Video Analytics for Detecting Motorcyclist Helmet Rule Violations

Chun-Ming Tsai[1], Jun-Wei Hsieh[2], Ming-Ching Chang[3], Guan-Lin He[1], Ping-Yang Chen[2], Wei-Tsung Chang[1], Yi-Kuan Hsieh[2]

[1]Department of Computer Science, University of Taipei, Taipei 10048, 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
cmtsai@go.utaipei.edu.tw, jwhsieh@nycu.edu.tw, mchang2@albany.edu.

Abstract

The use of helmets is essential for motorcyclists' safety, but non-compliance with helmet rules remains a common issue. In this study, we extend the frontier of AI video analytic technologies for detecting violations of helmet rules among motorcyclists. Our method can handle highly challenging conditions for traditional methods, including occlusions, fast vehicle movement, shadows, large viewing angles, poor illumination and weather conditions. We adopt the widely used YOLOv7 object detector and develop a first baseline using YOLOv7-E6E. We further develop two improved versions, namely YOLOv7-CBAM and YOLOv7-SimAM that better address the challenges. Experiments are performed on the 2023 AI City Challenge Track 5 contest benchmark. Evaluation on the 100 test videos of the contest demonstrates the effectiveness of our approach. The baseline YOLOv7-E6E model trained with image size 1920 achieves 0.6112 mAP. The YOLOv7-CBAM achieves 0.6389 mAP, and YOLOv7-SimAM achieves 0.6422 mAP, where both are trained with image size 1280. These models rank sixth, fifth, and fourth on the public leaderboard, respectively, which outperforms over 36 global participating teams. The code for our models is available at: https://github.com/cmtsai2023/AICITY2023_Track5_DVHRM.

1. Introduction

The AI City Challenge (AIC) have spurred research into vision problems in recent years, particularly in the realm of Intelligent Transportation Systems (ITS). These challenges have addressed a range of issues, including vehicle re-identification, vehicle tracking, vehicle anomaly detection, tracked vehicle retrieval using natural language, naturistic driver data analytics, and traffic safety [1].

AI and computer vision are increasingly being applied in the field of ITS. One critical issue that can be addressed using this technology is the accurate and automatic detection of motorcyclists without helmets, which is essential for enforcing strict regulatory traffic safety measures. This is the focus of AIC Track 5 Contest, which aims to detect violations of the helmet rule for motorcyclists [1].

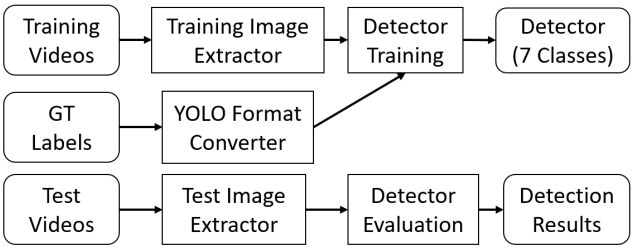

Figure 1. The flow diagram of the proposed method.

Motorcycles are widely used in many countries, especially in Asia and other tropical regions [17]. In developing countries like India, they are one of the most popular modes of transportation. However, due to their small size and lack of protection, motorcycle riders are at a greater risk of accidents compared to drivers of standard vehicles. Therefore, wearing helmets is mandatory as per traffic rules. However, accurately detecting motorcyclists without helmets can be challenging in real-world scenarios due to factors such as occlusion, movement, illumination, shadows, different viewing angles, and weather conditions. Additionally, motorcycles, drivers, and passengers vary widely in size, and the training and testing videos used in the AIC 2023 Track 5 Contest come from diverse traffic cameras in India.

For the AIC Track 5 task, we need to identify motorcycles and their riders, with or without helmets. Specifically, we need to determine whether each rider (i.e., the driver, passenger 1, and passenger 2) in the 100 test videos is wearing a helmet. Participating teams will need to identify seven

classes: (1) Motorbike: the bounding box of the motorcycle. (2) DHelmet: the bounding box of the motorcycle driver wearing a helmet. (3) DNoHelmet: the bounding box of the motorcycle driver not wearing a helmet. (4) P1Helmet: the bounding box of passenger 1 on the motorcycle, wearing a helmet. (5) P1NoHelmet: the bounding box of passenger 1 on the motorcycle, not wearing a helmet. (6) P2Helmet: the bounding box of passenger 2 on the motorcycle, wearing a helmet. (7) P2NoHelmet: the bounding box of passenger 2 on the motorcycle, not wearing a helmet.

The automatic detection of motorcyclists without helmets is not only an image classification task but also involves locating where and when the seven classes appear in the testing videos. Therefore, it is essential to train a seven-class detection model using 100 training videos from Track 5 to accurately detect these classes.

