Electronic Letters on Computer Vision and Image Analysis 21(2):40-61, 2022

Robust Pedestrian Detection and Path Prediction using Improved YOLOv5

Kamal Hajari[*] , Ujwalla Gawande[*] and Yogesh Golhar[+]

[*] *Research Scholar, IT Department, Yeshwantrao Chavan College of Engineering, Nagpur, India*
[*] *Dean R & D and Associate Professor, IT Department, Yeshwantrao Chavan College of Engineering, Nagpur, India*
[+] *Assistant Professor, CE Department, St. Vincent Palloti College of Engineering and Technology, Nagpur, India*

Received 17th of January, 2022; accepted 14th of August 2022

Abstract

Pedestrian detection and path prediction are critical concerns in vision-based surveillance systems. Advanced computer vision applications are challenged by discrepancies in pedestrian postures, scales, backdrops, and occlusions. In the state-of-the-art approach, these challenges cannot be addressed because of limitations in postures and multiscale information missing in the features of the region of interest. In order to address these challenges, we present an improved YOLOv5-based deep learning approach. Occupied pedestrians are detected on multi-scales using the improved YOLOv5 model. We improved the YOLOv5 detection method in three ways: 1) a new feature fusion layer has been added to capture more shallow feature information of small size pedestrians, 2) features from the backbone network have been brought into the feature fusion layers to reduce feature information loss of small size pedestrians; and 3) Scale Invariant Cross-stage Partial Network (SCSP) has been added to detect the pose and scale invariant pedestrians. At last, the proposed path prediction method used to estimate the pedestrians path based on motion data. The proposed method deals with partial occlusion circumstances to reduce object occlusion-induced progression and loss, and links recognition results with motion attributes. It uses motion and directional data to estimate the movements and orientation of pedestrians. The proposed method outperforms the existing methods, according to the results of the experiments. Finally, we conclude and look into future research directions.

Key Words: CNN, Deep Learning, YOLOv5, Pedestrian detection, Tracking, Path prediction.

1 Introduction

Because of the growing concern about security, many surveillance systems are now placed in significant locations all over the world. Human resources had to monitor a lot of videos, which took a long time. As a result, one of the key shortcomings of the traditional approach is that it is difficult to detect and forecast a pedestrians path from films. It is a lengthy procedure. Because of current system constraints, an active video surveillance system that can identify and forecast pedestrian paths in real-time is required. COVID-19 viral infections are currently causing a pandemic all over the planet. In congested and densely populated places, the viruss chances

Correspondence to: <kamalhajari123@gmail.com>

Recommended for acceptance by <Angel D. Sappa>
https://doi.org/10.5565/rev/elcvia.1413
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

of spreading are boosted. However, in this case, an intelligent surveillance system is required to monitor the congested scene. The system should be able to detect pedestrians and routes in order to prevent people from congregating and spreading the COVID-19 virus. Pedestrian detection can also be used in computer vision applications such as homeland security, child monitoring, elder fall detection and monitoring, public and private banks, ATMs, and airports, among others. Deep learning architectures are employed in many computer vision systems because of their accuracy. In many complicated conditions, such as complex backdrop backgrounds, occlusion, object deformation, illumination, and scale fluctuation, the deep learning model surpasses conventional methods. The real-time identification of pedestrians is one of the most important study areas in computer vision. Despite the fact that deep learning models improve pedestrian identification accuracy, there is still potential for improvement in both human and machine perception [1]. The model's accuracy will be harmed by complex backdrops, low-resolution photos, lighting conditions, and obstructed and small objects in the distance. Most studies in this field have solely worked on color picture object detection so far. Pedestrian recognition in real time is still slower and less precise than human eyesight. The classic Viola-Jones detector [2], Deformable Part Model (DPM) [3], Histogram of Orientated Gradient (HOG) [4], and multiscale gradient histograms [5] technique is time-consuming and computationally difficult, and it necessitates manual intervention. Such strategies have proven quite popular in the recent evolution of deep learning methods, and deep CNN-based pedestrian detection algorithms have outperformed conventional methods [6]. R-CNN [7] is the first deep learning model for object detection. This method uses a selective search window to construct a region of interest for object detection based on deep learning, as implemented in all R-CNN series. Two-stage detectors, such as R-CNN [8], SPPNet [9], Fast-R-CNN [10], Faster R-CNN [11], and Mask-R-CNN [12], and single-stage detectors, such as SSD [13] and YOLO [14], are examples of deep learning techniques. As a result, these approaches are ineffective in real-time pedestrian detection. Deep learning-based pedestrian recognition systems have a slow processing speed, making them unsuitable for self-driving cars' real-time requirements. As a result, Joseph Redmon et al. [15] introduced the YOLO network, which is a single end-to-end object regression architecture, to increase detection speed and accuracy. Later, the researchers presented other variants of YOLO, such as YOLO (v5, v4, v3, v2, v1) [16][17], to increase both detection accuracy and speed when identifying smaller and densely distributed pedestrians.

In many cases, images of natural scenes usually vary in proportion and at different orientations. These artifacts make pedestrian detection and classification a challenging task. Several other problems include 1) uneven lighting; 2) blurry and hazy appearance; 3) changes in size with different posture changes, etc. Fig. 3 describes the key ideas for a new improved YOLOv5 framework, which is built on the Faster R-CNN pipeline [12]. There are three enhancements in the proposed improved YOLOv5 approach compared to the original YOLOv5 method. First, a new fusion layer is added, resulting in a large-scale feature map with a dimension of 152x152x255 pixels is extracted to detect the small size pedestrian efficiently. Second, additional fusion layer used to carry feature data from the backbone into the feature fusion layers. The SCSP modules are then added to feature fusion layers as the third step. It is useful for the detection of multiscale and occluded pedestrians. In comparison to the original YOLOv5 model in Fig. 2, the improved YOLOv5 model has four fusion layers. Fig. 3 shows the structure of the new fusion layer. A image pyramid-based approach designed in backbone network for detecting pedestrians based on the improved YOLOv5 detection method. The fourth step is to relocate the detection results into the original input image and eliminate duplicate ones. Once all the detection results are mapped to the original input image, the Non-maximum Suppression (NMS) algorithm is applied to eliminate duplicate detection results. Therefore, improved YOLOv5 can achieve excellent detection performance at a wide range of input scales. Since improved YOLOv5 shares the convolutional features of the entire image with different target proposals, it is very effective in terms of training and testing time. At last, proposed pedestrian direction prediction algorithm applied for efficient moment tracking of pedestrian in video. The experiment were preformed on the proposed academic environment dataset and other benchmark dataset. Small-scale unchanging pedestrians can be detected efficiently using the proposed improved YOLOv5 approach. The following are the contributions to this paper:

1. The proposed deep learning system detects and predicts pedestrian routes using improved YOLOv5 and motion information.

2. For pedestrian detection, we first employ an improved YOLOv5. It may be used to detect small, scale-invariant pedestrians in a scene.

3. For path prediction, we extract the pedestrian's motion characteristics and their intercommunication within a frame.

4. Finally, the routes prediction algorithm uses the motion pattern in the frame to estimate the pedestrian movement's direction.

5. When compared to existing state-of-the-art approaches, the experimental results reveal a significant performance improvement in pedestrian identification and path prediction.

The remainder of the research paper is organized in the following manner. In section 2, related work such as pedestrian recognition and path prediction is discussed, along with the methods drawbacks. In Section 3, a new proposed pedestrian dataset is discussed. In Section 4, a new improved YOLOv5 model and path prediction method is discussed. Section 5 contains experimental data as well as a comparison of the suggested strategy to other approaches. The final section concluded with a research direction for the future.

2 Related Work

Pedestrian detection has used a variety of technologies over the last few decades, many of which have already had a substantial influence. Some techniques try to improve the basic features employed [18, 19, 20], while others aim to improve the detection algorithm [21, 22], while some combine DPM [23] or exploit the scene context [23,24]. Many aspects and methodologies were evaluated by Benenson et al. [18]. Bensenson et al. proposed the fastest method for achieving a frame rate of 100 frames per second (FPS) for pedestrian identification once more in [20]. The deep enhanced pedestrian identification accuracy after 2012 [21][22]. However, the time it takes them to process each image is longer, lasting a few seconds. On CNNs, a variety of astonishing methodologies are presently used. P. Sakrapee et al. [25] proposed new features based on low-level vision features that included spatial pooling to improve translation invariance and the resilience of the pedestrian identification process. The author of [27] employs convolutional sparse coding to begin each layer and then fine-tunes it to perform object detection. The merging of Region Proposal Networks (RPN) and Boosted Forest Classifiers is used in the [28] technique. RPN generates candidate bounding boxes, high-resolution feature maps, and confidence scores, as proposed in Faster R-CNN [12]. The Real-boost technique is also used to design the Boosted Forest Classifier in order to leverage the information gained from the RPN. On the pedestrian test data set, this two-stage detector performs admirably. Murthy et al. [29] investigated pedestrian detection using a variety of custom-made deep learning approaches.

