PRB-FPN+: Video Analytics for Enforcing Motorcycle Helmet Laws

Bor-Shiun Wang[1]† Ping-Yang Chen[1]† Yi-Kuan Hsieh[2] Jun-Wei Hsieh[2]
Ming-Ching Chang[3] JiaXin He[3] Shin-You Teng[2] HaoYuan Yue[4]
Yu-Chee Tseng[12]

[1]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan
[2]College of AI and Green Energy, National Yang Ming Chiao Tung University, Tainan 71150, Taiwan
[3]Department of Computer Science, University at Albany, State University of New York, NY 12222 USA
[4]Department of Computer Science, Vanderbilt University, USA
[5]Department of Computer Science, The Chinese University of Hong Kong, China

eddiewang.cs10@nycu.edu.tw, pingyang.cs08@nycu.edu.tw, khjhsnaughty.ai10@nycu.edu.tw,
jwhsieh@nycu.edu.tw, mchang2@albany.edu, jiaxin.he@vanderbilt.edu,
tengyoyo.ai10@nycu.edu.tw, 1155157271@link.cuhk.edu.hk

† Equal contribution in this paper.

Abstract

We present a video analytic system for enforcing motorcycle helmet regulation as a participation to the AI City Challenge 2023 [18] Track 5 contest. The advert of powerful object detectors enables real-time localization of the road users and even the ability to determine if a motorcyclist or a rider is wearing a helmet. Ensuring road safety is important, as the helmets can effectively provide protection against severe injuries and fatalities. However, monitoring and enforcing helmet compliance is challenging, given the large number of motorcyclists and limited visual input such as occlusions. To address these challenges, we propose a novel two-step approach. First, we introduce the PRB-FPN+, a state-of-the-art detector that excels in object localization. We also explore the benefits of deep supervision by incorporating auxiliary heads within the network, leading to enhanced performance of our deep learning architectures. Second, we utilize an advanced tracker named SMILEtrack to associate and refine the target tracklets. Comprehensive experimental results demonstrate that the PRB-FPN+ outperforms the state-of-the-art detectors on MS-COCO. Our system achieved a remarkable rank of 8 on the AI City Challenge 2023 [18] Track 5 Public Leaderboard. Code implementation is available at: https://github.com/NYCU-AICVLab/AICITY_2023_Track5.

1. Introduction

Road safety is a critical priority for governments, traffic authorities, and citizens across the globe. In several countries, particularly in Asia, where motorcycles are the primary mode of transportation, the enforcement of helmet usage for motorcyclists and riders is a critical regulation. Helmets play a crucial role in reducing the risk of severe injuries or fatalities in the event of accidents. However, the enforcement of helmet policies can be challenging due to the large number of motorcyclists and limited resources available to enforcement personnel.

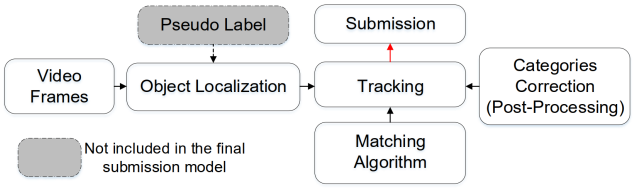


Figure 1. The overview diagram of the proposed method for the robust helmet detection and tracking for motorcyclists and riders.

With the advance of object detection in AI and computer vision, it is now possible to effectively monitor and identify motorcyclists and their helmet usage in real-time. The state-of-the-art detectors [3, 5, 8, 25] are capable of locating targets within a scene and recognizing whether or not the helmets are worn by the motorcyclists or riders. This video analytic technology offers a smart and non-intrusive solution to address the challenges faced in the monitoring and enforcing of helmet rules. The developed system can be deployed to run on energy-efficient edge devices, where traffic authorities and enforcement personnel can better monitor the compliance of helmet regulations. Such technology can ultimately contribute to road safety improvement and smart transportation.

(a)  (b) 

Figure 2. Comparisons of object detection between YOLOv7-E6E [25] and our PRB-FPN+. (a) shows the detection results obtained using YOLOv7-E6E [25], which were unable to accurately detect a motor in a foggy scene. (b) shows the detection results obtained using our proposed PRB-FPN+, which successfully recognized the objects in the foggy scene. These results demonstrate the superior performance and robustness of our approach compared to YOLOv7-E6E [25] in challenging environments such as fog.

Despite the advantages of video analytic systems, how best to visually recognize whether a person is wearing a helmet or not from the real-time video streams is challenging in the complex street environment. Typical challenges including occlusions, motion blur, diverse appearances, complex background, and environmental factors can greatly degrade the image-based helmet detection performance. Furthermore, discerning the subtle differences between a helmeted and non-helmeted motorcyclist, or identifying the specific positions of passengers on a moving motorcycle, requires a high level of precision and robustness from the detection model. Robust estimation can be achieved by associating information across frames through association of detection into tracklets, and perform Multiple Object Tracking (MOT) and information fusion. However, a robust tracker is required for such spatial-temporal analysis.