YOLOv7 [17] surpasses all known object detectors in terms of speed and accuracy in the range from 5 FPS to 160 FPS, achieving the highest accuracy of 56.8% AP among all known real-time object detectors with 30 FPS or higher on GPU V100. When applying the YOLOv7-E6E [17] pre-trained model to the test images from AI City CHALLENGE Track5, specifically video 008 frame 84 and video 014 frame 118, the detection results are shown in Figures 2 (a) and 2(b). Figure 2(a) shows a clear weather day where eight motorcycles and nine persons are detected correctly using the YOLOv7-E6E [17] pre-trained model. However, there are many non-motorcycle drivers that should be removed, and the detected person cannot be classified into DHelmet, DNoHelmet, P1Helmet, P1NoHelmet, P2Helmet, and P2NoHelmet. Therefore, we will use the YOLOv7-E6E [17] model to train the object detector to identify these seven classes.

In Figure 2(b), the foggy night and very bright headlights make it challenging to see the motorcycle and driver clearly. Although the YOLOv7-E6E [17] pre-trained model detects a false positive person in the advertisement on the right side of the image, it fails to detect the motorcycle and driver. To address these challenges, we propose training the YOLOv7-E6E [17] model on the Track5 training dataset.

Despite the high performance of YOLOv7-E6E [17] in terms of speed and accuracy, accurately detecting the seven objects in Track5 testing images and videos still presents challenges. Therefore, further research is needed to develop more robust and accurate object detection models for the seven classes of motorcycle and rider detector in the Track5 challenge.

Figure 1 shows the proposed method, which includes the following steps: first, extracting each frame image from the 100 training videos. Second, converting the ground truth labels to the YOLO format. Third, training the proposed model that combines YOLOv7-E6E [17] with CBAM [20] and SimAM [21] to detect the seven classes of motorcycle and rider. Fourth, using the seven classes' motorcycle and rider detector to detect the 100 testing videos. Finally, we obtain the detection results for the seven classes.

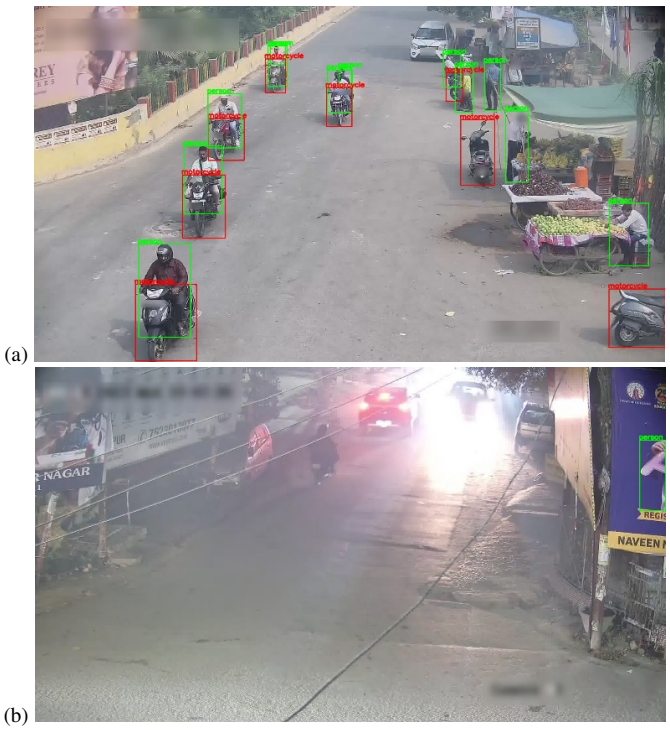

Figure 2. Examples of person and motorcycle detection using the YOLOv7-E6E model. (a) Frame 84 of video 008 with satisfactory detection results. (b) Frame 118 of video 014 with a false and many miss detections.

The remainder of this paper is structured as follows. Section 2 describes the related works that have been done in the field of motorcycle and rider detection. Section 3 presents the proposed method, which combines YOLOv7-E6E [17] and CBAM [20] and combines YOLOv7-E6E [17] and SimAM [21] to improve the accuracy of the seven-class motorcycle and rider detector. In Section 4, we present the experimental results and discuss the findings. Finally, Section 5 summarizes our findings and presents our conclusions.

2. Related Works

This section provides the current related research in object detection and without helmet detection.

2.1. Object Detection

Object detection is one of the most popular tasks in image processing and computer vision. It involves identifying the position of objects in an image by predicting their bounding boxes and classes. The task of detecting seven specific classes is a specialized branch of object detection

that focuses on these specific classes in images or videos. This can be achieved through various backbones, including CNN-based object detectors [2,5,8,12,14] and transformer-based detectors [3, 13, 23].