Zhenxue Chen [30] proposed a network for precise real-time localization of small-scale pedestrians that combines both region creation and prediction parameters. A scale-aware Fast-R-CNN approach for identifying pedestrians of various scales is proposed by Jianan et al. in [31], and the anchor frame method is used to numerous feature layers. In addition, Wanli et al. [32] suggested using a unified deep neural network to learn four components for pedestrian detection: feature extraction, deformation, occlusion, and classification. Yanwei et al. [33] proposed a mask-guided attention network for identifying occluded pedestrians, which modulates entire body attributes to emphasis just viewable portions while suppressing occluded regions. However, this strategy fails to deliver adequate results when dealing with obstructed pedestrians. A simple and compact technique is recommended by Shanshan Zhang et al. [34]. A channel-based attention network is added to the Faster R-CNN detector while detecting obstructed pedestrians. Tao et al. [35] suggested a unique method for detecting small-scale pedestrians that are far from the camera by combining somatic topological line localization with temporal feature aggregation. By adopting a post-processing strategy based on Markov Random Fields, this method also reduces ambiguity in occluded pedestrians (MRF). For recognising small-scale and obstructed pedestrians, Yongqiang et al. [36] developed a key point-guided super-resolution network (KGSNet). They first trained the network to generate a super-resolution pedestrian image, after which a section of the estimation module encoded the semantic information of four human body parts. Chunze et al. [37] suggested a method for detecting small-scale and obstructed pedestrians using granular perception feature learning. The attention technique is utilised to build graininess-aware feature maps, and then a zoom-in-zoom-out module is used to enhance the features. To increase the identification accuracy of small-scale pedestrians, Jialian et al. [38] proposed a novel self-mimic loss learning approach. Wei-Yen et al. [39] presented a new ratio-and-scale-aware YOLO (RSA-YOLO) that detects small pedestrians more accurately.

For dealing with occlusion other research group uses an adaptive learning method. In [51] author proposed an adaptive learning system to segment an object robustly. By using the on-line adaptation of color probabilities, method presents several specific features. It handles the illumination changes even in the outdoor area in real-time. Bayes rule and Bayesian classifier is employed to calculate the probability of an object color. In [52] author proposed a methodology for motion analysis and hand tracking based on adaptive probabilistic models. In this method a deterministic clustering framework and a particle filter together in real time. The skin color of a human hand is firstly segmented. A Bayesian classifier and an adaptive process are utilized for determining skin color probabilities. The methodology can be used to deal with luminance changes. After that, author determine the probabilities of the fingertips by using semicircle models for fitting curves to fingertips. Following this, the deterministic clustering algorithm is utilized to search for regions of interest, and at last the Sequential Monte Carlo is also used to track the fingertips efficiently. Bing et al. [40] also suggested a new small-scale sensing (SSN) network that can produce some proposal regions and is effective at recognizing small-scale pedestrians. Two-stage deep learning-based object detectors, in particular, have advantages in terms of localization accuracy and precision. Despite the fact that the method necessitates a large number of resources, its computational efficiency is low. One-stage detectors are faster than two-stage detectors due to the unified network structures, even though model precision lowers. Furthermore, with deep learning-based object detectors, the amount of training data is critical. We introduce a real-time adaptive deep neural network inspired by YOLOv5 for recognising smaller and densely distributed pedestrians. YOLOv5 [15] is a feature extraction, bounding, box extraction, object classification, and detection end-to-end single deep neural network. The YOLOv5 model was used as the base model in order to obtain greater accuracy and speed when recognizing smaller and more densely scattered pedestrians. It was chosen for the accurate detection of smaller and scattered pedestrians after improvements to the YOLOv5 network structure and hyper-parameters. The suggested method improves YOLOv5 by using the YOLOv5 deep learning framework as a basis model and adjusting hyper-parameters in real time to improve detection accuracy. Additionally, the suggested model eliminates several unneeded repetitive convolution layers at the higher end, consuming less computational time than the YOLOv5 Model. As a result, the updated YOLOv5 model is the most accurate method for detecting smaller and densely distributed pedestrians in real time. The proposed model performance is compared to that of the YOLOv5 and YOLO variants on benchmark datasets. Improved YOLOv5 is additionally examined on both INRIA [37] and Caltech [38] pedestrian datasets to test the robustness of the proposed model. Next, We describe the most relevant pedestrian datasets. In addition, we discuss the state-of-the-art approaches of pedestrian detection, along with issues and challenges.

2.1 Benchmark Pedestrian Datasets and analysis

In this section, we describe the commonly used pedestrian datasets by researchers.

2.1.1 Caltech Dataset

This dataset consists of 2,300 pedestrians along with 350,000 annotated bounding boxes. It was recorded on city roads with normal traffic during the day. The video was recorded with a resolution of 640x480 from a camera mounted on the vehicle to capture pedestrians walking on the street. Pedestrian annotation is used to verify the performance and accuracy of different pedestrian detection algorithms [38].

2.1.2 MIT Dataset

This dataset is the first pedestrian dataset; it is quite small and relatively high quality. This data set contains 709 sample pedestrian images. Whether in front view or back view, the range of pose images taken in city streets [42] is relatively limited.

2.1.3 Daimler Dataset

This dataset captures humans walking on the street through cameras installed on vehicles in an urban environment during the day. The dataset includes pedestrian tracking attributes, annotated labeled bounding boxes, ground truth images, and floating disparity map files. The training set contains 15,560 pedestrian images and 6,744 annotated images. The testing set comprises 21,790 pedestrian images with 56,492 annotated images [43].

2.1.4 ATCI Dataset

This dataset is a rear-view pedestrian database captured using a vehicle-mounted standard automotive rear-view camera for assessing rear-view pedestrian detection at different locations, such as in-house and outside parking lots, city roads, and private driveways. The data set contains 250 video clips of 76 minutes each, and 200,000 marked pedestrian bounding boxes captured in both day and night scenarios, with varying weather [44].

2.1.5 ETH Dataset

This dataset was utilized to observe traffic scenes from inside vehicles. The pedestrian behavior was captured from stereo equipment fixed on a stroller attached to a car. The dataset can be used for pedestrian detection and tracking from movable platforms in an urban scenario. The dataset includes different traffic agents such as cars and pedestrians. [40].

2.1.6 TUD-Brussels Dataset

This data set was created using a mobile platform in an urban environment. Crowded urban street behavior recorded using a camera mounted on the front side of the vehicle. It can be used in car safety scenarios in urban environments [45].

2.1.7 INRIA Dataset

This dataset is one of the popular static pedestrian detection dataset. It includes human behavior with significant changes in posture, appearance, clothing, background, lighting, contrast, etc., plus a movable camera and complex background scenes. [37].

2.1.8 PASCAL Visual Object Classes (VOC) 2017 and 2007 Dataset

This dataset consists of static objects with different views and poses in an urban environment. The aim behind the creation of this dataset was to recognize the visual object classes in real-world scenes. The 20 different categories in this dataset include animals, trees, road signs, vehicles, and pedestrians [46].

2.1.9 MS COCO Dataset

Microsoft Common Object in Context (COCO) 2018 dataset [39] created the Common Object in Context (COCO) 2018 dataset was recently used for stimulus object detection, while focusing on detecting different objects in the context. The annotations include different instances of objects related to 80 categories of objects and 91 different categories of human segmentation. There are key point annotations for pedestrian instances and five image labels per sample image. The COCO 2018 dataset challenges include (1) real-scene object detection with segmentation mask, (2) panoptic segmentation, (3) pedestrian key point evaluation, and (4) dense pose estimation in a crowded scene [39].

2.1.10 Mapillary Vistas Research Dataset

This dataset is used for real scene segmentation of street images [47]. Panoptic segmentation resolves both pedestrian and different non-living classes, consolidating the concept of semantic and instance segmentation tasks efficiently. Table 1 shows a comparative analysis of the pedestrian databases with their purposes for video surveillance. We have also included our proposed dataset, discussed later in the paper. The association is performed in terms of the use of the dataset, size of the dataset, real-world environment scenarios, type of labeling, and annotation. These details are used for validating the object detection and tracking algorithm performance.

3 Proposed Academic Environment Pedestrian Dataset

In this section, we discuss the proposed dataset development framework and its advantage over the existing state-of-the-art benchmark dataset.

