Fig 1: Block Diagram of the Proposed System

3.2. Yolov5

Yolov5 (you only look once) is the computer vision model for detecting the objects in the video or images. The model comes in various versions. Depending on the time taken to train the dataset, different versions of enhanced yolov5 are used. There are four different sizes available: small, medium, big, and extra-large. Here, an extra-large version of yolov5 is used for the proposed work. Yolov5 architecture is designed so that first, it creates the various features from the input image where that feature is passed to the prediction system, making the box around the feature for predicting the classes. The figure shows the overall working of yolov5 in the autonomous vehicle as a first-step input taken in the video format, where it will be converted into frames. Later, the frame is sent to the three main parts of the Yolo algorithm, i.e. Backbone, Neck, and Head. Where Backbone uses CNN and creates the various features from the images. In the next stage, the Neck combines the multiple features and forwards the class prediction to the next stage, the Head. Head the Head uses features from the Neck and does the final prediction step [10].

Algorithm

Step 1: Divide the input image into the smaller grid-like boxes. So, the grid box will detect the objects inside it.

Step 2: Mark the bounding box inside the Object-bounding box regression.

$Y= (p_c, b_x, b_y, b_n, b_w, b_h, c)$

Where p_c is the centre point of the image, b_x, b_y is the two coordinates of the Point in the image,

b_w is the width of the bounding box, and c is the detected object class.

Step 3: Calculate the intersection over the union to measure the model's accuracy [11].

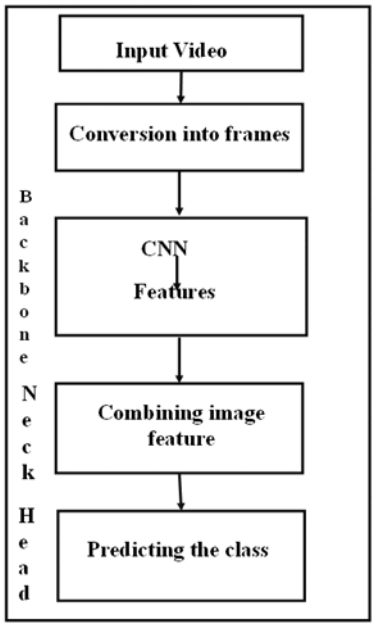


Fig 2: Working of the Proposed Model

Following YOLOv3, YOLOv5 has become the widely used one-stage object identification algorithms due to its exceptional stability, great universality, and extraordinary performance. Moreover, there are specific problems that need to be fixed. First, a considerable percentage of failed detections in tiny object detection are caused by the loss of edge information resulting from 32-fold down sampling and deep convolution in two-dimensional pictures, where smaller objects have smaller pixel proportions than larger ones. Researchers developed a unique feature extraction structure to solve this problem, one that maintains edge information as possible for small objects devoid of requiring a lot of floating-point arithmetic or parameter increases. With this enhancement, small-item identification becomes more precise. In conventional detection techniques, more convolutional layers make semantic information more readable for larger objects; however, too many convolutions cause oversaturation. Thus, for large-size objects, 32-fold down-sampling layer is most transparent conceptual layer, but shallow convolutional layers yield more transparent goal conceptions for small items. To improve the performance of small object detection even further, researchers presented the external network fusion technique. Anchor boxes are used in the object detection results to show the locations of the objects in the head section during detection tasks, along with confidence levels. The enhanced YOLOv5 simultaneously handled localization and classification tasks, which reduced both classification and localization accuracy while shortening the training inference time. On the other hand, the accuracy of small object identification may be significantly increased, and error rates can be substantially decreased with the anchor-based Decoupled Head [12,13].

The enhanced of YOLOv5 incorporates the bottleneck structure suggested in Resnet in addition to C3. It was discovered that the C3 module preserved full-size object recognition precision levels using the original YOLOv5s configuration file but reduced model by about 5.7% relative to Bottleneck-CSP. Nevertheless, the low degree of detection precision persists, yet the ability to extract features from small objects has remained the same. The Efficient Aggregation Layer Module (ELAN) was introduced, which maximizes the deep network's feature extraction capability by improving its gradient propagation efficiency. The ELAN module shows a more robust capacity for model learning and successfully addresses the problem of model convergence brought on by scale. That complicates the network topology because of its long gradient update path. Furthermore, inference ease and speed are hampered by the large number of parameters. Enhanced model has simple structure while enhancing its small object identification capabilities, drawing on insights from previous research. Identifying objects in densely populated locations or with small pixel dimensions takes time and effort. Proposed

model created and used an improved CB structure that improves feature extraction for edge information by applying several CBS structures to address object detection in various environmental condition. To avoid gradient anomalies, proposed model included a residual structure. To lower parameter values, it used bottleneck modules. Further CB structure refer to +CB in the experimental segment, finds a balance between detection precision and training efficiency [15].