Thanks to the success of convolutional networks, CNN-based detectors have made significant progress, including Faster-RCNN [14], SSD [12], Cascade-RCNN [2], Yolox [8], PRB-Net [5], and YOLOv7 [17]. SSD, Yolox, PRB-Net, and YOLOv7 are one-stage detectors that prioritize speed and accuracy to run in real-time, while Faster-RCNN and Cascade-RCNN are two-stage detectors that are usually more accurate and flexible but time-consuming. Another branch of object detection is transformer-based detectors, which draw inspiration from the success of natural language processing.

The transformer architecture can learn sequences using a self-attention mechanism [3]. DETR [23] and Swin Transformer [13] are examples of such object detectors that introduced vision transformers to achieve competitive performance on object detection benchmarks by treating an image as a series of patches. Typically, a CNN-based detector can capture spatial information within each patch, allowing it to handle spatially local patches well, while a transformer-based detector is better suited for capturing long-distance pixel relationships.

2.2. Detecting Motorcyclists Not Wearing Helmets

Dahiya *et al.* [7] proposed an approach for the automatic detection of motorcycle drivers without helmets using surveillance videos in real-time. In their approach, they first detect motorcycle drivers from surveillance videos using background subtraction and object segmentation. Then, they determine whether the motorcycle drivers are using helmets or not using visual features and an SVM binary classifier. From their experimental results, it shows that the detection accuracy is 93.80% on the real-world surveillance data. However, this method cannot detect the motorcycle passengers, including passenger 1 and passenger 2.

Soni and Singh [16] have developed a system in the field of computer vision based on Tensorflow and Keras. Their system can detect in real-time whether motorcyclists are wearing a helmet or not. If a motorcyclist is detected without a helmet, the system will accurately identify the situation and flag the rule violation. However, the size of their testing dataset is limited to only 50 samples.

Chairat *et al.* [4] proposed an automated system for detecting helmet violations to identify riders and passengers not wearing helmets. Their system employed YOLO for motorcycle detection, Kristan's method for tracking, GoogleNet for classification, and a specific system architecture for processing multiple cameras to detect helmet violations. The violation class and non-violation class contain 960 and 931 images, respectively. They trained and tested the top half of the motorcycle bounding box. In real-world testing, their system detected 97% of helmet violations with a false alarm rate of 15%. However, their approach only considered motorcycles with two riders.

Singh *et al.* [15] proposed a framework for automatic detection of motorcyclists without helmets. Their method includes a motorcyclist detector, person localization, and head and helmet classifier. The first dataset contains sparse traffic, which is a two-hour surveillance video data collected at 30 frames per second. The first hour of the video is used for training, which contains 42 motorcycles, 13 cars, and 40 humans. The second hour is used for testing, which contains 63 motorcycles, 25 cars, and 66 humans. Their second dataset contains dense traffic, which is a 1.5-hour video collected at 25 frames per second. The first half-hour of the video is used for training the model, which contains 1261 motorcyclists and 4960 non-motorcyclists. The remaining video is used for testing, which contains 2312 motorcycles and 9112 non-motorcyclists. Their experimental results demonstrate the efficacy of their proposed approach. However, this method cannot detect motorcycle passengers, including passengers 1 and 2.

Giron *et al.* [9] developed an approach to classify motorcycle riders as either wearing a helmet or not using deep machine learning, specifically convolutional neural network, and by utilizing different pre-trained models on a gathered dataset. However, their study was limited by the small size of the dataset, which consisted of only 400 images, with 320 images for training and 80 images for testing. These 400 images were divided equally into two classes: 200 images of riders wearing helmets and 200 images of riders not wearing helmets.

Jia *et al.* [11] proposed an automatic method to detect helmet-wearing motorcyclists based on deep learning. Their method involves two stages: (1) using an improved YOLOv5 detector that incorporates triplet attention and soft-NMS instead of NMS to detect motorcycles (including motorcyclists) in video surveillance, and (2) using the same detector to detect whether the motorcyclists wear helmets. They also introduced a new motorcycle helmet dataset (HFUT-MH) that is larger and more comprehensive than existing datasets derived from multiple traffic monitoring in Chinese cities. The proposed method was validated through experiments and compared to other state-of-the-art methods. Their method achieved a mAP of 97.7%, a F1-score of 92.7%, and 63 FPS, outperforming other detection methods. However, this method cannot detect motorcycle passengers, including passenger 1 and passenger 2.