3.1 Image/Video Acquisition framework

In proposed dataset the students behavior in college premises recorded using a high-quality DSLR camera from a different viewing angle. We recorded video at 30 f/s, enabled 4K recording, with a resolution of $3840x2160$, and H.264 compressed to .mp4 format. The database includes more or less 100 sample videos. The duration of each example video is 20-30 minutes. Fig. 1 shows an example frame of a video sequence available in the dataset. The camera tilt angle varies from 45° to 90°. Pedestrians are students of Yeshwantrao Chavan College of Engineering in Nagpur, aged 22-27 years old, over 90%, of which 65% are male and 35% are female, mainly of Indian ethnicity. The key features of the data acquisition settings summarized in Table 2.

The proposed pedestrian database consists of different behaviors of students in academic activities under different conditions, such as students studying in practical laboratories, exam hall scenarios, classroom, a student doing the cheating in the exam hall, a student taking answer book outside the exam hall, student stealing the mobile phone or other electronic devices such as mouse, keyboard, student stealing the lab equipment, student dispute in the college premises, student disturbing another student, student threatening another student, etc.

3.2 Pedestrian Annotation

The proposed dataset completely annotated at the frame of video, by human specialists. We provide a csv file for each video sequence, using the same file naming protocol as videofilename.csv. The labeling process is divided into three stages: 1) human detection; 2) tracking, 3) suspicious activity recognition and soft biometric features. First, the Mask R-CNN [12] method is used to provide an initial estimate of the position of each pedestrian in the scene, and the data obtained are manually verified and corrected. Next, the deep sort method [14] provided the preparatory tracking information, which was again corrected manually. As a result of these two initial steps, we obtain a rectangular bounding box representing the region of interest (ROI) for each pedestrian in each frame. The final stage of the annotation process is carried out manually, where a human expert

Table 1: Comparative analysis of state-of-the-art dataset.

Dataset Title	**Purpose**	**Source Data**	**Annotation**	**Environment**	**Year**	**Ref.**
Caltech Pedestrian dataset	Pedestrian detection and tracking	Frames:250000 (137 min.)	Bounding boxes:350,000 Pedest.:2300	City Environment	2012	[38]
MIT dataset	Pedestrian segmentation, detection,	Pedestrian:709 Train: 509 Test: 200	No annotated pedestrian	Day light scenario	2000, 2005	[42]
Daimler	Detection and Tracking of pedestrian	Pedestrian: 15,560, Negative:6744	2D bounding box and a ground truth	City Environment	2016	[43]
GM-ATCI	Pedestrian segmentation, detection, and tracking	250 video sequences	200K annotated pedestrian bounding boxes	Day and complex weather and lighting	2015	[44]
ETH	Segmentation, Detection, Tracking	Videos	Traffic agents Ex. cars and pedestrians	City Environment	2010	[40]
TUD Brussels	Detection, Tracking	1092 images	Annotated Ped.:1776	City Environment	2009	[45]
INRIA	Detection, Segmentation	498 images	Manual Annotations	City Environment	2005	[37]
PASCAL VOC 2012	Detection, Classification, Segmentation	11,530 images,20 obj., classes	27,450 ROI annotated 6929 segmentation	City Environment	2012	[46]
MS COCO 2017	Recognition, Segmentation	328,124 images, 1.5 million obj.	Segmented people obj.	City Environment	2017	[39]
KITTI	Recognition, Segmentation	328,124 images, 80 obj. cat.	Segmented people obj.	City Environment	2015	[41]
Mapillary Vistas dataset 2017	Semantic understanding street scenes	25,000 images, 152 object ,categories	Pixel accurate instance specific Pedestrian annotations	City Environment	2017	[47]

Table 2: Proposed dataset video Acquisition configuration.

Parameter of Camera	**Setting**
Camera:DSLR High resolution camera, Effective pixels:12.4M	Video frame resolution:3840x2160
Lens FOV 94 20 mm f/2.8 focus	ISO Range: 100-3200
Camera tilt angle	45° to 90°
Video recording format	.mp4

who personally knows the students of the college sets up ID information and characterizes the samples based on soft labels. Table 3 shows the detailed information of the label annotated for each pedestrian instance in the frame, as well as the ID information, the bounding box that defines the ROI, and the frame information. For each label, we also provide a list of its possible values.

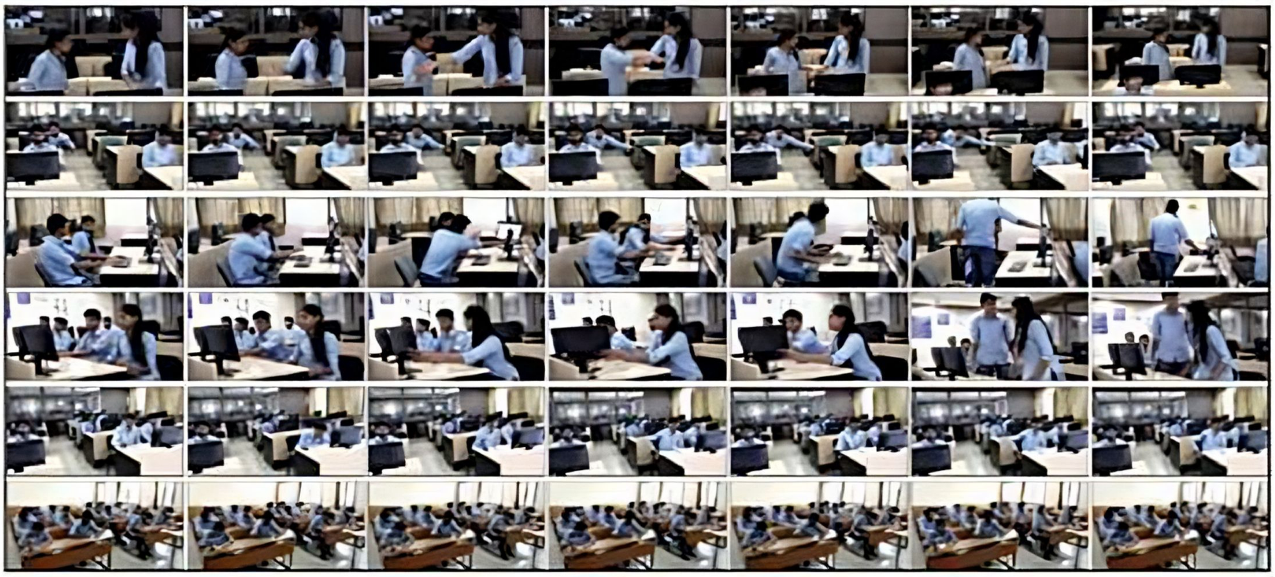


Figure 1: Sample image of the proposed database. The first row illustrates two girls dispute in the lab. The second row illustrate the scenario stealing the mobile phone in lab. The third row illustrate a scenario of a student threatening. The fourth row shows the same threatening scenario(front view). The fifth row shows the scenario of students stealing the lab equipment. The sixth row shows the scenario cheating in the exam hall.

First row illustrates two girls dispute in the lab. Second row illustrate the scenario stealing the mobile phone. Third row illustrate a scenario of a student threatening. Fourth row shows the same threatening scenario. Fifth row shows the scenario of students stealing the lab equipment. Sixth row shows scenario cheating in the exam. The proposed data set is annotated by human experts at the frame level. The labeling process is divided into three stages: 1) human detection; 2) tracking. 3) suspicious activity recognition. First, the Mask R-CNN [12] method is used to provide an initial estimate of the position of each pedestrian on the scene, and the data obtained are manually verified and corrected. Next, the deep sort method [14] provided preparatory tracking information, which was again corrected manually. Because of these two initial steps, we obtain a rectangular bounding box representing the region of interest (ROI) for each pedestrian in each frame. The final stage of the annotation process is carried out manually, where a human expert who personally knows the students of the college sets up ID information and characterizes the samples based on soft labels. The label annotated for each pedestrian instance in the frame includes pedestrian height, age, bounding box Id, feet, frame, body volume, hairstyle, hair color, head accessories, clothing, moustache beard action, and accessories. Next, the proposed improved YOLOv5 framework is detailed in depth in the next section.

4 Proposed Methodology - YOLOv5 and Improved YOLOv5

In this section, we first describe the existing YOLOv5 architecture (as shown in Fig. 2), followed by a proposed method for pedestrian detection based on the improved YOLOv5 architecture (as shown in Fig. 3). Finally, the proposed path prediction has been described. We adapt the YOLOv5 architecture for detection in order to successfully detect small-scale unchanging pedestrians in the scene. The moving pedestrian directional and motion component information is then utilized to anticipate the pedestrian's progress in subsequent frames until the pedestrian is visible in the view scope. Our method works for single individuals as well as multiple people in the scene. In YOLOv5, the compact feature corresponds to the image's greater area. It can distinguish between larger and smaller size pedestrian using the 19x19x255 and 76x76x255 features. These properties are used by YOLOv5 to detect and classify objects. In the focus module, the input image is separated into sections and then

Table 3: The proposed dataset 16 annotated attribute with other soft biometric labels.