4. Result Analysis and Discussion

This section evaluates the proposed model performance by looking at their recall, accuracy, precision. Jupiter and Google Colab are considered for the experiment. TensorFlow Object Detection API and Robofow are used to provide a comparison examination of various environmental condition. Tensorboard is an interactive platform created by Tensorfow to visualize the training and assessment data [16]. MATLAB 2020a is used to design the graphical representation. A system running Windows 11 with 16GB RAM and a GTX GPU is considered. Two different data formats are used for the experiment. For image, KITTI is considered, whereas for video, BDD and Indian traffic datasets with various environments are considered. The enhanced Yolov5 model is used for both image and video datasets. For analysis, the first video is converted into frames and is considered for further study. Total 2000 images are considered for the study. There are total 5 classes in the work i.e car, bus, traffic light, truck, person. For each nine different condition 67% is taken for training, 23% is taken for validating, 10% is taken for the testing.

Table 2 is showing the average object detection rate of the proposed enhanced yolov5 model in all nine-weather condition. Overall model is working for all the condition with the improved accuracy. Comparing with the previous work accuracy is improved for rainy, fog, night combined with rain and low light. Figure 3 and 4 shows the object detection in low light before detection and after detection, which resulted 68% accuracy. Figure 5 and 6 shows the multi-object detection for the Indian traffic dataset in the normal condition. Figure 7 and 8 shows the multi-object detection in bright sunlight.

Figure 9 compares the accuracy concerning the three objects, i.e., 1) Car; 2) Person and 3) Traffic light. The blue line represents the car as the Object, the orange represents the person as the Object, and the grey represents the traffic light as the Object. Figure 10 shows a single shot object, that is, a car, is taken to compare the accuracy in all the extreme condition.

we take the average of the Average Precision. Box boundaries labelled by class are predicted by object identification algorithms, as is well known [17,18]. IOU, or intersection over union, is the name given to this statistic. As such, we employ an IOU Threshold to compute the algorithms' recall and precision measures. For the Average recall calculation equation 1 is used.

$$AR = 2\int_{0.5}^{1} recall\,(IoU)do \quad (1)$$

Recall is averaged over all IOU in the range of [0.5 – 1.0] to determine it where the matching recall value is recall, and o stands for IOU. The classification loss of 'N' classes is a weighted, straightforward soft-max loss[19]. It explains how the model operates in classifying assignments for every class. Table 3 Shows the accuracy, precision and recall of the proposed model.

Table 2: Comparisons of various conditions results in terms of Detection Rate.

Dataset	*Condition*	*Result (average detection Rate)*
Kitti	*Low Lighting*	*Car:68.85*
Kitti	*Normal*	*Car:94*
		Traffic light:75
		Bus: 100
		Person:96
BDD	*Rainy*	*Car:76*
		Person :71
BDD	*Sunny*	*Car :52*
		Traffic Light :57
BDD	*Normal*	*Person: 93*
		Truck :91
		Car : 98
BDD	*Night*	*Car:71*
BDD	*Night+Rain*	*Car :71*
		Person:68
		Bus :69
BDD	*Fog*	*Car : 76*
		Traffic Light : 75
		Person :78
BDD	*High Illumination*	*Person:70*
		Car :75
BDD	*High Speed*	*Car :84*
		Traffic Light :79