Goyal *et al.* [10] proposed an approach to detect, track, and count violations of motorcycle riding in videos captured from a vehicle-mounted dashboard camera. To tackle challenging scenarios such as occlusions, they employed a curriculum learning-based object detector. They also intro-

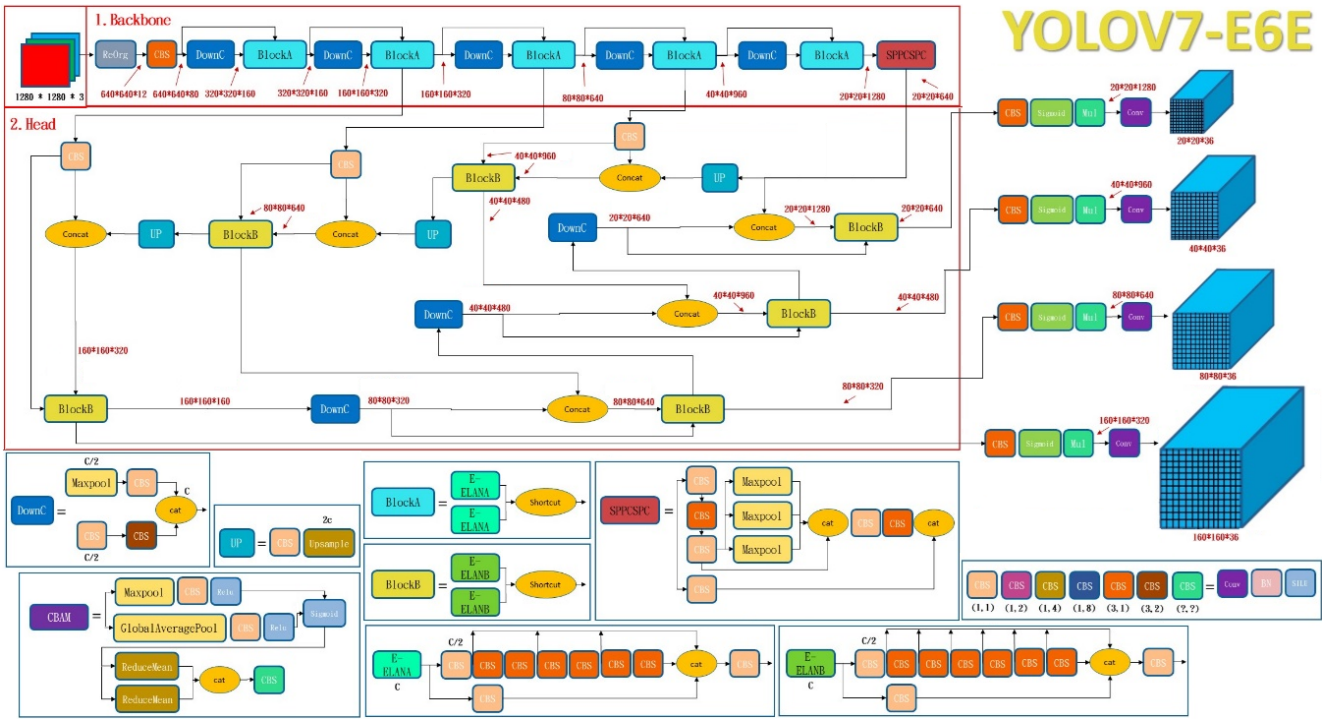

Figure 3. The YOLOv7-E6E architecture adopted from [17] for person and motorcycle detection.

duced a trapezium-shaped object boundary representation to increase robustness and handle the rider-motorcycle association. In addition, they introduced an amodal regressor that generates bounding boxes for occluded riders. The experimental results on a large-scale unconstrained driving dataset demonstrate the superiority of their approach compared to existing approaches and other ablative variants. However, this method cannot detect motorcycle passengers, including passenger 1 and passenger 2.

Wang *et al.* [18] proposed a safety helmet detection method that utilizes YOLOv5-CBAM-DCN with an attention mechanism and deformable convolution. This method addresses the issue of insufficient accuracy faced by traditional target algorithms due to complex site environments, uneven lighting, and irregular target shapes. However, it is unable to detect motorcycle passengers, such as passenger 1 and passenger 2.

Waris *et al.* [19] proposed a system for automatically detecting helmet violations from surveillance videos captured by roadside-mounted cameras. Their technique is based on a faster region-based convolutional neural network (R-CNN) deep learning model that takes videos as input and detects helmet violations to take necessary actions against traffic rule violators. Experimental analysis shows that their system achieves an accuracy of 97.69% and outperforms its competitors. Their dataset includes 13,631 images of drivers wearing helmets and 10,169 images of drivers not wearing helmets. However, this method cannot detect motorcycle passengers, including passenger 1 and passenger 2.

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. Their dataset comprised 4555 target images with helmets and 3164 target images without helmets. However, this approach does not account for the detection of motorcycle passengers, such as passenger 1 and passenger 2.

3. Methods

Based on the information presented in [17] and above, it is evident that the YOLOv7-E6E model, with a test size of 1280, achieves an APtest of 56.8% and an AP50test of 74.4%, surpassing all known real-time object detectors. Therefore, we have chosen the YOLOv7-E6E model as the baseline for our seven-class object detector. Furthermore, previous research on without-helmet detection has shown that attention mechanisms can significantly improve detection accuracy. Hence, in this paper, we propose two seven-class object detectors that combine the YOLOv7-E6E model with CBAM and SimAM, respectively. We provide a brief overview of these models below.