Attributes	**Values**
Height	0→Children, 1→Short, 2→ Medium, 3→ Tall, 4→Not known.
Age	0→0-11, 1→12-17, 2→18-24, 3→25-34, 4→35-44, 5→45-54, 6→55-64, 7→greater than 65, 8→Not known.
Bounding Box	[x→Top Left; y→Top left row; h→Height; w→Width]
ID	1, 2, 3, 4, . Not known.
Feet	0→Sport, 1→Classic, 2→High Heels, 3→Boots, 4→Sandals, 5→Nothing, 6→Not known.
Frame	1, 2, 3, 4, . n.
Body Volume	0→Thin, 1→Medium, 2→Fat, 3→Not known.
Hairstyle	0→Bald, 1→Short, 2→Medium, 3→Long, 4→Horse Tail, 5→Unknown.
Hair Color	0→Black, 1→Brown, 2→White, 3→Red, 4→Gray, 5→Occluded, 6→Not known.
Head Accessories	0→Hat, 1→Scarf, 2→Neckless, 3→Occluded, 4→Not known.
Upper Body Clothing	0→T-shirt, 1→Blouse, 2→Sweater, 3→Coat, 4→Bikini. 5→Naked, 6→Dress, 7→Uniform, 8→Shirt, 9→Suit, 10→Hoodie, 11→Cardigan.
Lower Body Clothing	0→Jeans, 1→Leggins, 2→Pants, 3→Shorts, 4→Skirt, 5→Bikini , 6→Dress, 7→Uniform, 8→Suit, 9→Not known.
Moustache	0→Yes, 1→No, 2→Not known.
Beard	0→Yes, 1→No, 2→Not known.
Action	0→Walk, 1→Run, 2→Standing, 3→Sit, 4→Cycle, 5→Exercise, 6→Pet, 7→Phone, 8→Leave Bag, 9→Fall, 10→Fight, 11→Date, 12→Offend, 13→Trade.
Accessories	0→Bag, 1→Backpack, 2→Rolling, 3→Umbrella, 4→Sport, 5→Market, 6→Nothing, 7→Unknown.

combined afterwards to extract the features more effectively during down-sampling. The block of convolutions Cross-stage partial network (CSP) [15] in YOLOv5 is composed of the modules for convolution, normalization, and Leaky Rectified Linear Unit (ReLU) activation functions [39]. It was utilized in the backbone network and the neck network, respectively. The front and rear layers are connected through cross-layer connection.

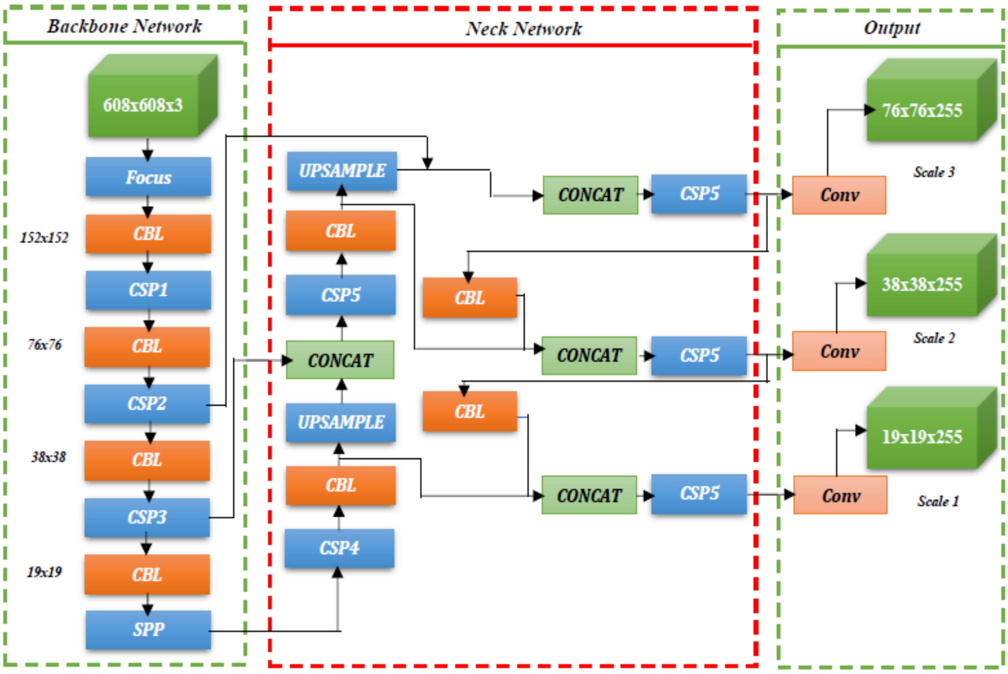


Figure 2: The architecture of the YOLOv5 model. (Backbone, Neck, and Output layers)

4.1 YOLOv5 for Pedestrian Detection

The YOLOv5 detector is a one-stage detector. The architecture of YOLOv5 is shown in Fig. 2. There are three parts to the YOLOv5 architecture: 1) A strong backbone. 2) The neck, and 3) the output. The backbone portion extracts the input picture features first. For extracting the features of the input image in varied lengths, the backbone network employs several Convolutional Neural Networks (CNN) and pooling [17]. The backbone network has four layers of feature map creation. Each layer produces a feature map with dimensions of 152x152, 76x76, 38x38, and 19x19 pixels. To gather additional contextual information and prevent information loss, the neck network combines feature maps of various levels. The Feature Pyramid Network (FPN) and Pixel Aggregation Network (PAN) feature pyramid structures are employed in the fusion process. The semantic features from the top feature maps are passed down to the lower feature maps via the FPN. Object localization features are transferred from lower feature maps to higher feature maps using the PAN structure.

We can see three feature fusion layers in the neck network, which yield three scales of new feature maps with sizes of 76x76x255, 38x38x255, and 19x19x255, where 255 is the pixel intensity range of the network. By lowering model size, the CSP network tries to enhance inference speed while retaining precision. The CSP network replaces the residual units with CBL modules in the neck. The Special Pyramid Pooling (SPP) module combines the features of the largest pooling with variable kernel sizes. The input feature map is primarily compressed. On the one hand, it reduces the size of the feature map and reduces the network's computational complexity; it compresses features and removes the most important ones. Following that, we went through the intricacies of the upgraded YOLOv5 architecture.

4.2 Improved YOLOv5 for Pedestrian Detection

To detect multiscale pedestrians, we improved the YOLOv5 detection method in three ways: 1) a new feature fusion layer with a grey background has been added to capture more shallow feature information of small pedestrians; 2) features from the backbone network have been brought into the feature fusion layers highlighted

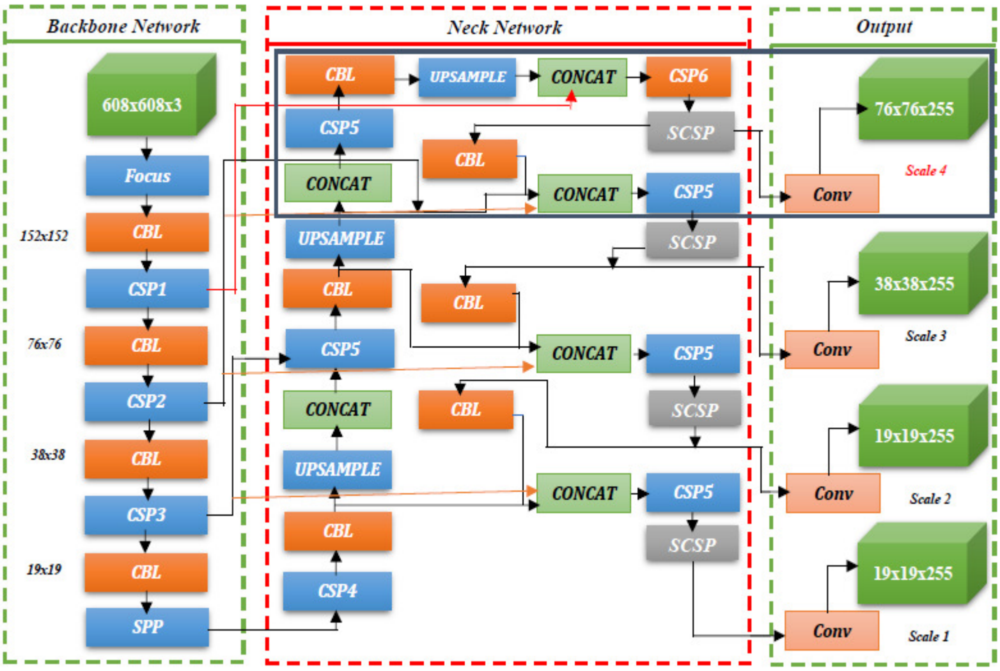


Figure 3: The architecture of the improved YOLOv5 model.

