# An Effective Method for Detecting Violation of Helmet Rule for Motorcyclists

Yunliang Chen[*], Wei Zhou[†], Zicen Zhou, Bing Ma, Chen Wang,
Yingda Shang, An Guo, Tianshu Chu

China Mobile Shanghai ICT Co.,Ltd

{chenyunliang, zhouweigy, zhouzicen, mabing}@cmsr.chinamobile.com

{wangchen, shangyingda, guoan, chutianshu}@cmsr.chinamobile.com

## Abstract

*Motorcycles are one of the most popular modes of transportation. However, motorcycle riders are exposed to a greater risk of crashes due to fewer safety protection measures compared to drivers of cars or other standard vehicles. Thus helmet plays an essential role in protecting the safety of motorcycle riders and passengers. Nevertheless, people pay little attention to it, particularly in developing countries such as India. Therefore, video surveillance-based automatic detection of motorcyclists without wearing helmets is one of the critical tasks to enforce strict regulatory traffic safety measures. To this end, we introduce a simple yet effective method for detecting violation of helmet rule for motorcyclists in this paper. Two transformer-based detectors, DETA and Co-DETR, are employed to detect motorbikes and riders not wearing helmets. We enhance the generalization ability of our models using several data augmentation techniques. Furthermore, we explore different fusion strategies merging the predictions from different detection models to improve the performance of our method. Our method achieves a mAP of 0.4824 on the public leaderboard of 2024 AI City Challenge Track 5 without any tracking or post-processing steps.*

## 1. Introduction

Riding motorcycles without wearing helmets significantly increases the safety risk of the drivers and passengers. As a result, lawmakers have enacted relevant legislation to regulate such behaviors. A video surveillance-based automatic detection method to determine whether motorcycle drivers and passengers are wearing helmets plays a crucial role in enhancing the efficiency of law enforcement officers. Developing a method to detect motorcycle helmet violations is considered to make a substantial contribution to improve public safety and road safety.

Generally, due to the distance and camera angle, helmets are small or even tiny objects which is difficult to locate. Moreover, large differences in weather conditions and illuminations pose a great challenge to the performance and robustness of object detection algorithms. In addition, this task further requires the identification and localization of motorcycles as well as the driver and passengers sitting in different relative positions.

In recent years, the performance of object detection methods have appeared to be greatly improved due to the rapid development of neural networks, specifically, from CNN-based networks to transformer-based networks. CNN-based methods rely on local patterns in the image and are limited by the size and complexity of the receptive field. On the other hand, the attention mechanism used by transformer-based frameworks can extract non-local information by global modeling, which helps to effectively identify the global contextual information in the image. In this paper, transformer-based frameworks are used for detecting motorcycles and helmet violations of drivers and passengers.

However, the task of robust motorcycle helmet detection is still quite challenging. For instance, different sensors and the environment of the surveillance cameras bring about large differences of the captured images. There are differences in the occlusion of passengers on motorcycles in different traveling directions at fixed camera altitudes and angles. In addition, under complex road conditions, some bicycles and motorcycle tricycles also have the likelihood of being recognized as motorcycles. In order to distinguish these complex scenarios and improve the robustness of the detection model, we introduce the idea of data augmentation to be used in the inference and fusion of the model detection results. The implementation details will be shown in Section 4. With these approaches, we finally obtain more

[*]Corresponding author: chenyunliang@cmsr.chinamobile.com
[†]Equal contribution
[‡]Equal contribution

accurate results.

In summary, the main contributions of this paper are shown as follows:

1. We illustrate transformer-based detectors for motorcycle helmet violation detection.

2. We introduce several data augmentation and model ensemble strategies to improve the accuracy and robustness of our method.

3. The proposed method achieves the second place (a 0.4824 mAP) in Track 5 of 2024 AI City Challenge without using any tracking or post-processing techniques. Comprehensive experiments are done to show the effectiveness of the method.

The rest of this paper is organized as follows. Section 2 reviews the related works in the field of object detection. In section 3, the implementation details of the proposed motorcycle helmet violation detection method are presented. Experiments are shown with a large number of experimental results and ablation studies in section 4, and Section 5 is the conclusion.