3.1. YOLOv7-E6E model

The architecture of the YOLOv7-E6E model is depicted in Figure 3. This model represents an improvement over several previous models, including YOLOv4, Scaled YOLOv4, and YOLO-R, and was developed through further experimentation, resulting in enhancements and new

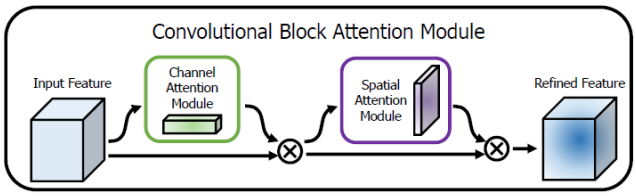

Figure 4. The CBAM structure [18].

features. The YOLOv7 backbone contains a computational block called E-ELAN (Extended Efficient Layer Aggregation Network), which employs expand, shuffle, and merge cardinality to improve the network's learning ability without compromising the gradient path.

Different applications require specific models, with some prioritizing accuracy and others prioritizing speed. YOLOv7 addresses these requirements by allowing model scaling to accommodate various computing devices. The scaling process considers parameters such as resolution (input image size), width (number of channels), depth (number of layers), and stage (number of feature pyramids).

The predicted outputs are in the head of YOLOv7, which consists of multiple heads. The Lead Head is responsible for the final output, while the Auxiliary Head is used to assist in training in the middle layers.

Overall, the YOLOv7-E6E model is highly efficient and effective for real-time object detection applications with limited resources. More information about YOLOv7-E6E can be found in reference [17].

3.2. CBAM

CBAM (Convolutional Block Attention Module) is an attention module proposed by Woo *et al.* [18] that applies attention mechanisms to enhance the representation power of CNNs by emphasizing important features and suppressing irrelevant ones. The CBAM structure is depicted in Figure 4, and it comprises two modules: channel attention and spatial attention. The channel attention module focuses on "what" is meaningful given an input image. It learns to weight the importance of each feature map channel based on its global distribution. On the other hand, the spatial attention module focuses on "where" the informative part of the image is located. It employs average-pooling and max-pooling operations along the channel axis to learn and highlight informative features, which are then concatenated to generate an efficient feature descriptor. CBAM can be integrated into any CNN architecture with negligible overheads and is end-to-end trainable along with the base CNNs. For more information about CBAM, please refer to reference [18].

3.3. SimAM

CBAM separately estimates 1-D channel attention and 2-D spatial attention and combines them, rather than directly generating true 3-D weights. Moreover, the two-step process in CBAM is computationally expensive. To address these limitations, Yang *et al.* [22] proposed SimAM (A Simple Attention Module), which is a simple and parameter-free attention module that can efficiently produce true 3-D weights. SimAM is based on well-known neuroscience theories and optimizes an energy function to determine the importance of each neuron. Yang *et al.* demonstrated that SimAM is a lightweight module that can be used for various vision tasks.

3.4. The proposed models

Based on the results presented in [17], the YOLOv7-E6E model achieves the highest performance among real-time object detectors, with an APtest of 56.8%, AP50test of 74.4%, and AP75test of 62.1% for a test size of 1280. Additionally, CBAM and SimAM attention modules have demonstrated improvements in detection accuracy for various vision tasks.

From above-mentioned, we know YOLOv7-E6E model has the highest performance in test size is 1280, APtest is 56.8%, AP50test is 74.4%, and AP75test is 62.1%. We also found that the channel and spatial attentions in CBAM which sequential arrangement gives a better result. Furthermore, the channel-first order is slightly better than the spatial-first. Thus, we propose a new model is called YOLOv7-CBAM model which combine YOLOv7-E6E model and CBAM. That is, we insert three CBAM module between three DownC and concat in the Head in Figure 3, respectively.

We propose two models that combine the YOLOv7-E6E model with attention mechanisms: YOLOv7-CBAM and YOLOv7-SimAM. In the YOLOv7-CBAM model, we insert three CBAM modules between three DownC layers and concat layers, respectively, in the Head part, as shown in Figure 3. The YOLOv7-SimAM model follows a similar structure, but with three SimAM modules instead of CBAM modules.

SimAM is a parameter-free attention module that achieves competitive results against other attention modules while maintaining efficiency in terms of speed and parameters. Therefore, the proposed YOLOv7-SimAM model provides a lightweight and effective solution for object detection tasks. For details, refer to the original papers of CBAM [18] and SimAM [22].