with black color to reduce feature information loss of small pedestrians; and 3) the Scale invariant Cross-stage Partial Network (SCSP) has been added to reduce feature information loss of small pedestrians. Figure 3 depicts the improved YOLOv5 architecture. There are three enhancements in the architecture of the upgraded YOLOv5 approach compared to the original YOLOv5 method. First, a new fusion layer is added, resulting in a large-scale feature map with a dimension of 152x152x255 pixels. Second, additional red-lined connections have been created to carry feature data from the backbone into the feature fusion layers. The SCSP modules are then added to feature fusion layers as the third step. To begin, we add a new fusion layer that generates a larger feature map with a size of 152x152x255 to improve the performance of YOLOv5 in recognizing little pedestrians. In comparison to the original YOLOv5 model in Figure 1, the upgraded YOLOv5 model has four fusion layers. Figure 2 shows the structure of the new fusion layer. The fused feature maps are further un-sampled and concatenated with the feature map of 152x152 pixels from the backbone network to construct an extra layer of fused feature maps as a result of the original network. The Cross Stage Partial (CSP) and Case based learning (CBL) modules are also employed in this process. CBL is used for the extraction of multiscale features of object of different size. Second, four red lines were added to bring feature information from the backbone network 152x152 pixels, 76x76 pixels, and 38x38 pixels, taking into account the large size of the images from the planets and the diverse scales of the pedestrian in the images, an image pyramid-based approach designed in this paper for detecting pedestrians based on the improved YOLOv5 detection method as shown in Fig. 4. Scale invariant Cross-stage Partial Network (SCSP) process the multiscale CNN feature paralleled to reduce the network overhead. The fourth step is to relocate the detection results into the original input image and eliminate the duplicate ones. Actually, there are two relocation steps. One is to relocate the detection results from each slice to their corresponding layer. Another is to further relocate the detection results of each layer to the original input image. Once all the detection results are mapped to the original input image, the NMS algorithm is applied to eliminate the duplicate detection results.

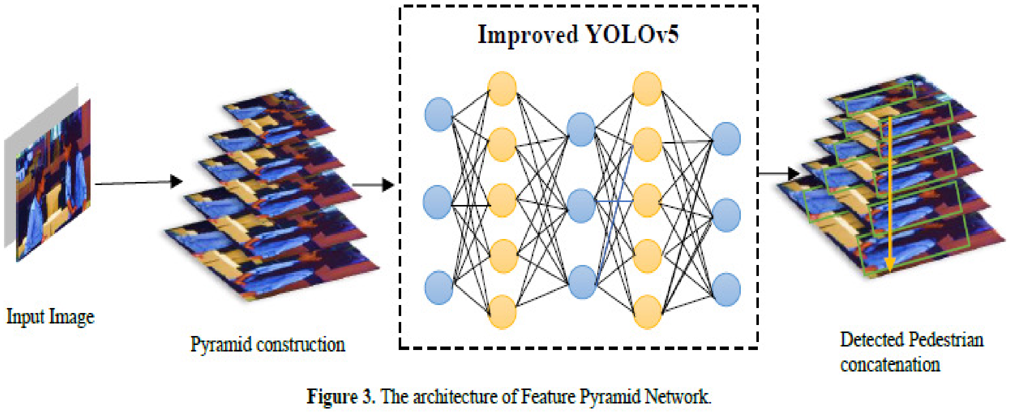

Improved YOLOv5
Input Image
Pyramid construction
Detected Pedestrian concatenation
Figure 3. The architecture of Feature Pyramid Network.

Figure 4: The architecture of Feature Pyramid Network

4.3 Path Prediction

The pedestrian's path is drawn in the frame by the proposed path prediction algorithm. The fundamental concept is to track pedestrian movement. The line is drawn until the pedestrian appears in the frame. As seen in Figure 4, the bounding box parameter is represented by the $x1$, $y1$, $x2$, $y2$ created in the detection output. Eq. 1 and 2 were used to calculate the centroid of the bounding box of each identified pedestrian:

$$cX = \frac{x1 + x2}{2} \tag{1}$$

$$cY = \frac{y1 + y2}{2} \tag{2}$$

We then draw a line from the current centroid to the previous centroid to plot the position of each pedestrian in the subsequent frame. We constructed two repositories: one for storing the pedestrian detection id and the other for storing the center position of each detected pedestrian object. On the initial frame, draw the start and end points afterwards. The following algorithm shows a step-by-step illustration of path prediction.

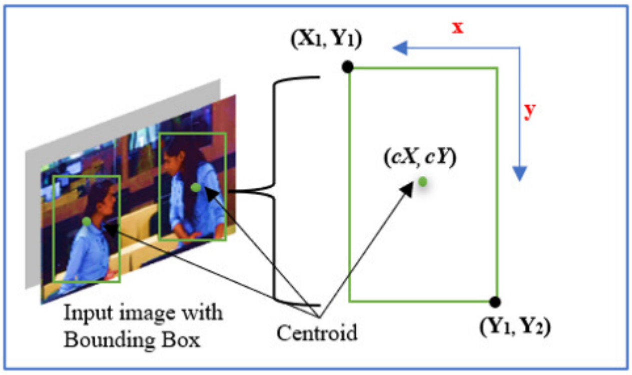


Figure 5: The Pedestrian representation with bounding box.

Algorithm 1: Pedestrian path prediction algorithm

```
Input   : $x1, x2, y1, y2$, Bonding Box parameters. $objArr$, Pedestrian detected object array.
Output  : Draw a line for path of pedestrian.

For each pedestrian in object array;
for $(objId, bboX)$ to $objArr$ do
    $bboX \leftarrow x1, x2, y1, y2$ Compute the centroid from the bounding box points. represented by:
     $(cX, cY)$
    Append centroid in centroid dictionary
    $cenDist[objId] \leftarrow (cX, cY)$
    if $objId \neq objIdList$ then
        $objIdList \leftarrow objId$
        start point
        $stPt \leftarrow (cX, cY)$
        end point
        $edPt \leftarrow (cX, cY)$
        Draw line on frame points $(stPt, edPt)$
    else
        for $pt$ to $cenDist[objId]$ do
            if $objId \neq objIdList$ then
                $stPt \leftarrow (cenDist[pt][0], cenDist[pt][1])$
                $edPt \leftarrow (cenDist[pt+1][0], cenDist[pt+1][1])$
                Draw line on frame points $(stPt, edPt)$
            end
        end
    end
    Draw rectangle on frame $x1, x2, y1, y2$
end
```

5 Experiment

On benchmark pedestrian databases such as Caltech [38], INRIA [37], MS COCO [39], ETH [40], and KITTI [41], as well as our own pedestrian database, we evaluated the performance of the proposed Improved YOLOv5 framework. The tests and proposed deep learning framework were run on a single NVIDIA GPU with a CPU Intel Core i5 3.4GHz, 16GB RAM, and a 16GB NVIDIA graphics card. The limitation of these datasets are: 1) the limited range of pedestrian poses recorded on the city streets in a controlled environment, 2) these datasets contain data with short lapses of time between successive observations of each ID in a single day, which allows to use clothing appearance features in identity matching, 3) All of these pedestrian data sets are recorded in various places such as streets and parking spaces, but they do not cover student behavior in an academic environment. This paper proposes a new dataset in an academic environment. Human experts annotated student pedestrian behavior on each frame sequence of the video, providing three types of information.

1. Pedestrian positioning with bounding box. The position of each pedestrian in the video frame is represented as a bounding box, and we can use this data for pedestrian detection, tracking, instance, and semantic segmentation.

2. Physical, behavioral, or adhered to human characteristics. Each pedestrian fully characterized by labels such as Face: eyes, eyebrows, forehead, nose, ears, mouth, facial hair, moustache, glasses, beard, hairstyle, hair color, age, body volume, gender, age, height, body accessories, ethnicity, head accessories, action and clothing data.

3. Annotated class label and ID. Each pedestrian has a unique identifier that is uniform across all video frames. These characteristics of the data set make it suitable for various recognition difficulties.

5.1 Data Pre-processing

In the proposed dataset, we classified the video information namely into three directories as Train, Test and Validation”. We describe the following three entities:

1. Annotations directory contains a XML file for each image. This file contains all the information about the image.

2. The frames are extracted from each video and divided into training and validation set in different directory.

3. The details about the frames are stored in the .txt file, it stores a unique identity number for each image.

However, for the classification purpose, we defined a improved YOLOv5 model, along with ReLU as an activation function. In the proposed approach, we used sparse categorical cross-entropy instead of categorical cross-entropy for the compilation of improved YOLOv5 model. The major advantage of using sparse classification cross-entropy is that it preserves time and computational effort, because it only uses a single value for a class label instead of the entire feature vector. We used simultaneous callbacks. A callback is a phenomenon that can execute processes at different stages of training phase. Each of these stages are describe as follows:

1. Early stopping stage: When the observed indicators stop improving the results, the training process stops.

2. Reduce Learning rate On Plateau:When the observed indicator stops improving the result, it is used to reduce the learning rate.