In this paper, we develop a video analytic approach to aid the enforcement of the helmet policy for motorcyclists and riders. Figure 1 overviews our helmet and motorcycle rider detection and tracking pipeline. We propose a new object detector, the **PRB-FPN+** that is an extension of our prior work of PRB-FPN [5], to accurately detect and recognize both the helmet and the motorcyclists and riders on a per-frame basis. Compared to the original PRB-FPN, the improved version of PRB-FPN+ can effectively fuse both the auxiliary and lead heads in parallel for fast and accurate one-shot object detection. Figure 2 shows a visual example comparing the popular YOLOv7-E6E detector and the PRN-FPN+, which can better localizing objects, especially for tiny and heavily occluded targets. We next incorporate an effective strategy is to fuse information across frames using a **SiMIlarity LEarning based tracker** (SMILEtrack) [27], to estimate the trajectory of each individual and determine their position on the vehicle. This two-step fusion approach can better deal with challenges associated with complex real-world scenes and occlusions, in capturing the subtle difference between the individuals with and without wearing helmets.

In summary, contributions of this work include:

- We propose a new PRB-FPN+ object detector that can recognize tiny objects such as small helmets in challenging environments such as foggy scenes or targets under heavy occlusions or clutters. Results also show great generalization ability on various object sizes and types. The comprehensive experiments on multiple tasks have demonstrated that our method outperforms state-of-the-art detectors on MS COCO dataset [16].
- We incorporate the SMILEtrack tracker to perform trajectory association and multiple target tracking. The combination of target detection and tracking enables spatial-temporal analysis for the motorcyclist helmet detection.
- Our system achieves the rank of 8 on the AI City Challenge 2023 [18] Track 5 Public Leaderboard [1].

2. Related Works

2.1. Motorcyclist helmet detection

Real-time helmet detection systems. Real-time helmet detection is important for ensuring the safety of motorcyclists on the road. Dahiya et al. [7] introduced a real-time method for detecting helmetless motorcycle drivers in surveillance videos. They employed background subtraction and object segmentation to identify drivers, and an SVM binary classifier using visual features to determine helmet usage. Jia et al. [14] proposed a deep learning-based automatic helmet detection method for motorcyclists using an improved YOLOv5 detector with triplet attention and soft-NMS for both motorcycle and helmet detection.

Accurate helmet detection systems. Although real-time detection is important, accurately recognizing the presence of a helmet, as well as the position of the passenger, are also crucial factors for ensuring the safety of motorcyclists on the road. Chairat et al. [4] developed an automated helmet violation detection system using YOLO,

Kristan's tracking method, and GoogleNet for classification. They processed multiple cameras and utilized 960 violation and 931 non-violation class images. Singh et al. [22] presented a framework for detecting helmetless motorcyclists using a detector, person localization, and head/helmet classifier. Goyal et al. [13] developed a method to detect, track, and count motorcycle riding violations in dashcam videos. They employed a curriculum learning-based detector for challenging situations like occlusions and used a trapezium-shaped boundary representation for robustness and rider-motorcycle associations. Additionally, they integrated an amodal regressor to create bounding boxes for occluded riders. Wang et al. [26] proposed a safety helmet detection method using YOLOv5-CBAM-DCN, incorporating attention mechanisms and deformable convolutions. This method tackles the accuracy issues in traditional target algorithms caused by complex site environments, uneven lighting, and irregular target shapes. Chen et al. [6] utilized YOLOv5 object detector, an attention module, a super-resolution reconstruction network, and a classifier to address the problem of helmet detection for riders.

However, the above-mentioned methods are unable to simultaneously recognize the presence of a helmet and detect the position of the passenger.

2.2. Object Detection

Object detection [17,19,21] is a key computer vision task that localizes and classifies objects in images. Detection methods are mainly one-stage [19], such as YOLO [19] and SSD [17], or two-stage detectors [21], like R-CNN [12] and Faster R-CNN [21]. One-stage methods prioritize speed while maintaining reasonable accuracy, making them suitable for real-time applications with resource or latency constraints. In contrast, two-stage methods focus on high accuracy through separate region proposal and object classification stages, catering to applications that prioritize precision over computational complexity and inference speed.

The recent advancements in one-stage object detectors [3, 5, 8, 20, 25] have paved the way for real-time object detection, offering a promising solution for monitoring helmet usage by accurately identifying motorcyclists and their adherence to helmet rules.

2.3. Multiple Object Tracking

The Tracking-By-Detection (TBD) approach comprises two primary stages: detection and tracking. In the detection stage, the system locates objects of interest within individual video frames. Following this, the tracking stage utilizes data association techniques to connect detected objects to existing tracks or to establish new tracks when required.