Overall, these two proposed models, YOLOv7-E6E + CBAM and YOLOv7-E6E + SimAM, offer promising enhancements to the YOLOv7-E6E model, which already achieves state-of-the-art performance in real-time object detection. By incorporating attention mechanisms, these models aim to improve the ability of the network to identify important features and suppress irrelevant ones, leading to improved detection accuracy. The choice between CBAM

and SimAM depends on the Track5 requirements of the task at hand, with CBAM offering a more complex attention module with greater flexibility in weight computation, while SimAM provides a simple and lightweight solution with competitive performance. Further experimental evaluation will show in Experiments Section to determine the effectiveness of these proposed models on Track 5 of seven classes object detection datasets and scenarios.

4. Experimental Results

We conducted experiments to evaluate the performance of our proposed YOLOv7-CBAM and YOLOv7-SimAM models on the testing images of Track 5 in the 2023 AI City CHALLENGE. Our proposed methods were implemented on an Ubuntu 18.04.6 LTS operating system using NVIDIA RTX A5000 24GB*2 GPUs and GeForce 3090 24GB*2 GPUs, with Nvidia driver version 470.74, CUDA 11.4 driver, and Python 3.8.16 for training. For testing and detection, we used an Ubuntu 18.04.6 LTS operating system with NVIDIA GeForce 2080 11GB*4 GPUs, Nvidia driver version 510.54, CUDA 11.6 driver, and Python 3.8.13.

4.1. Datasets

The training dataset for AIC 2023 Track 5 Contest consists of 100 videos with ground truth bounding boxes of motorcycles and motorcycle riders, with or without helmets. Each video is 20 seconds long and recorded at a frame rate of 10 fps, with a video resolution of 1920x1080. The annotations provide bounding boxes for each motorcycle and up to three riders, with helmet information for each rider. The object classes in this dataset are labeled as follows: (1) motorbike - bounding box of the motorcycle, (2) DHelmet - bounding box of the motorcycle driver wearing a helmet, (3) DNoHelmet - bounding box of the motorcycle driver without a helmet, (4) P1Helmet - bounding box of passenger 1 wearing a helmet, (5) P1NoHelmet - bounding box of passenger 1 without a helmet, (6) P2Helmet - bounding box of passenger 2 wearing a helmet, and (7) P2NoHelmet - bounding box of passenger 2 without a helmet.

The test dataset for this track also consists of 100 videos, each 20 seconds long and recorded at 10 fps. The objective is for participating teams to identify motorcycles and motorcycle riders with or without helmets. Similar to the training dataset, each rider (i.e. driver, passenger 1, and passenger 2) in a motorcycle must be identified separately, with helmet information for each rider.

4.2. Evaluation

The performance of each team will be ranked based on the mean Average Precision (mAP) across all frames in the test videos. The mAP metric measures the mean of average precision (the area under the Precision-Recall curve) over all object classes, as defined in the PASCAL VOC 2012 competition.

To meet the labeling standard, an object must have at least 40% visibility. The minimum height and width of the bounding boxes are 40 pixels. Objects smaller than 40 pixels will not be considered in the test accuracy results. Objects that overlap with the redacted area (blurred region) will also be ignored because the blurred region can obscure important object features. Any objects that overlap with the redacted areas in the test dataset will be ignored and will not affect the test accuracy.

4.3. Implementation Details

In order to establish a baseline object detector for the seven classes, we trained the YOLOv7-E6E model with an image size of 1920, using the training parameters specified in hyp.scratch.p6.yaml. The training dataset and validation dataset included 100 training videos from AIC Track 5 Contest. After completing the YOLOv7-E6E model training, we applied the seven-class object detector to detect the seven classes in the test images for Track 5. The resulting detections were uploaded to the evaluation system to determine the mAP score. To improve the mAP score, we adjusted the confidence and IoU threshold values for detecting the test images in Track 5.

We then trained the proposed YOLOv7-CBAM model with an image size of 1280 to obtain the seven-class object detector. During training, we used the 100 training videos for the training dataset and 001 025 and 075 100 training videos for the validation dataset. The training parameters were identical to those specified in hyp.scratch.p6.yaml. After completing the YOLOv7-CBAM model training, we utilized the seven-class object detector to detect the seven classes in the test images for Track 5. The resulting detections were submitted to the evaluation system to determine the mAP score. To improve the mAP score, we adjusted the confidence and IoU threshold values when detecting the test images of AIC Track 5 test set.

Finally, we trained the proposed YOLOv7-SimAM model with an image size of 1280 to obtain the seven-class object detector. During training, we used the 100 training videos for the training dataset and 001 025 and 075 100 training videos for the validation dataset. The training parameters were the same as those specified in hyp.scratch.p6.yaml. After completing the YOLOv7-SimAM model training, we utilized the seven-class object detector to detect the seven classes in the test images for Track 5. The resulting detections were submitted to the evaluation system to determine the mAP score. To improve the mAP score, we adjusted the confidence and IoU threshold values when detecting the test images in Track 5.