## 2. Related Work

Object detection has always been one of the fundamental tasks in computer vision, which aims to localize and classify the objects of interest in images or videos, and eventually return the bounding boxes and categories of targets. Researchers have applied a variety of deep learning methods to solve the task of object detection. In 2014, R-CNN[5] introduced Convolution Neural Networks(CNNs) into the field of object detection for the first time by pre-selecting a series of region proposals, extracting features of region proposals and identifying the category. To avoid redundant computation, Fast-R-CNN[4] extracts features from the original image once. Then the local features are cropped and used to predict the positions and categories of the target boxes. Such an end-to-end framework greatly reduces the complexity of computation and boosts its performance. Faster-RCNN[8] integrates feature extraction, candidate region extraction, bounding box regression (rect refine), and classification in a single network, which further improves the comprehensive performance and detection speed of the network. The above methods based on region proposals are called two-stage approaches. Without region proposals, one-stage methods such as YOLO[7] directly output grid-wise classification and regression for bounding boxes according to their label assignment.

Instead of CNN-based backbone, the transformer-based network is also employed as backbone of the object detection framework. Transformer is first introduced by Facebook for natural language processing, which is still essentially an encoder-decoder structure, but employs a self-attention mechanism in encoder and decoder for feature extraction and reconstruction to efficiently extract contextual feature, and introduces multi-head attention to extract features from multiple perspectives. DETR[2] proposed a two-dimensional spatial location coding method, which employs position coding in two-dimensional space, and introduces multi-head attention to extract features from multiple perspectives. Attention mechanism in encoder and decoder for feature extraction and reconstruction, efficiently extracting correlations in context, and introducing multi-head attention to extract features from multiple perspectives. DETR[2] proposes a two-dimensional spatial location coding method, which incorporates location coding with encoder's self-attention and decoder's cross-attention, while object queries are added to multi-head attention of decoder. Deformable DETR[13] combines DCN[3] and self-attention[10] to design multi scale deformable attention module that it improves the problem of sparse attention weight matrix of DETR[2] and enhances the detection performance of small targets. DETA[6] adopts fixed overlap-based assignment (IoU-based) label-GT assignment train strategy to accelerate the model training speed to improve the prediction accuracy(map). Especially for the scenario of whether a motorcycle rider is wearing a helmet or not, the rider's head region is small and the information is difficult to capture, DETA[6] brings tremendous gain for fine-grained feature extraction in this region.

This paper proposes a transformer-based approach to detect motorcycles and passengers, especially for detecting and classifying whether the passenger violates the helmet rule or not. Our proposed framework contains a two-stage processing, detection and fusion module. The final results are merged the obtained after inference of the data-enhanced video, improving the model's detection robustness to different motion scenarios.

## 3. Method

In this section, a simple yet effective method for motorcycle helmet violation detection is proposed. Several data augmentation strategies are applied to improve diversities of input images. Except for general image processing, e.g. rotation, translation, and flipping, Mosaic[1], Mixup[12], blurring and brightness operations are also necessary to tackle low-light, fog, and background chaos. Then considering different camera angles and distances, the augmented inputs are cropped and resized randomly for detecting multi-scale objects later. The augmented images $I'$ are indicated as follows:

$$I' = f(I) \tag{1}$$

where the input images $I$ are normalized and cropped and resized to multi-scales and $f(\cdot)$ represent the sequential combination of image processing.

Moreover, as shown in Fig. 1, two state-of-the-art transformer-based models, Co-DETR[14] and DETA[6],

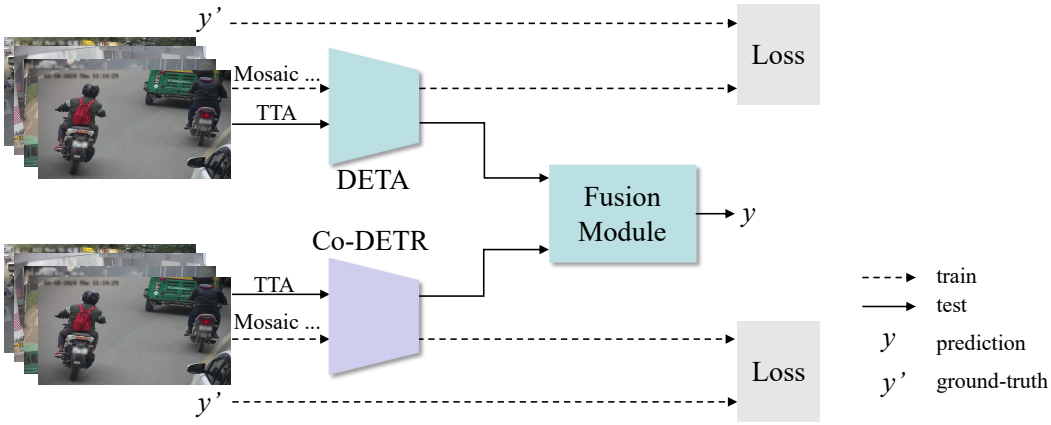


Figure 1. Pipeline of the proposed method for motorcycle helmet violation detection.