Detection Models. Adapted for multi-object tracking (MOT) applications, popular YOLO object detection models [3, 20, 25] excel in real-time processing and accurately detecting objects in cluttered scenes. However, the anchor-based detector's case-specific hyperparameter adjustment is challenging, and the Intersection Over Union (IOU) calculation during training is time-consuming and memory-intensive. To address these issues, anchor-free detectors [8, 15, 33] offer an alternative. YOLOX [8] transitions the YOLO series [3, 19, 20] from an anchor-based to an anchor-free detector and employs decoupled heads to enhance detection accuracy. While existing methods struggle to detect both large and small objects, our detection process utilizes the PRB-FPN approach [5] to tackle this challenge.

Data Association Methods. Data association in multi-object tracking (MOT) systems is often complicated by numerous challenges, including object occlusion, crowded scenes, and motion blur. Several methods have been proposed to address these limitations, such as SORT [2], which uses the Kalman filter for predicting object locations, and Deep SORT [29], which employs a pre-trained CNN model to extract appearance features. JDE [28] combines the Detector and Embedding models for real-time processing and high accuracy. FairMOT [32] enhances performance by utilizing an anchor-free method built on top of CenterNet [9]. Despite these improvements, JDE still struggles with feature conflicts.

We aim to use a state-of-the-art tracker [27] to post-process the tracking of drivers and passengers, which can improve the accuracy and reliability of monitoring and enforcing helmet rules for motorcyclists in complex real-world scenarios.

3. Methods

3.1. PRB-FPN+

The newly proposed **Parallel Residual Bi-fusion Feature Pyramid Network Plus (PRB-FPN+)** object detector is an extension of the original PRB-FPN architecture [5], with modifications made to the P5 model to incorporate the P6 model. The design of the PRB-FPN+ architecture includes two main features: (1) model scaling to adapt to large input images [23, 24], and (2) the parallel use of both auxiliary and lead heads, which enables efficient feature capture for identifying and localizing objects of varying sizes without compromising efficiency.

Model Scaling. Inspired from the model scaling methods [23, 24] and the original PRB-FPN [5] from our previous work, we propose the PRB-FPN+ as a newly renovated object detector. PRN-FPN+ outperforms the SoTA object detection approaches, includint YOLOv7 [25] and YOLOX [8].

Let P_i denote the features obtained from the backbone, BF_j denote the j-th BiFusion module [5], and $CORE_k^j$ and BFM_k^j represent the Pyramidal Layer within the j-th BiFusion module. The input configuration for bottom-up feature

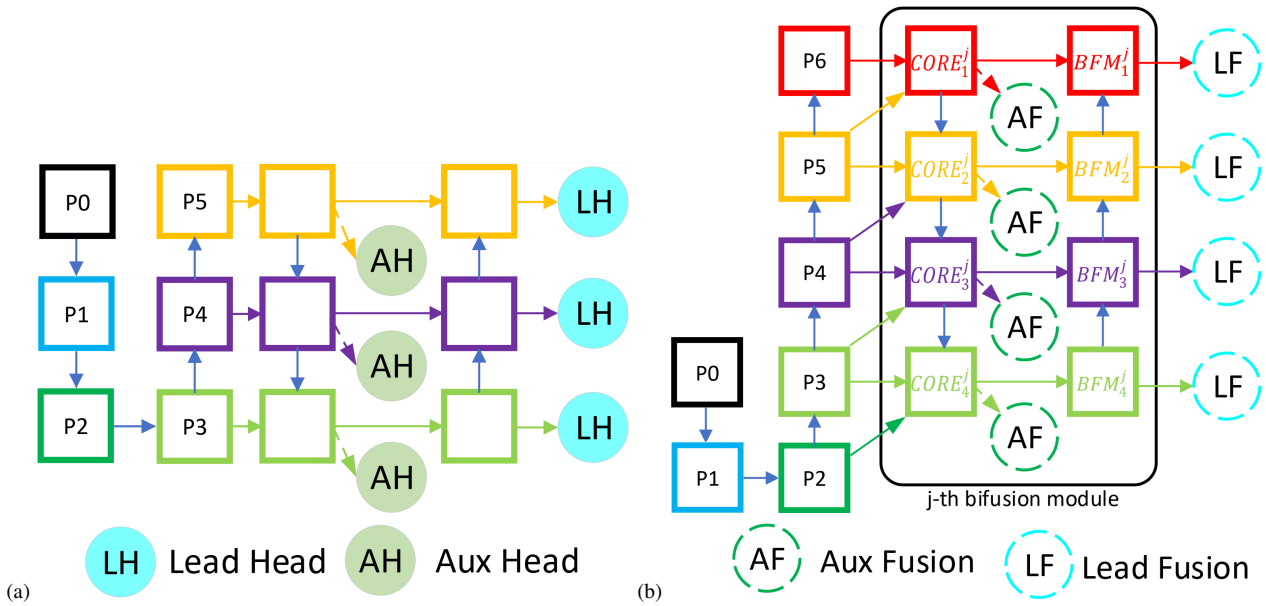


Figure 3. Coarse for auxiliary and fine for lead head fusion. (a) Model with auxiliary and lead head. (b) PRB-FPN+ with parallel auxiliary and lead fusion head.