The detailed configuration of the the proposed improved YOLOv5 model illustrated in Table 4. We can see that most of the images classified by our model are correct. Also, as there is always room for improvement, this model can perform better with a larger dataset. The summary of layers with their required parameter while training and validation phase for the proposed improved YOLOv5. The model got a notable accuracy of 96% on the testing set and 95% of accuracy on the validation set. Next, we have computed the training loss and accuracy and validation loss accuracy. The $accuracy$ represented training accuracy. The $valLoss$ represented validation accuracy, and the $valAcc$ represented validation accuracy.

5.2 Experimental results and analysis

AP (Average Accuracy) is a commonly used index to measure the accuracy of object detectors (such as Faster R-CNN, Fast R-CNN SSD, etc.). Average precision calculates the average precision value of the recall value in the range of 0 to 1. We considered the following criteria in our experiments.

1. IF IoU $\geq$ 0.5, classify the object detection as True Positive (TP). We have calculated AP at IoU .50 and .75. represented as AP_{50}, AP_{75}. Also, at across scales small, medium, and large, represented as AP_{S}, AP_{M}, AP_{L} as shown in Table 5.

2. If IoU $<$ 0.5, then it is a wrong detection and classifies it as False Positive (FP).

3. When ground truth is present in the image and the model failed to detect the object, we classify it as False Negative (FN).

Table 4: The proposed improved YOLOv5 model summary.

Layers type	**Output Shape**	**Params**
InputLayer	(None, 224, 224, 3)	0
Conv2D	(None, 224, 224, 64)	1792
Conv2D	(None, 224, 224, 64)	36928
MaxPooling2D	(None, 112, 112, 64)	0
Conv2D	(None, 112, 112, 128)	73856
Conv2D	(None, 112, 112, 128)	147584
MaxPooling2D	(None, 56, 56, 128)	0
Conv2D	(None, 56, 56, 256)	295168
Conv2D	(None, 56, 56, 256)	590080
Conv2D	(None, 56, 56, 256)	590080
MaxPooling2D	(None, 28, 28, 256)	0
Conv2D	(None, 28, 28, 512)	2359808
Conv2D	(None, 28, 28, 512)	2359808
Conv2D	(None, 14, 14, 512)	0
MaxPooling2D	(None, 14, 14, 512)	2359808
Conv2D	(None, 14, 14, 512)	2359808
Conv2D	(None, 14, 14, 512)	2359808
Conv2D	(None, 7, 7, 512)	0
MaxPooling2D	(None, 25088)	0
Flatten	(None, 4096)	102764544
Dense	(None, 4096)	16781312
Dense	(None, 2)	8194

4. True Negative (TN) is every part of the image where we did not predict an object. This metrics is not useful for object detection, hence we ignore TN.
   We have computed the precision, recall, and mean accuracy precision (mAP) using the mathematical model as:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} \mathrm{AP_i} \tag{5}$$

We compare improved YOLOv5 to the state-of-the-art techniques in the area of instance segmentation in Table 6. Our proposed model outperform baseline variants of previous state-of-the-art models. This involves YOLOv3[16], YOLOv4[14], and other method include MNC [48] and FCIS [49], which are the winners of the COCO 2015 and 2016 challenges, respectively. improved YOLOv5 with ResNet-101-FPN backbone is better than FCIS [47]. We achieve a benchmark AP of 79%, AP_{50} of 68%, and AP_{75} of 64%. Again, we validate our system at multiple scale and obtained the results as $\mathrm{AP_S}$ of 16.5%, $\mathrm{AP_M}$ of 39.5%, and $\mathrm{AP_L}$ of 54.5%.

Table 5: Evaluation metrics used for improved YOLOv5 pedestrian detector

Metric	Annotation	Description
Avg. Precision	AP	AP at IoU=.50:.05:.95
$\mathrm{AP}_{\mathrm{IoU}=.50}$	AP_{50}	AP at IoU=.50
$\mathrm{AP}_{\mathrm{IoU}=.75}$	AP_{75}	AP at IoU=.75
$\mathrm{AP}_{\mathrm{Small}}$	AP_{S}	AP for small objects area $< 32^2$
$\mathrm{AP}_{\mathrm{Medium}}$	AP_{M}	AP for medium objects $< 32^2 < \text{area} < 96^2$
$\mathrm{AP}_{\mathrm{Large}}$	AP_{L}	Area for large objects $\text{area} > 96^2$

Table 6: Improved YOLOv5 average precision results. Comparative analysis of average precision of proposed model to the state of the art approaches available in the literature.

Methodology	Backbone	AP	AP_{50}	AP_{75}	AP_{S}	AP_{M}	AP_{L}
MNC [48]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [47]	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
Mask R-CNN [13]	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN [13]	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN [13]	ResNeXt-101-FPN	37.1	60.0	39.4	16.9	39.9	53.5
R-CNN [12]	ResNeXt-101-FPN	61.61	50.13	44.79	-	-	-
pAUCEnsT [35]	ResNeXt-101-FPN	65.26	54.49	48.60	-	-	-
FilteredICF [38]	ResNeXt-101-FPN	67.65	56.75	51.12	-	-	-
DeepParts [37]	ResNeXt-101-FPN	70.49	58.67	52.78	-	-	-
CompACT-Deep [32]	ResNeXt-101-FPN	70.69	58.74	52.71	-	-	-
Regionlets [29]	ResNeXt-101-FPN	73.14	61.15	55.21	-	-	-
3DOP [30]	ResNeXt-101-FPN	77.93	65.01	60.42	-	-	-
SAF R-CNN [31]	ResNeXt-101-FPN	77.93	65.01	60.42	-	-	-
Improved YOLOv5 without SCSP	ResNeXt-101-FPN	76	63.21	61.22	14.63	35.51	47.21
Improved YOLOv5 with SCSP	ResNeXt-101-FPN	79.31	68.42	64.1	16.5	39.5	54.5

5.3 Comparison with state-of-the-Art Pedestrian Detection methods

5.3.1 Caltech

Caltech training and testing photos were used to train the suggested system. Figure 6 depicts the outcomes (a). Techniques like TA-CNN [33], Checkerboards [34], CompACT-Deep [32], and SAF R-CNN [31] are compared to the suggested method. Improved YOLOv5 surpasses the other approaches by a significant margin, and we get the lowest miss rate of 8.32%. Using Mask R-CNN, it achieves state-of-the-art performance for object detection.

5.3.2 INRIA and ETH

The INRIA and ETH database images were also used to train and test the improved YOLOv5. Figure 6(b) and Figure 6(c) illustrate the comparison results (c). First, the suggested technique has a miss rate of 7.32% for the INRIA dataset picture, which is better than the previous method [35]. Second, the suggested model has a miss rate of 32.64% for the ETH dataset, compared to 34.98% for [41] and 37.37% for [47]. In general, the suggested strategy has a greater detection rate and a lower miss rate on the dataset.

5.3.3 KITTI

The improved YOLOv5 was also put to the test on the KITTI dataset. Figure 6 (d) shows the improved YOLOv5 pedestrian identification findings and performance comparisons with other existing techniques [48] [49] [50]. On the KITTI dataset, the proposed technique yields promising results of 76%, 64%, and 60%.