are employed to localize the bounding boxes of motorcycles, drivers, and passengers respectively. With pretraining on MS-COCO dataset, these DETR's[2] variants efficiently build the spatial and semantics relationship between objects from the global context. Their one-to-many label assignment strategies contribute to not only fast and stable convergence but also impressive performance. Their differences in model structure and label assignment strategies make the following model ensemble effective.

During testing, Test-Time Augmentation (TTA) can effectively improve the performance of both models by multi-scale inference. TTA achieves this by augmenting the test data with various transformations or perturbations, generating multiple augmented samples, and leveraging the model's predictions on these augmented samples to obtain a more reliable and accurate final prediction. Moreover, to merge the predictions from diverse models, we explore several fusion approaches, e.g. weighted box fusion (WBF)[9] and non-maximum suppression (NMS)[4].

$$\mathbf{B} = Fusion(D_1(g_1(I)) \cup D_2(g_2(I))) \tag{2}$$

where $g_i$ indicates Test-Time Augmentation (TTA) for the $i$-th model. $D_1$ and $D_2$ represents DETA[6] and Co-DETR[14] respectively. The final prediction $B$ is obtained by filtering those bounding boxes with low scores.

## 4. Experiments

### 4.1. Datasets

As a mean of transportation, motorcycle is very popular particularly in developing countries such as India. Without adequate protection, both of motorcycle riders and passengers suffer from a greater risk of crashes. Therefore, the Track 5 dataset of 2024 AI City Challenge [11] is obtained from various roads of an Indian city and used to detect motorcycles, drivers and passengers for their safety. This dataset aims to devise an algorithm that detects motorcycles and classifies motorcycle riders with respect to their position and whether they are wearing helmets or not, i.e., distinguishing rider positions as in Fig. 2 and determining whether a helmet is worn or not for each rider.

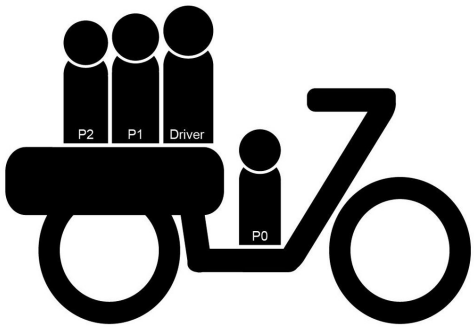


Figure 2. Illustration of positions of motorcycles, riders and passengers.

The training and testing datasets both contain 100 video clips. Recorded at 10 fps and 1920×1080 resolution, each video is 20 seconds duration. The ground-truth bounding boxes of motorcycles, drivers and passengers are all annotated individually, the height and width of which are all greater than 40 pixels. Specifically, a motorcycle could carry no more than 4 riders. In other word, there are 9 categories annotated, i.e. (0) *motorbike*, (1) *DHelmet* (Driver with helmet), (2) *DNoHelmet* (Driver without helmet), (3) *P1Helmet* (Passenger 1 with helmet), (4) *P1NoHelmet* (Passenger 1 without helmet), (5) *P2Helmet* (Passenger 2 with helmet), (6) *P2NoHelmet* (Passenger 2 without helmet), (7) *P0Helmet* (Child sitting in front of the driver of the motorcycle with helmet), (8) *P0NoHelmet* (Child sitting in front of the driver of the motorcycle without helmet). Some examples of annotations are visualized in Fig. 3. As shown in the example images, severe occlusion of motorcycles and


Figure 3. Illustration of some annotations. Parts of motorbike or riders are occluded.