fusion can be written as:

$$\mathrm{CORE}_k^j = \{\mathrm{P}_{7-k}, \mathrm{P}_{6-k}, \mathrm{CORE}_{k-1}^j\}, \quad (1)$$

where j = 1, 2, 3 and k = 1, 2, 3, 4, respectively. The input configuration for top-down feature fusion can be written as:

$$\mathrm{BFM}_k^j = \{\mathrm{CORE}_k^j, \mathrm{BFM}_{k+1}^j\}, \quad (2)$$

where j = 1, 2, 3 and k = 1, 2, 3, 4, respectively. By employing this hierarchical approach, our method effectively combines multi-scale features and leverages the advantages of each BiFusion module, leading to improved performance on object detection and recognition tasks.

Coarse for auxiliary and fine for lead loss. Deep learning networks have revolutionized many computer vision tasks, but how best to effectively train them still remains challenging. To address this issue, **deep supervision** has emerged as a popular technique for guiding the training of deep networks. We adopt the approach in YOLOv7 [25], to incorporate the lead head and auxiliary head to the network. Specifically, we leverage the lead head prediction as guidance to generate coarse-to-fine hierarchical labels, which are used for the training of the auxiliary head and lead head.

In addition, we propose a novel approach that introduces parallelization to these heads, enabling us to more effectively capture the necessary features for object detection tasks by better focusing on regions of interest. This parallelization design improves the feature representation, by efficiently capturing features to identify and localize objects of varying sizes without compromising efficiency. Figure 3 (a) illustrates the variant of the FPN object detector architecture, and Figure 3 (b) shows the version with the parallel multi-scale features fusion incorporated.

Following the parallelization design, the output from each BFM [5] module is concatenated to form a lead fusion before lead head for each level k as:

$$\text{Lead Fusion}_k = \text{cat}\left(\mathrm{BFM}_k^1, \mathrm{BFM}_k^2, \mathrm{BFM}_k^3\right), \quad (3)$$

where k = 1, 2, 3, 4. Moreover, the output from each CORE [5] module is concatenated to form an auxiliary fusion before auxiliary head for each level k as:

$$\text{Aux Fusion}_k = \text{cat}\left(\mathrm{CORE}_k^1, \mathrm{CORE}_k^2, \mathrm{CORE}_k^3\right), \quad (4)$$

where k = 1, 2, 3, 4. Our newly introduced designs offer significant improvements to the performance of our proposed model, by enabling it to learn from more informative signals during training. A key innovation in our design lies in the incorporation of parallelization into both the lead and auxiliary heads. By implementing this parallel design, we can more efficiently represent features and capture the essential information required for identifying and localizing objects of varying sizes, without sacrificing efficiency. These advancements have greatly enhanced the detection capabilities of our model, providing it with a powerful tool for accurately identifying and localizing objects in a variety of contexts.

3.2. SMILEtrack

In this section, we utilize the **SiMIlarity LEarning based tracker (SMILEtrack)** [27], a cutting-edge tracker that fuses a detector with a Similarity Learning Module

(SLM) to tackle the challenges in Multiple Object Tracking (MOT). This tracker presents three essential contributions to a new state-of-the-art (SoTA) MOT system: an efficient object detector, a lightweight self-attention mechanism, and a robust tracker. The "PRB-Net" [5] serves as our object detector, adept at localizing both large and small objects. Moreover, drawing inspiration from the model scaling method [23, 24], we propose PRB-FPN+, which outperforms the aforementioned SoTA approaches [8, 25].

3.3. Overall System

This section presents our overall system and its rationale. The system's workflow is illustrated in Figure 1.

The first step of our system is a detection model (Section 3.1), which forms the basis of the overall system by obtaining bounding boxes and preliminary classification results from videos. To improve performance, we use the pseudo-label technique to reduce model confusion. Given the labeling rule, objects with pixels below 40 are not labeled, yet our model can detect smaller objects. This can cause confusion, classifying similar features into different categories. To address this, we create a pseudo-label by combining the ground truth label and our model's predictions. Although pseudo-labels improve training set accuracy, test set scores do not follow suit. In our final submission, we discard this technique and use a model trained with given labels.

The next step is a tracking model designed to stabilize classification results from the detection model (Section 3.2). We observed that the detection model correctly predicts larger objects but may fail with occluded or small objects, causing inconsistent object classifications. To tackle this, we use SMILEtrack [27] that matches objects across frames. The tracking model references different frames and corrects the current result, even for occluded or distant objects. However, problems may arise when objects are misclassified initially.

In this step, we aim to rectify such errors. For example, when Passenger 1 is absent, there should be no bounding box for Passenger 2. The tracking model can reference multiple frames to correct these inconsistencies and enhance the overall performance of our system.