Fig 3: Low light object detection before

Fig 4: Low light object detection

Fig 5: Multi-object detection in bright light Before

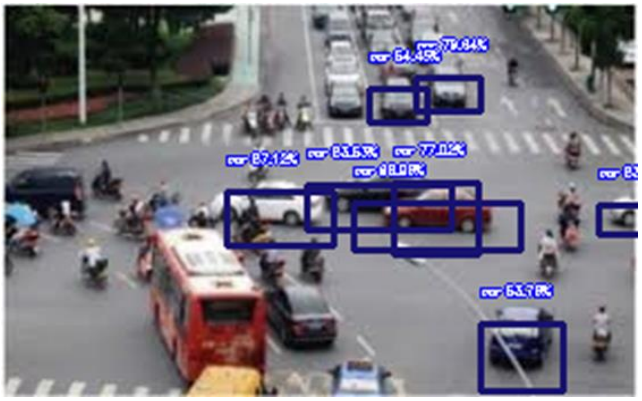


Fig 6: Multi-object detection in bright light after

Fig 7: Multi -object detection in the bright sun light before

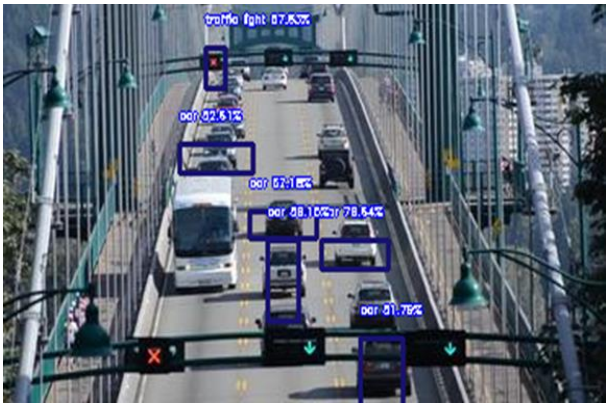
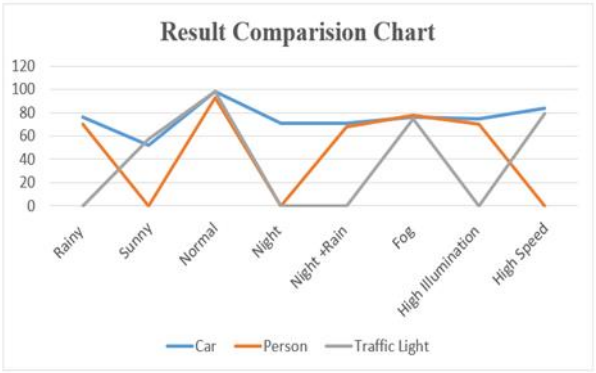


Fig 8: Multi -object detection in the bright sun light after



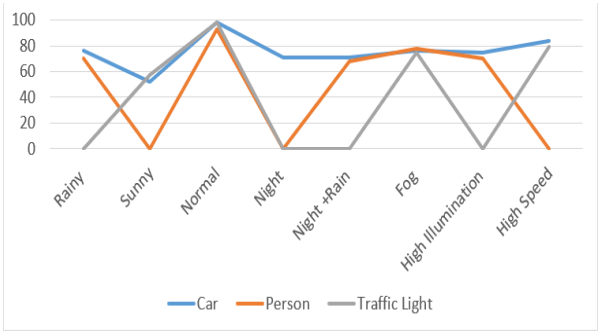


Fig 9: Shows the accuracy comparison of various conditions among three conditions

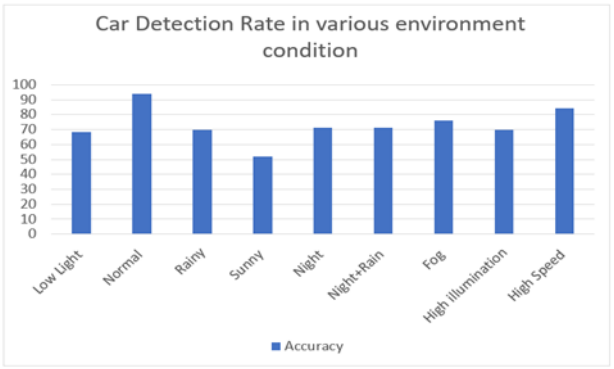


Fig 10: Comparative results on car detection rates in various environmental conditions

4.1 Comparative Analysis

Many models are developed using deep learning framework for segmentation, object detection and classification in the autonomous vehicle in the normal weather condition. Which yield excellent results in object detection. But when object need to be detected in the extreme condition like night, rainy, high illumination most of the models fails or yields very less accuracy. In the proposed model resulted in better accuracy compared to the previous work which is discussed in the related work. Table 3 shows the comparison of various conditions results in terms of accuracy, precision and recall. As the overall observation from the work night vision, low light and bright conditions require more advanced algorithms to improve the detection rate or accuracy.

Table 3: Comparisons of various conditions results in terms of accuracy, precision, recall

Condition	*Accuracy*	*Precision*	*Recall*
Normal	98.7%	0.98	0.97
Rainy	76%	0.72	0.79
Night	71%	0.69	0.72
High Speed	7 9%	0.80	0.81
High illumination	75%	0.76	0.79
Fog	76%	0.78	0.79
Low Light	68.85%	0.69	0.70

5. Conclusion

All the environmental conditions need to be considered for working with autonomous vehicles. In the proposed work, two types of datasets are considered, i.e. image (KITTI) and video (BDD). The video consists of 9 various environmental conditions. For all the conditions, the deep learning model enhanced YOLOV5 is applied for object detection and classification. For this work, the objects are traffic signs, persons, cars, and buses. From the proposed model resulted in 98.7%, 76%, 79%,76%,75%,71%, 71% ,68.5 %for normal vision, Rainy, accelerated speed, fog, high illumination, night vision, night combined with rain, low light. This resulted in low light, and night vision requires more improvement. The proposed model intends to categorize the classes accurately with improved accuracy where the model is confident in predicting the object in single shot. The training and validation metrics are visualized with better metrics and gives promising outcomes. However, the model encounters some complexity in predicting the objects in single shot which can be resolved with the integration of hybridization approach. In future work, hybrid deep learning models can be applied to light images and night vision for better object detection and classification.