Table 1. The AI City Challenge 2023 Track 5 Public Leaderboard.

Rank	Team ID	Team Name	mAP
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

4.4. Evaluation Results

The detection results of our proposed models were submitted to Track 5 of the AI City CHALLENGE 2023 for evaluation. As shown in the AIC 2023 Track 5 Public Leaderboard in Table 1, we achieved mAP scores of 0.6112, 0.6389, and 0.6422 for the baseline YOLOv7-E6E model training with an image size of 1920, YOLOv7-CBAM model training with an image size of 1280, and YOLOv7-SimAM model training with an image size of 1280, respectively. These scores ranked us sixth, fifth, and fourth on the public leaderboard among over 36 teams worldwide.

4.5. Qualitative Results

Figure 5(a) presents the detection results of the baseline model, YOLOv7-E6E, which was trained with an image size of 1920, on an image extracted from video 008, frame 84. The figure displays the detection of five moving motorcycles as motorbikes, but it also shows a false positive detection of a non-moving motorcycle in the bottom-right corner. The drivers of the first and fourth moving motorcycles have been correctly identified as DHelmet, whereas the driver of the other moving motorcycle is identified as DNoHelmet. However, the drivers of the third and fifth moving motorcycles have also been erroneously detected as DHelmet. The image also includes a false positive detection of DNoHelmet in the bottom-right. In contrast, Figure 5(b) depicts the detection results of the same baseline model, YOLOv7-E6E, trained with an image size of 1920, on an image extracted from video 014, frame 118. In this case, only one motorcycle has been detected as a motorbike, and the driver has not been detected.

Figure 6(a) presents the detection results of the proposed model, YOLOv7-CBAM, trained with an image size of 1280, on an image extracted from video 008, frame 84. The figure displays the detection of five moving motorcycles as motorbikes, but it also shows a false positive detection of a non-moving motorcycle in the bottom-right corner. The drivers of the first and fourth moving motorcycles have been correctly identified as DHelmet. The driver of the second moving motorcycle has been correctly identified as DNoHelmet. However, the driver of the third moving motorcycle has been erroneously identified as DHelmet. The driver of the fifth moving motorcycle is identified as both DHelmet and DNoHelmet. Figure 6(b) depicts the detection results of the same proposed model, YOLOv7-CBAM, with an image size of 1280, on an image extracted from video 014, frame 118. In this case, one motorcycle has been detected as a motorbike, and the driver has also been correctly identified as DHelmet.

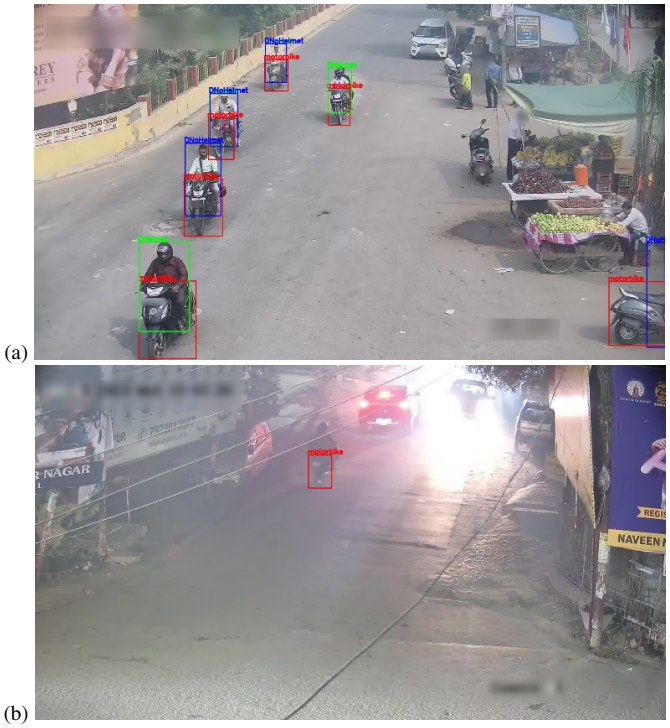

Figure 5. Examples of the seven classes detected in Track 5 using the baseline model: YOLOv7-E6E trained with an image size of 1920. (a) The image, taken from video 008, frame 84, shows five detected motorbikes and one false positive, two DHelmets, two false positive DHelmets, and three DNoHelmets. (b) The image, taken from video 014, frame 118, shows only one detected motorbike.