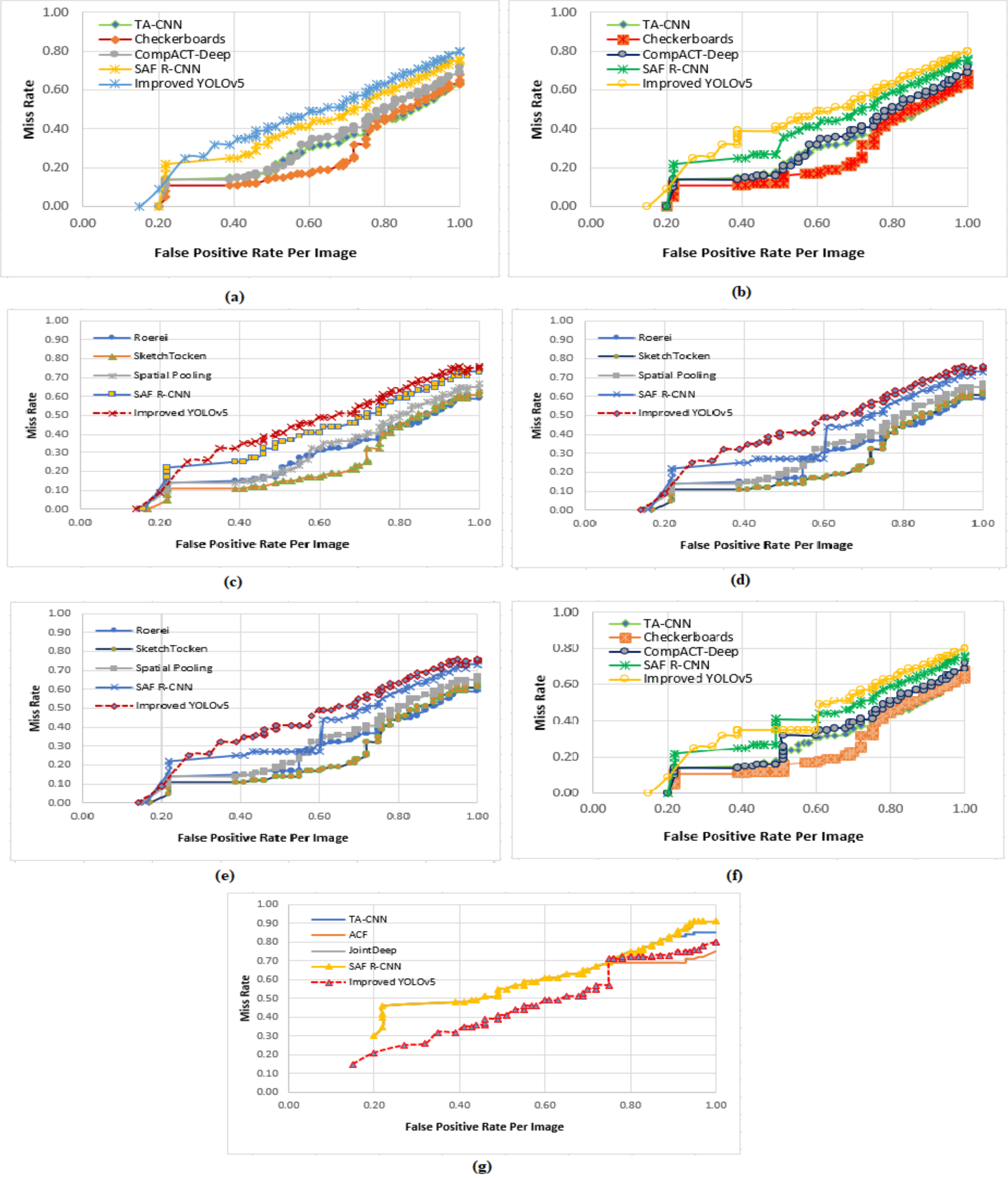
5.3.4 MS COCO and proposed pedestrian dataset

Figures 6(e) and 6(f) depict the outcomes. When compared to existing strategies [44][45][46]47], the proposed methodology outperforms them. It has an 8.57% miss rate on MS COCO and an 8.69% miss rate on the planned pedestrian dataset. The receiving operating characteristics curve was used to represent the outcome. The suggested Improved YOLOv5-based pedestrian detector is compared to current pedestrian detection techniques. Figure 6(g) shows how the proposed model compares to TA-CNN [33], Checkerboards [48], CompACT-Deep [32], and SAF R-CNN [31] in terms of recall and precision. In terms of accuracy, speed, and time required for pedestrian detection, the Improved YOLOv5 surpassed existing techniques.

6 Conclusion and Future Scope

For accurate detection and path prediction of small and distributed pedestrians, a new and enhanced YOLOv5 and an approach based on motion and direction information are proposed. While identifying smaller and more dispersed pedestrians, the proposed network structure enhances detection accuracy. The network's feature extraction capabilities were improved by removing the repetitive convolutional layer and adding a scale-invariant cross-stage partial network layer to recognize pedestrians of various sizes. In YOLOv5, we've implemented a new feature fusion layer to gather more information about little pedestrians. We additionally use the backbone network's shallow features into the feature fusion layer to further reduce feature information loss for small pedestrians. A pyramid-based approach was devised to recognise pedestrians of different scales from multiple layers of photos with different resolutions using multi-scale pedestrian examples images. According to the evaluation results, the precision of YOLOv5 has increased by 3.4%. The current object detection technology of YOLOv5 for detecting varied size pedestrians could improve the proposed methodology. On benchmark datasets such as Caltech [38], INRIA [37], MS COCO [39], ETH [40], KITTI [41], and our proposed academic environment database, the revised YOLOv5 technique produces competitive results.

On the Caltech dataset, the experimental findings indicated that the suggested technique has 1) the lowest miss rate of 8.31%. 2) the INRIA dataset has the lowest log-average miss rate of 7.31%, 3) the ETH dataset has a miss rate of 32.63%, the KITTI dataset has a pedestrian detection accuracy of 79%, and the suggested database has a miss rate of 8.68%. When compared to existing new techniques such as YOLOv4[16], Mask R-CNN [12], SAF R-CNN [31], YOLOv5[17], and SSN[40], the suggested method is superior in identifying varied sizes and varying lighted pedestrians, as well as variation in orientation. The following parts of the suggested framework can be improved in the future. Despite the fact that the model operates in real time, it still has space for improvement in terms of speed, missed detection rates on the INRIA test dataset, and missed detection of small, similar, and concealed pedestrians. The suggested model can also detect and block pedestrians by detecting human poses and trajectories of various sizes.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Figure 6: Improvements to the YOLOv5 pedestrian detector are compared to state-of-the-art approach and datasets. (a) Caltech: Lowest log-average miss rate of 8.32%, the improved YOLOv5 surpasses other approaches. (b) INRIA: Log-average miss rate of 7.32%, the improved YOLOv5 surpasses previous approaches. (c) ETH: The improved YOLOv5 Miss rate of 32.64% for the suggested model. (d) KITTI dataset: The upgraded YOLOv5 yields encouraging results of 76%, 64%, and 60%, respectively. (e) MS COCO: the Improved YOLOv5 outperforms other approaches, with an 8.57% miss rate. (f) Proposed pedestrian dataset: improved YOLOv5 outperforms other approaches, with an 8.69% miss rate. (g) A comparison of the Improved YOLOv5 and state-of-the-art approaches.

References

[1] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, Bernt Schiele, Towards reaching human performance in pedestrian detection, TPAMI, vol. 40, no. 4, pp. 973986, April 2018.

[2] Paul Viola. Michael Jones, ”Rapid object detection using a boosted cascade of simple features,” Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, pp. I-I, DOI: 10.1109/CVPR.2001.990517.

[3] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, Deva Ramanan, ”Object Detection with Discriminatively Trained Part-Based Models,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627-1645, Sept. 2010, DOI: 10.1109/TPAMI.2009.167.

[4] Navneet Dalal and Bill Triggs, ”Histograms of oriented gradients for human detection,” 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 2005, pp. 886-893 vol. 1, DOI: 10.1109/CVPR.2005.177.

[5] Najah Muhammad, Muhammad Hussain, Ghulam Muhammad, George Bebis, ”Copy-Move Forgery Detection Using Dyadic Wavelet Transform,” 2011 Eighth International Conference Computer Graphics, Imaging and Visualization, 2011, pp. 103-108, DOI: 10.1109/CGIV.2011.29.

[6] Ghulam Muhammad, M. Shamim Hossain, Neeraj Kumar, ”EEG-Based Pathology Detection for Home Health Monitoring,” in IEEE Journal on Selected Areas in Communications, vol. 39, no. 2, pp. 603-610, Feb. 2021, DOI: 10.1109/JSAC.2020.3020654.

[7] Ghulam Muhammad, Mohammed F. Alhamid, Xiaomi Long, ”Computing and Processing on the Edge: Smart Pathology Detection for Connected Healthcare,” in IEEE Network, vol. 33, no. 6, pp. 44-49, Nov.-Dec. 2019, DOI: 10.1109/MNET.001.1900045.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 580587, 2014. DOI:10.1109/CVPR.2014.81.

[9] Kaiming He; Xiangyu Zhang; Shaoqing Ren; Jian Sun, ”Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015, DOI: 10.1109/TPAMI.2015.2389824. 10. Ross Girshick, ”Fast R-CNN,” 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440-1448, DOI: 10.1109/ICCV.2015.169.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, ”Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, DOI: 10.1109/TPAMI.2016.2577031.

[11] Kaiming He, Menlo Park, Georgia Gkioxari, Piotr Dollr, Ross Girshick, ”Mask R-CNN,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 386-397, 1 Feb. 2020, DOI: 10.1109/TPAMI.2018.2844175. 13. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, SSD: Single shot multibox detector”, European Conference on Computer Vision, Cham, Springer, pp. 2137, 2016.

[12] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, ”You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788, DOI: 10.1109/CVPR.2016.91.

[13] Joseph Redmon, Ali Farhadi, YOLO9000: Better, faster, stronger, IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, pp. 65176525, Dec 2016.

[14] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection, CVPR, 23rd April 2020.