References

[1] Joel Janai, Fatma Guney, Aseem Behl, Andreas Geiger," Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art", arXiv: 1704.05519v2 [cs.CV. 2019].

[2] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang and Bernt Schiele," How Far are We from Solving Pedestrian Detection?", arXiv:1602.01237v2 [cs.CV] .2016.

[3] Joel Janai, Fatma G uney, Anurag Ranjan, Michael Black, and Andreas Geiger. \Unsupervised Learning of Multi-Frame Optical Flow with Occlusions". In: Proc. of the European Conf. on Computer Vision (ECCV). 2018.

[4] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele.\Multi-person Tracking by Multicut and Deep Matching". In: Proc. Of the European Conf. on Computer Vision (ECCV) Workshops. 2016.

[5] Qasem Abu Al-Haija , Manaf Gharaibeh ,Ammar Odeh, "Detection in Adverse Weather Conditions for Autonomous Vehicles via Deep Learning", MDPI, 2022..

[6] Mazin Hnewa and Hayder Radha," Object Detection Under Rainy Conditions for Autonomous Vehicles: A Review of State-of-the-Art and Emerging Techniques", arXiv:2006.16471v4 [cs.CV] 12 Feb 2021.

[7] Kim, Y., Hwang, H., Shin, J.:Robust object detection under harsh autonomous-driving environments. IET Image Process.16, 958–971 (2022).

[8] Abhishek Balasubramaniam, Sudeep Pasricha, "Object Detection in Autonomous Vehicles: Status and Open Challenges, 2021.

[9] Top 10 Popular Datasets for Autonomous Driving Projects (analyticsindiamag.com)

[10] Gene Lewis, “Object Detection for Autonomous Vehicles”, 2015.

[11] *https://blog.roboflow.com/yolov5-improvements-and-evaluation/*.

[12] *https://www.exxactcorp.com/blog/DeepLearning/YOLOv5-PyTorch-Tutorial*

[13] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang and Bernt Schiele,” How Far are We from Solving Pedestrian Detection?”, arXiv:1602.01237v2 [cs.CV] .2016.

[14] *Joel Janai, Fatma G uney, Anurag Ranjan, Michael Black, and Andreas Geiger. \Unsupervised Learning of Multi-Frame Optical Flow with Occlusions". In: Proc. of the European Conf. on Computer Vision (ECCV). 2018.*

[15] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M. Gavrila. \The EuroCity Persons Dataset: A Novel Benchmark for Object Detection". In: IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) (2019).

[16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. \The Cityscapes Dataset for Semantic Urban Scene Understanding". In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2016.

[17] Laura Leal-Taix_e, Cristian Canton-Ferrer, and Konrad Schindler. \Learning by Tracking: Siamese CNN for Robust Target Association". In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops. 2016.

[18] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele.\Multi-person Tracking by Multicut and Deep Matching". In: Proc. Of the European Conf. on Computer Vision (ECCV) Workshops. 2016.

[19] Ahmed I, Ahmad M, Ahmad A, Jeon G (2021) IoT-based crowd monitoring system: Using SSD with transfer learning. Comput Electr Eng 93:107226.

[20] Wang Q et al (2023) Deep convolutional cross-connected kernel mapping support vector machine based on SelectDropout. Inf Sci 626:694–709.

[21] Ding L, Xu X, Cao Y, Zhai G, Yang F, Qian L (2021) Detection and tracking of infrared small target by jointly using SSD and pipeline flter. Digit Signal Process 110:102949.

[22] Yundong LI et al (2020) Multi-block SSD based on small object detection for UAV railway scene surveillance. Chin J Aeronaut 33(6):1747–1755

[23] Li X, Li Y, Shen C, Dick A, Hengel AVD (2013) Contextual hypergraph modeling for salient object detection. In: 2013 IEEE International Conference on Computer Vision, Sydney, pp. 3328–3335.

[24] Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S (2013) Salient object detection: a discriminative regional feature integration approach. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, pp 2083–209

[25] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat and Pierre Alliez. “Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark”. IEEE International Geoscience and Remote Sensing Symposium (IGARSS). 2017.

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. \SSD: Single Shot MultiBox Detector". In: Proc. of the European Conf. on Computer Vision (ECCV). 2016.

[27] Mohana, S.D., Bharathi, R.K. (2021). Face Template Security: LBP-Based LSB Watermarking Technique for Multi-class SVM Classification Using HoG. In: Komanapalli, V.L.N., Sivakumaran, N., Hampannavar, S. (eds) Advances in Automation, Signal Processing, Instrumentation, and Control. i-CASIC 2020. Lecture Notes in Electrical Engineering, vol 700. Springer, Singapore. https://doi.org/10.1007/978-981-15-8221-9_148