Figure 7(a) presents the detection results of the proposed model, YOLOv7-SimAM, trained with an image size of 1280, on an image extracted from video 008, frame 84. The figure displays the detection of five moving motorcycles as motorbikes, but it also shows a false positive detection of a non-moving motorcycle in the bottom-right corner. The drivers of the first and fourth moving motorcycles have been correctly identified as DHelmet. The drivers of the second and fifth moving motorcycles have also been correctly identified as DNoHelmet. However, the driver of the third moving motorcycle is identified as a false positive DHelmet. Additionally, there is a false positive DNoHel-

Figure 6. Illustrates examples of the seven detected classes in Track 5 using the proposed model, YOLOv7-CBAM, trained with an image size of 1280. In (a), the image extracted from video 008, frame 84 displays the detection of five motorbikes, one false positive motorbike, two DHelmets, two false positive DHelmets, and two DNoHelmets. In (b), the image extracted from video 014, frame 118, shows the detection of one motorbike and one DHelmet.

Figure 7. Examples of the seven detected classes in Track 5 using the proposed model, YOLOv7-SimAM, trained with an image size of 1280. (a) The image is taken from video 008, frame 84, and displays the detection of five motorbikes and one false positive motorbike, two DHelmets and one false positive DHelmet, and two DNoHelmets and one false positive DNoHelmet. (b) The image is taken from video 014, frame 118, and shows the detection of only one DHelmet. The motorbike was not detected.

met detected in the bottom-right corner. Figure 7(b) depicts the detection results of the same proposed model, YOLOv7-SimAM, with an image size of 1280, on an image extracted from video 014, frame 118. In this case, while one motorcycle was not detected, the driver has been correctly identified as DHelmet.

5. Conclusion

This paper presents two new deep learning models, YOLOv7-CBAM and YOLOv7-SimAM, which incorporate YOLOv7-E6E, CBAM, and SimAM. The YOLOv7-E6E model was trained on images of size 1920, while the YOLOv7-CBAM and YOLOv7-SimAM models were trained on images of size 1280. These models were employed to detect the test images in Track 5, and the results were submitted to the AI City CHALLENGE Track 5 evaluation system. The experimental results on the 100 test videos of the 2023 AI City CHALLENGE Track 5 demonstrate the effectiveness of our methods, with mAP scores of 0.6112, 0.6389, and 0.6422 for YOLOv7-E6E, YOLOv7-CBAM, and YOLOv7-SimAM, respectively. Our proposed methods ranked sixth, fifth, and fourth on the public leaderboard, out of over 36 participating teams. However, YOLOv7-CBAM produced one false positive motorbike and two false positive DHelmets in Figure 6(a), while YOLOv7-SimAM generated one false positive motorbike, one false positive DHelmet, and one false positive DNoHelmet in Figure 7(a). Moreover, YOLOv7-SimAM failed to detect one motorbike in Figure 7(b). In the future, these issues regarding false positives and non-detected objects should be addressed.

Overall, the proposed models demonstrate the potential of deep learning techniques in improving traffic safety measures, and they highlight the importance of continued research in this area. The code for these models is publicly available, enabling others to build upon these advancements and further improve the state-of-the-art.

Acknowledgements. This work was supported by the Ministry of Science and Technology, Taiwan, under Grants MOST 108-2221-E-845-003-MY3.

References

[1] 2023 AI City Challenge, Challenge Track 5: Detecting violation of helmet rule for motorcyclists, 2023. https://

www.aicitychallenge.org/2023-challenge-tracks/. 1

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high-quality object detection. In *CVPR*, pages 6154–6162, 2018. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 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. 3

[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. 3

[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. 4

[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. 3

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

[9] NNF Giron et al. Motorcycle rider helmet detection for riding safety and compliance using convolutional neural networks. In *HNICEM*, pages 1–6. IEEE, 2020. 3

[10] 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

[11] 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. 3

[12] 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

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 3

[14] 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

[15] 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

[16] A Soni and AP Singh. Automatic motorcyclist helmet rule violation detection using tensorflow & keras in opencv. In *SCEECS*, pages 1–5. IEEE, 2020. 3

[17] 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, 2022. 1, 2, 3, 4, 5

[18] L Wang et al. Investigation into recognition algorithm of helmet violation based on yolov5-cbam-dcn. *IEEE Access*, 10:60622–60632, 2022. 4, 5

[19] Tasbeeha Waris, Muhammad Asif, Maaz Bin Ahmad, Toqeer Mahmood, Sadia Zafar, Mohsin Shah, and Ahsan Ayaz. Cnn-based automatic helmet violation detection of motorcyclists for an intelligent transportation system. *Mathematical Problems in Engineering*, 2022:11, 2022. 4

[20] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 2

[21] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. Proceedings of the 38th international conference on machine learning. In *PMLR*, volume 139, pages 11863–11874, 2021. 2

[22] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. SimAM: A simple, parameter-free attention module for convolutional neural networks. *PMLR*, 139:11863–11874, 2021. 5

[23] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3