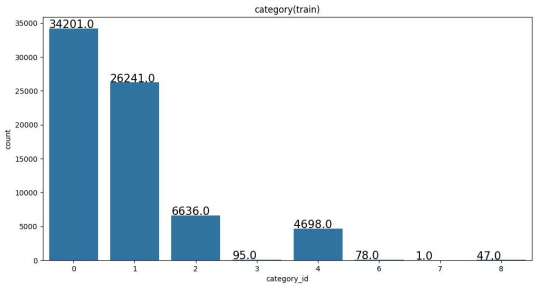


Figure 4. Statistics of label distributions of the training dataset.

passengers brings a great challenge to accurate model predictions.

According to the statistics of label distributions of the training dataset, severe data imbalance can be observed from Fig. 4. Particularly, the number of *P1Helmet*(3), *P2Helmet*(5), *P2NoHelmet*(6) and *P0Helmet*(7), *P0NoHelmet*(8) instances are all less than 100. Actually, the only annotation for *P0Helmet*(7) is an incorrect annotation (see Fig. 5). There are even no annotations of *P2Helmet*(5) in the training dataset.

The metric used to evaluate the performance of the model is the mean Average Precision (mAP) across all frames in the test videos. mAP measures the mean of average precision (the area under the Precision-Recall curve) over all the object classes, as defined in PASCAL VOC 2012 competition.

### 4.2. Implementation Details

After numerous comparative experiments, we choose DETA[6] and CO-DETR[14] with Swin-Large backbones as our base detectors.

Our algorithms are implemented with pytorch framework. At training phase, we first split 20% of the training dataset for validation and select an optimal training configuration using 80% of the training dataset. Then we use the full training dataset for training. We train our models with batchsize 4 on 4 Tesla V100 GPUs. We set init learning rate as 5e-5 and weight decay as 1e-4. In the feedforward layers of transformer block, the embedding dimension of input is 2048 and number of attention heads of multi-heading attention is 8. 900 object queries are used for all the transformer-based models. For data augmentations, the image scale during training phase is randomly selected from $[720, 768, 816, 864, 912, 960, 1008, 1056, 1104, 1152, 1200]$ pixels for the shorter side, and the longer side does not exceed 2000 pixels. We also do random horizontal flip in training phase. The shorter side is resized to 1200 pixels at testing phase. The model is finetuned from a pretrained weight that has been pretrained on the Objects365 dataset.

At testing phase, We extract frames from the 100 videos of the testing dataset. Then we get the test results from the two transformer-based models respectively. Finally, NMS is used for fusion to get the final results. It is worth mentioning that although TTA is mentioned above, it is not used in our final submissions due to time constraints.

### 4.3. Experimental Results

We achieved a 0.4824 mAP using a simple NMS fusion with two transformer-based models, CO-DETR and DETA, which ranked the second place on the public leaderboard of Track5 of the AI City Challenge 2024, as shown in Table 1. We also observed that a single CO-DETR model achieved a 0.4788 mAP with only 8 epochs of training.

### 4.4. Ablation study

In this section, we compare several object detection models with different configurations for motorcycle helmet violation detection. Results of different detection models with 80% training dataset at 1333×800 resolution are shown in Table 2. The results are evaluated on the remaining 20%


Figure 5. an annotation of P0Helmet in the training dataset.

| Rank | Team ID | Team Name | Score |
|---|---|---|---|
| 1 | 99 | Helios | 0.4860 |
| **2** | **76** | **CMSR_PANDA(Ours)** | **0.4824** |
| 3 | 9 | VNPT AI | 0.4792 |
| 4 | 155 | TeleAI | 0.4675 |
| 5 | 5 | SKKU Automation Lab | 0.4644 |
| 6 | 228 | DIDANO1 | 0.4621 |
| 7 | 57 | BUPT_MCPRL | 0.3940 |
| 8 | 247 | CHTTL_IOTLAB | 0.3650 |
| 9 | 154 | aio_tts | 0.3547 |
| 10 | 90 | Graph@FIT+Comenius | 0.3465 |

Table 1. Top 10 Leaderboard of Track5 in the AI City Challenge 2024

training dataset. As a classic two-stage detector, Faster-RCNN scores 0.244 and 0.257 with a ResNet-50 and a ResNet-101 backbone respectively. Cascade-RCNN scores 0.263 with a ResNet-101 backbone. DINO with different scales can get a score of 0.281 and 0.302 with a ResNet-50 and a Swin-L backbone respectively, which indicates the strong ability of Swin-transformer backbone for feature extraction. At 1333×800 input resolution with 80% training dataset, DINO achieves the best score.