4. Experimental Results

Our evaluation consists of two parts. Firstly, we compare the performance of PRB-FPN+ with state-of-the-art object detectors on the MS COCO dataset [16]. Secondly, we leverage the proposed PRB-FPN+ in combination with SMILEtrack [27] to develop a video analytics system for enforcing motorcycle helmet laws in the AI City Challenge 2023 [18] Track 5 [1] contest.

4.1. Dataset and Settings

This comprehensive dataset enables the development and evaluation of object detection models focusing on the complex task of helmet rule compliance detection.

Datasets. The training dataset consists of 100 videos, each 20 seconds long, recorded at 10 frames per second with a resolution of 1920x1080. These videos feature motorcycles and their riders, who may or may not be wearing helmets. Ground truth bounding boxes are provided for motorcycles and their riders, with up to three riders per motorcycle.

Each annotated frame includes bounding box annotations, and the dataset comprises 7 different classes representing motorcycles, drivers, and passengers with or without helmets. The challenge lies in the fact that the models need to accurately recognize whether a person is wearing a helmet or not.

The test dataset for this track comprises 100 videos, each with a duration of 20 seconds and recorded at 10 fps. The participating teams' objective is to identify motorcycles and motorcycle riders while discerning if they are wearing helmets. Similar to the training dataset, it is crucial to distinguish each rider (*i.e.*, driver, passenger 1, and passenger 2) on a motorcycle and determine their respective helmet information.

Metrics. The evaluation metric used for the object detection tasks is the mean Average Precision (mAP), which was defined in PASCAL VOC 2012 [10]. The mAP is calculated by averaging the Average Precision (AP) values for each class. AP for a specific class is derived from the Precision-Recall curve, which is generated by varying the detection confidence threshold. Let True Positive TP represent the number of correctly detected objects of the class, False Positive FP denote the number of incorrect detections, and False Negative FN indicate the number of objects of the class that were not detected. The Precision (P) and recall (R) are defined as $P = \frac{TP}{TP+FP}$ and $R = \frac{TP}{TP+FN}$. To compute the AP, the area under the Precision-Recall curve is calculated, typically using the 11-point interpolation method or the integration of the interpolated curve. The final mAP score represents the mean AP across all object classes, providing an overall assessment of the object detection model's performance.

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

where N is the number of object classes, and AP_i is the average precision for the i-th class.

Labeling criteria and bounding box constraints.

To ensure consistent labeling standards, two key requirements have been established. First, objects must have at least 40% visibility to be considered. Second, the minimum height and width of the bounding boxes are set at 40 pixels.

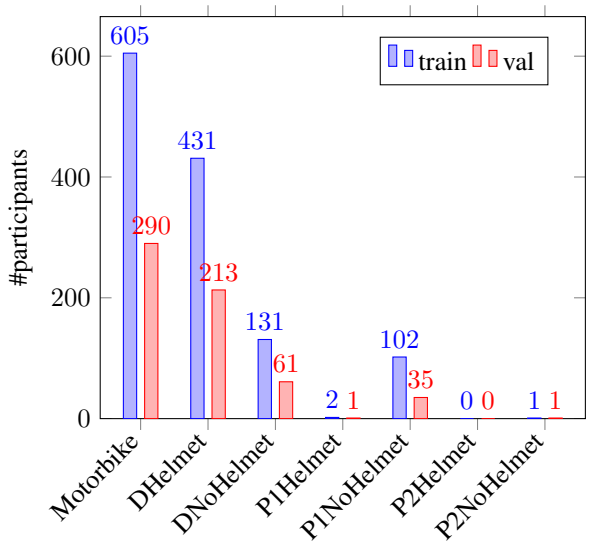


Figure 4. The object counting of the separated dataset.

Objects smaller than 40 pixels will not be taken into account when determining test accuracy results. Objects that overlap with redacted areas (blurred regions) will also be disregarded, as these obscured regions can conceal vital object features. Consequently, any objects overlapping with redacted areas in the test dataset will not influence test accuracy.

4.2. Implementation Details

Dataset distribution. Before the experiment, to make sure the training and validation dataset have a similar distribution of object categories we analyzed the dataset of 100 videos. The dataset has 895 motorbikes, 644 DHelmets, 192 DNoHelmets, 3 P1Helemts, 137 P1NoHelmets, 0 P2Helmet, and 2 P2NoHelmets. We split the dataset into 1-70 videos for training and 71-100 for validation and this way there is a similar ratio of classes in the training and validation dataset respectively. The distribution of the splitting is shown in Figure 4.

Detector training. The training process is split into two parts. The first part trains the model with the given dataset. The second part trains the model with the pseudo label. We will discuss the pseudo label later.