[15] Qingqing Xu, Zhiyu Zhu, Huilin Ge, Zheqing Zhang, Xu Zang, ”Effective Face Detector Based on YOLOv5 and Superresolution Reconstruction”, Computational and Mathematical Methods in Medicine, vol. 2021, Article ID 7748350, 9 pages, 2021. https://doi.org/10.1155/2021/7748350.

[16] Rodrigo Benenson, Mohamed Omran, Jan Hosang, Bernt Schiele, Ten years of pedestrian detection, what have we learned, European Conference on Computer Vision, Cham, Springer, pp. 613627, Nov 2014.

[17] Arthur Daniel Costea, Sergiu Nedevschi, ”Word Channel Based Multiscale Pedestrian Detection without Image Resizing and Using Only One Classifier,” 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2393-2400, DOI: 10.1109/CVPR.2014.307.

[18] Rodrigo Benenson, Markus Mathias, Radu Timofte, Luc Van Gool, ”Pedestrian detection at 100 frames per second,” 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2903-2910, DOI: 10.1109/CVPR.2012.6248017.

[19] Ping Luo, Yonglong Tian, Xiaogang Wang, Xiaoou Tang, ”Switchable Deep Network for Pedestrian Detection,” 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 899-906, DOI: 10.1109/CVPR.2014.120.

[20] Dollr, Piotr and Appel, Ron and Kienzle, Wolf (2012) Crosstalk Cascades for Frame-rate Pedestrian Detection. In: Computer Vision ECCV 2012. Vol.2. Springer, Berlin, pp. 645-659. ISBN 978-3-642-33708-6.

[21] Junjie Yan, Xucong Zhang, Zhen Lei, Shengcai Liao, Stan Z. Li, ”Robust Multi-resolution Pedestrian Detection in Traffic Scenes,” 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3033-3040, DOI: 10.1109/CVPR.2013.390.

[22] Wanli Ouyang, Xiaogang Wang, ”Joint Deep Learning for Pedestrian Detection,” 2013 IEEE International Conference on Computer Vision, 2013, pp. 2056-2063, doi: 10.1109/ICCV.2013.257.

[23] Sakrapee Paisitkriangkrai, Chunhua Shen, Anton van den Hengel, Strengthening the effectiveness of pedestrian detection with spatially pooled features, European Conference on Computer Vision, Cham, Springer, pp. 546561, July 2014.

[24] Xingyu Zeng, Wanli Ouyang, Xiaogang Wang, ”Multi-stage Contextual Deep Learning for Pedestrian Detection,” 2013 IEEE International Conference on Computer Vision, 2013, pp. 121-128, DOI: 10.1109/ICCV.2013.22.

[25] Christian Wojek, Stefan Walk, Bernt Schiele, ”Multi-cue onboard pedestrian detection,” 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 794-801, DOI: 10.1109/CVPR.2009.5206638.

[26] Liliang Zhang, Liang Lin, Xiaodan Liang, Kaiming He, Is faster R-CNN doing well for pedestrian detection?, European Conference on Computer Vision, Cham, Springer, pp. 443457, July 2016.

[27] Murthy, Chinthakindi B., Mohammad F. Hashmi, Neeraj D. Bokde, and Zong W. Geem. 2020. ”Investigations of Object Detection in Images/Videos Using Various Deep Learning Techniques and Embedded PlatformsA Comprehensive Review” Applied Sciences 10, no. 9: 3280. https://doi.org/10.3390/app10093280.

[28] Zhenxue Chen, Q. M. Jonathan Wu, Chengyun Liu, Real-time pedestrian detection with deep supervision in the wild, Signal Image and Video Processing, Springer, vol. 13, no. 4, pp. 761769, June 2019. DOI:10.1007/s11760-018-1406-6

[29] Jianan Li, Xiaodan Liang, Shengmei Shen, Tingfa Xu, Jiashi Feng, Shuicheng Yan, ”Scale-Aware Fast R-CNN for Pedestrian Detection,” in IEEE Transactions on Multimedia, vol. 20, no. 4, pp. 985-996, April 2018, DOI: 10.1109/TMM.2017.2759508.

[30] Wanli Ouyang, Hui Zhou, Hongsheng Li, Quanquan Li, Junjie Yan, Xiaogang Wang, ”Jointly Learning Deep Features, Deformable Parts, Occlusion and Classification for Pedestrian Detection,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 8, pp. 1874-1887, 1 Aug. 2018, DOI: 10.1109/TPAMI.2017.2738645.

[31] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, Ling Shao, Mask-guided attention network for occluded pedestrian detection, IEEE International Conference on Computer Vision, Seoul, Korea, pp. 49664974, October 2019.

[32] Shanshan Zhang, Jian Yang, Bernt Schiele, ”Occluded Pedestrian Detection Through Guided Attention in CNNs,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6995-7003, DOI: 10.1109/CVPR.2018.00731. 35. Tao Song, Leiyu Sun, Di Xie, Haiming Sun, Shiliang Pu, Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation, ECCV 2018 Computer Vision and Pattern Recognition, July 2018.

[33] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Shibiao Xu, Bernard Ghanem, ”KGSNet: Key-Point-Guided Super-Resolution Network for Pedestrian Detection in the Wild,” in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 5, pp. 2251-2265, May 2021, DOI: 10.1109/TNNLS.2020.3004819.

[34] Chunze Lin, Jiwen Lu, Gang Wang, Jie Zhou, ”Graininess-Aware Deep Feature Learning for Robust Pedestrian Detection,” in IEEE Transactions on Image Processing, vol. 29, pp. 3820-3834, 2020, DOI: 10.1109/TIP.2020.2966371. 38. Jialian Wu, Chunluan Zhou, Qian Zhang, Ming Yang, Junsong Yuan, Self-mimic learning for small-scale pedestrian detection, MM ’20: Proceedings of the 28th ACM International Conference on Multimedia, pp. 20122020, October 2020. https://doi.org/10.1145/3394171.3413634

[35] Wei-Yen Hsu, Wen-Yen 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.

[36] Bing Han, Yunhao Wang, Zheng Yang, Xinbo Gao, ”Small-Scale Pedestrian Detection Based on Deep Neural Network,” in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 7, pp. 3046-3055, July 2020, DOI: 10.1109/TITS.2019.2923752.

[37] INRIA dataset, Available at: http://pascal.inrialpes.fr/data/human/.

[38] Caltech dataset, Computer Vision Lab, http://www.vision.caltech.edu/.

[39] MS COCO (2018) dataset, Microsoft, https://cocodataset.org/download.

[40] ETH dataset, http://www.vision.ee.ethz.ch/‘aess/dataset/.

[41] KITTI dataset, http://www.cvlibs.net/datasets/kitti/.

[42] MIT Dataset. http://cbcl.mit.edu/software-datasets/PedestrianData.html

[43] ATCI dataset, Computer Vision Lab, https://sites.google.com/site/rearviewpeds1/

[44] NICTA dataset, https://www.nicta.com.au/category/research/computervision/tools/ automap-datasets/

[45] TUD-Brussels dataset, Computer Vision Lab, https://www.mpiinf.mpg.de/departments/ computer-vision-and-machinelearning/publications

[46] PASCAL dataset, CV Lab, http://pascallin.ecs.soton.ac.uk/challenges/VOC/ databases.html/

[47] Nicolai W., A. Bewley, Dietrich P., Simple online and real-time tracking with a deep association metric, IEEE ICIP, pp. 36453649, March 2017.

[48] Yonglong Tian, Ping Luo, Xiaogang Wang, Xiaoou Tang, ”Pedestrian detection aided by deep learning semantic tasks,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5079-5087, DOI: 10.1109/CVPR.2015.7299143.

[49] Shanshan Zhang, Rodrigo Benenson, Bernt Schiele, ”Filtered channel features for pedestrian detection,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1751-1760, DOI: 10.1109/CVPR.2015.7298784.

[50] Mohammad Saberian Zhaowei Cai and Nuno Vasconcelos, ”Learning complexity-aware cascades for deep pedestrian detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence PP(99):1-1, September 2020. DOI:10.1109/TPAMI.2019.2910514.

[51] Kerdvibulvech, Chutisant. (2011). Real-time Adaptive Learning System using Object Color Probability for Virtual Reality Applications. SIMULTECH 2011 - Proceedings of 1st International Conference on Simulation and Modeling Methodologies, Technologies and Applications, Noordwijkerhout, pp. 200-204, The Netherlands, 29 - 31 July, 2011.

[52] Chutisant Kerdvibulvech, A methodology for hand and finger motion analysis using adaptive probabilistic models. EURASIP J. Embed. Syst. 2014: 18 (2014), and Hybrid model of human hand motion for cybernetics application. SMC 2014: 2367-2372.DOI: 10.1186/s13639-014-0018-7