Results of several transformer-based models with a Swin-L backbone are shown in Table 3. The results are evaluated on the full test dataset. For single model performance using full training dataset for training, CO-DETR model at 2048×1280 resolution achieved the best score of 0.4788 with only 8 epochs of training.

Different fusion strategies are explored. During inference, the prediction results of a Co-DETR[14] model and four DETA[6] models (testing at epoch 8,12,16,20) are merged with WBF[9] and NMS[4] respectively. The results are shown in Table 4. For our final submission (scoring 0.4824), two transformer-based models, DETA[6] (scoring 0.3858 on the testing dataset) and Co-DETR[14] (scoring 0.4788 on the testing dataset), are utilized with a NMS strategy for model ensemble. If we use our best DETA [6] model (scoring 0.4504 on the testing dataset) for model ensemble, we can achieve better results on the public leaderboard of the challenge.

## 5. Conclusion

Automatic detection of motorcyclists without wearing helmets based on video surveillance is a critical task to enforce strict regulatory traffic safety measures. In this paper, a simple yet effective method is proposed to detect violation of helmet rule for motorcyclists. We have explored several state-of-the-art object detection models and chosen two transformer-based models, DETA and Co-DETR. Finally, NMS and WBF strategies are utilized to merge the prediction results for performance improvement. Our method achieves a 0.4824 mAP on 2024 AI City Challenge Track 5 dataset and ranks the second place on the leaderboard without using any tracking or post-processing strategies, which demonstrates the effectiveness of our method.

| Method | Backbone | Epoch | Score |
|---|---|---|---|
| Faster-RCNN | ResNet-50 | 12 | 0.244 |
| Faster-RCNN | ResNet-101 | 12 | 0.257 |
| Cascade-RCNN | ResNet-101 | 24 | 0.263 |
| DINO-4scale | ResNet-50 | 12 | 0.281 |
| DDQ-detr-4scale | Swin-L | 30 | 0.291 |
| **DINO-5scale** | **Swin-L** | **12** | **0.302** |

Table 2. Results of different detection models with 80% training dataset at 1333×800 resolution. -4scale and -5scale indicate features at 4 and 5 scales from the backbone network are used. The results are evaluated on the remaining 20% training dataset.

| Method | Resolution | Epoch | Score |
|---|---|---|---|
| DINO | 1333×800 | 12 | 0.4029 |
| DINO* | 1333×800 | 12 | 0.4162 |
| H-DINO* | 1333×800 | 12 | 0.3147 |
| DINO* | 1666×1000 | 12 | 0.4350 |
| DETA | 2000×1200 | 2 | 0.3858 |
| DETA* | 2000×1200 | 20 | 0.4504 |
| **Co-DETR*** | **2048×1280** | **8** | **0.4788** |

Table 3. Results of transformer-based models based on a Swin-L backbone with different configurations. * indicates training with full training dataset. The results are evaluated on the full test dataset.

| Models | Fusion strategy | Score |
|---|---|---|
| DETAs | WBF | 0.4719 |
| DETAs | NMS | 0.4633 |
| Co-DETR+DETA | WBF | 0.4625 |
| **Co-DETR+DETA** | **NMS** | **0.4824** |

Table 4. Results of different fusion strategies and models.

## References


[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *Arxiv*, abs/2004.10934, 2020. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 213–229, 2020. 2, 3

[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong

Figure 6. Visualization of results from the testing dataset.


Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 764–773, 2017. 2

[4] Ross B. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society, 2015. 2, 3, 5

[5] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587. IEEE Computer Society, 2014. 2

[6] Jeffrey Ouyang-Zhang, Jang Hyun Cho, Xingyi Zhou, and Philipp Krähenbühl. NMS strikes back. *Axriv*, abs/2212.06137, 2022. 2, 3, 4, 5

[7] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788, 2016. 2

[8] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015. 2

[9] Roman A. Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image Vis. Comput.*, 107:104117, 2021. 3, 5

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 2

[11] Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh Chakraborty, Sanjita Prajapati, Quan Kong, Norimasa Kobori, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Ganzorig Batnasan, Fady Alnajjar, Ping-Yang Chen, Jun-Wei Hsieh, Xunlei Wu, Sameer Satish Pusegaonkar, Yizhou Wang, Sujit Biswas, and Rama Chellappa. The 8th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[12] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 2

[13] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 2

[14] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 6725–6735, 2023. 2, 3, 4, 5