Both parts of the model are trained with the same setting but using different labels. We train our model on 4 NVIDIA 3090 GPUs with 16 batch sizes. To get better performance from batch normalization, we adopt synchronized batch normalization to avoid degradation by the small number of batches on a single GPU. The image is resized to 1280×1280 and adopts a powerful data transformation technique Mosaic. The stochastic gradient descent (SGD) is adopted as the optimizer with a learning rate starting from 0.01 and decreasing by one cycle with a cosine function. The overall epoch is set to 300 with 3 warmup epochs.

Tracker training. The tracker's adaptability and robustness are essential for accurately detecting helmet violations. Its performance depends on configurable parameters that can be optimized for different scenarios.

The tracking confidence threshold (0.3) determines if a detected object is reliable for tracking. A higher value may reduce false positives but increase false negatives. The lowest detection threshold (0.05) filters low-confidence detections, balancing false positives and negatives. The new track threshold (0.4) affects track initialization and can improve tracking consistency.

The track buffer parameter (30) retains frames for lost tracks, enhancing tracking performance when objects are occluded or undetected. The matching threshold (0.7) governs detection and track association, improving tracking accuracy and reducing identity switches.

The aspect ratio threshold (1.6) removes detections with unrealistic aspect ratios, ensuring only plausible detections are considered. The minimum box area parameter (10) filters out tiny boxes, reducing false positives. When enabled, the score and IoU fusion feature (set to False) combines detection score and IoU for association, further improving tracking performance.

4.3. Evaluation Results

Quantitative results. As shown in Tables 1, we compare our PRB-FPN+ against other SoTA object detection methods with respect to accuracy and efficiency. To begin with, we conducted a comparative analysis between our PRB-FPN+ and other existing models [11, 23–25, 31] using the MS COCO dataset [16]. Tables 1 shows that our proposed PRB-FPN+ outperforms the other SoTA object detection methods in terms of accuracy. Furthermore, we leveraged PRB-FPN+ in combination with SMILEtrack [27] to develop a video analytics system for enforcing motorcycle helmet laws in the AI City CHALLENGE 2023 [18] Track 5. Our proposed system achieved a rank of 8 on the Leaderboard, as shown in Tables 2. These results demonstrate the effectiveness and potential of our proposed approach for object detection and video analytics applications.

Qualitative results. We compared the object detection performance of YOLOv7-E6E [25] with our proposed PRB-FPN+. Figure 2(a) shows the detection results obtained using YOLOv7-E6E [25], which were unable to accurately detect a motor in a foggy scene. In contrast, Figure 2(b) shows the detection results obtained using our proposed PRB-FPN+, which successfully recognized the objects in the foggy scene. These results demonstrate the superior performance and robustness of our approach compared to YOLOv7-E6E [25] in challenging environments such as fog.

In addition to the superior performance of PRB-FPN+ compared to YOLOv7-E6E [25] in foggy scenes, our de-

Method	Size	FPS	AP	AP50	AP75	APS	APM	APL
YOLOv4-P6 [24]	1280	32	54.5	72.6	59.8	36.8	58.3	65.9
EfficientDet-D7 [23]	1536	8	53.7	72.4	58.4	35.8	57.0	66.3
SM-NAS: E5 [31]	1333x800	9	45.9	64.6	49.6	27.1	49.0	58.0
NAS-FPN [11]	1024	13	44.2					
YOLOv7-E6E [25]	1280	36	56.8	74.4	62.1	39.3	60.5	69.0
PRB-FPN+ [Ours]	1280	17	56.9	74.1	62.3	39.0	60.5	70.0

Table 1. Comparisons on the MS COCO test-dev set with SoTA models on Nvidia Volta V100.

Rank	Team ID	Team Name	Score
1	58	CTC-AI	0.8340
2	33	SKKU Automation Lab	0.7754
3	37	SMARTVISION	0.6997
4	18	UT_He	0.6422
5	16	UT_NYCU_SUNY-Albany	0.6389
6	45	UT_Chang	0.6112
7	192	Legends	0.5861
8	**55**	**NYCU - Road Beast**	0.5569
9	145	WITAI-513	0.5474
10	11	AIMIZ	0.5377

Table 2. The AI City Challenge 2023 [18] Track 5 Public Leaderboard, where our method ranks the 8-th among all participant teams.

Detection	Tracking	Score
✓		0.3685
✓	✓	0.3759

Table 3. The ablation of our method with and without tracking. The comparison is under 0.5 confidence and evaluate by the submission system.

Training	Validation	Score
70%	30%	0.3548
100%	100%	0.3685

Table 4. The ratio of different sets. The percentage represents the usage over the given dataset. The Score is evaluated on the leaderboard using our detection model with 0.5 confidence.

sign includes the addition of parallelization to the lead and auxiliary heads, which enables the detection of objects in the bottom of the image (as shown in Figure 2(b)). In contrast, YOLOv7 [25] is unable to detect these objects in the same image ((as shown in Figure 2(a))). These results further highlight the advantages of our proposed approach over YOLOv7 for object detection tasks in challenging environments.

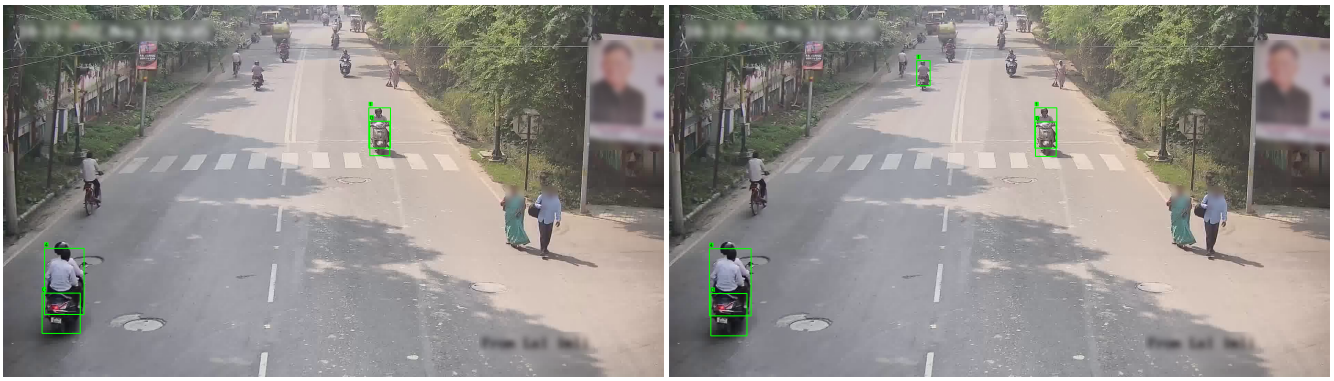
4.4. Ablation study

Data distribution. To evaluate whether the model will decrease the performance by using the whole data to train, we test on two different ratio splitting of the dataset and evaluate by the submission system. The result of two ratio splitting is shown in Table 4. The first separation is 70% of the training set and 30% of the validation set. The number of each class is shown in Figure 4 which is a fairness ratio. The second splitting is 100% of the training and validation set i.e. the training set is equal to the validation set. In our experiment, using all the data to train the model will get better performance. However, this will be different when adopting the pseudo label as discussed in the next section.

SMILEtrack [27] for post-processing. We performed an ablation study of our method with and without tracking, evaluating the results using the submission system at a confidence threshold of 0.5. As shown in Table 3, our findings suggest that incorporating tracking leads to improved performance. This implies that post-processing the tracking of drivers and passengers can enhance the accuracy and reliability of monitoring and enforcing helmet rules for motorcyclists in complex real-world situations.

4.5. Pseudo label analysis

According to the AIC Track 5 contest data labeling guidelines, any object that measures less than 40 pixels will not be labeled, nor will it impact the evaluation score. Nonetheless, our experiments have shown qualitative findings, illustrated in Figure 5, that our model can reliably detect small objects less than 40 pixels. However, this scenario may cause the model to confuse the object's features with those of the background. To overcome this issue, we adopt pseudo labeling, by utilizing the model's predictions to retrieve objects that were not labeled but can be detected. In other words, the pseudo labels are created by combining the ground truth label and the predictions made by our model. For each bounding box, we calculate the Intersection of Union (IoU) between ground truth and predictions. We exclude highly overlapping objects already present in the ground truth, but include those with low overlap, particularly non-overlapping boxes. This approach ensures that our model is trained on high-quality, representative data, improving its accuracy and performance in real-world scenarios.



(a) Ground Truth (b) Pseudo label

Figure 5. (a) The given dataset label. (b) The pseudo label composite ground truth label and the model prediction.

Pseudo label	mAP@.5	mAP@.5.95	Score
	0.954	0.785	0.3548
✓	0.9796	0.8755	0.3041

Table 5. The comparison between with and without the pseudo label. The result is evaluated on the detection model with 0.5 confidence.

Table 5 shows the evaluation results after incorporating pseudo labels. During the initial stages of model training, utilizing the pseudo label yields improved performance in terms of mAP@.5 and mAP@.5.95 metrics compared to the model trained without them. However, we observed inferior model performance of the test evaluation scores. We suspect that this discrepancy may be attributed to overfitting of the training model. Further investigation is needed to better understand and address this issue.

5. Conclusions and Future Works

We present PRB-FPN+, an innovative method for efficient and accurate single-shot object detection, which surpasses current state-of-the-art models. PRB-FPN+ has achieved an unprecedented level of performance on the challenging MS COCO dataset [16]. Our method employs both auxiliary and lead heads in parallel, enabling us to effectively extract features for recognizing and localizing objects of varying sizes, without compromising on efficiency. We utilize SmileTrack [27] to enhance the tracking of drivers and passengers, marking a practical application for enforcing motorcycle helmet laws. Our system achieves rank 8 in the AI City Challenge 2023 [18] Track 5 contest [1]. This demonstrates the efficacy and potential of our method in real-world smart city applications.

Limitations: Despite the significant advancements in our approach, there remains a limitation in the object detector's ability to accurately detect and differentiate between individuals with and without helmets. This limitation can be attributed to the inherent challenges in distinguishing subtle visual differences between the two categories.

Future Works: As a potential avenue for future work, we will propose incorporating attention mechanisms, such as the Convolutional Block Attention Module [30] (CBAM), to enhance the model's focus on relevant features and improve its discriminative power. This could potentially lead to more accurate detection and recognition of individuals with and without helmets, further advancing the state-of-the-art in multiple object tracking.

6. Acknowledgements

This work was supported by the National Science and Technology Council, Taiwan, under Grants NSTC109-2221-E-009-116-MY3.

References

[1] AI City Challenge, Challenge 2023 Track 5 Contest: Detecting violation of helmet rule for motorcyclists, 2023. https://www.aicitychallenge.org/2023-challenge-tracks/. 2, 5, 8

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, pages 3464–3468. IEEE, 2016. 3

[3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 1, 3

[4] A Chairat, MN Dailey, S Limsoonthrakul, M Ekpanyapong, and DR KC. Low cost, high performance automatic motorcycle helmet violation detection. In *WACV*, pages 3549–3557. IEEE, 2020. 2

[5] Ping-Yang Chen, Ming-Ching Chang, Jun-Wei Hsieh, and Yong-Sheng Chen. Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection. *IEEE Transactions on Image Processing*, 30:9099–9111, 2021. 1, 2, 3, 4, 5

[6] S Chen, J Lan, H Liu, C Chen, and X Wang. Helmet wearing detection of motorcycle drivers using deep learning network with residual transformer-spatial attention. *Drones*, 6(12):1–26, 2022. 3

[7] K Dahiya, D Singh, and CK Mohan. Automatic detection of bike-riders without helmet using surveillance videos in real-time. In *IJCNN*, pages 3046–3051. IEEE, 2016. 2

[8] Xiaohan Ding, Xiangyu Zhang, Zhaowei Cai, Ding Liang, Guo-Jun Qi, and Jianping Shi. YOLOX: Exceeding yolo series in 2021, 2021. 1, 3, 5

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, pages 6569–6578, 2019. 3

[10] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. PASCAL VOC 2012. `http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html`, 2012. 5

[11] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. NAS-FPN: Learning scalable feature pyramid architecture for object detection. In *CVPR*, June 2019. 6, 7

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 3

[13] A Goyal, D Agarwal, A Subramanian, CV Jawahar, RK Sarvadevabhatla, and R Saluja. Detecting, tracking and counting motorcycle rider traffic violations on unconstrained roads. In *CVPR Workshop*, pages 4302–4311. IEEE, 2022. 3

[14] Wei Jia, Shiquan Xu, Zhen Liang, Yang Zhao, Hai Min, Shujie Li, and Ye Yu. Real-time automatic helmet detection of motorcyclists in urban traffic using improved YOLOv5 detector. *IET Image Processing*, 15(14):3623–3637, 2021. 2

[15] Hei Law and Jia Deng. CornerNet: Detecting objects as paired keypoints. In *ECCV*, September 2018. 3

[16] Tsung-Yi Lin et al. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2, 5, 6, 8

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 3

[18] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Sanjita Prajapati, Alice Li, Shangru Li, Krishna Kunadharaju, Shenxin Jiang, and Rama Chellappa. The 7th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023. 1, 2, 5, 6, 7, 8

[19] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 3

[20] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 3

[22] D Singh, C Vishnu, and CK Mohan. Real-time detection of motorcyclist without helmet using cascade of cnns on edge-device. In *ITSC*, pages 1–8. IEEE, 2020. 3

[23] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *CVPR*, pages 10778–10787, 2020. 3, 5, 6, 7

[24] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In *CVPR*, pages 13029–13038, 2021. 3, 5, 6, 7

[25] Chien-Yao Wang, Hong-Yuan Liao, and Yueh-Hua Wu. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *arXiv 2207.02696*, 2022. 1, 2, 3, 4, 5, 6, 7

[26] L Wang et al. Investigation into recognition algorithm of helmet violation based on YOLOv5-CBAM-DCN. *IEEE Access*, 10:60622–60632, 2022. 3

[27] Yu-Hsiang Wang, Jun-Wei Hsieh, Ping-Yang Chen, and Ming-Ching Chang. SMILEtrack: Similarity learning for multiple object tracking, 2022. 2, 3, 4, 5, 6, 7, 8

[28] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *ECCV*, 2020. 3

[29] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, pages 3645–3649, 2017. 3

[30] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 8

[31] Lewei Yao, Hang Xu, Wei Zhang, Xiaodan Liang, and Zhenguo Li. SM-NAS: Structural-to-modular neural architecture search for object detection. *AAAI*, 34:12661–12668, Apr. 2020. 6, 7

[32] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 129:3069–3087, 2021. 3

[33] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 3