Robust Motorcycle Helmet Detection in Real-World Scenarios: Using Co-DETR and Minority Class Enhancement

Hao Vo [1,2], Sieu Tran [1,2], Duc Minh Nguyen [1,2]
Thua Nguyen [1,2], Tien Do [1,2], Duy-Dinh Le [1,2], Thanh Duc Ngo [1,2]

[1] University of Information Technology, VNU-HCM, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
{21520832, 21520097, 21520730}@gm.uit.edu.vn
{thuann, tiendv, duyld, thanhnd}@uit.edu.vn

Abstract

Motorcycle helmet detection is a crucial task in intelligent traffic systems (ITS), as it enhances traffic safety consciousness and guides individuals towards legal compliance. Numerous challenges are tied to this problem, particularly regarding data from the real world. In addition to requiring resilience to environmental fluctuations, such as diverse camera angles and lighting conditions, the solution must also address the problem of unbalanced data distribution across object classes. This study presents a system that utilizes Co-DETR to address the difficulties of dealing with changing perspectives on real-world data. Additionally, we propose to use the Minority Optimizer and the Virtual Expander to enhance the accuracy of rare classes in imbalanced data. With a mean average precision (mAP) of 0.4860, our method achieved Rank 1 in the AI City Challenge 2024 Track 5 competition, demonstrating its high effectiveness.

1. Introduction

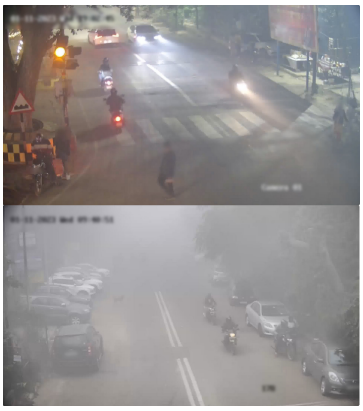
In developing countries, motorcycles stand as a standard mode of transportation due to their cost-effectiveness and maneuverability. However, the minimal protection offered by motorcycles heightens the risk of accidents for riders. To counteract this risk, authorities have instituted laws that mandate helmet usage to reduce the severity of injuries in crashes. The World Health Organization highlights the essential role of helmets in significantly lowering the risk of head injuries and fatalities. However, despite these regulations, ensuring adherence poses a challenge, especially in developing areas, underlining the necessity for innovative solutions.

Acknowledging the abovementioned need, the AI City Challenge introduces Track 5: Detecting Violation of Helmet Rule for Motorcyclists [14]. This initiative emphasizes the importance of helmets in safeguarding motorcycle riders, who are particularly vulnerable due to the limited protection their vehicles provide. Intending to enforce traffic safety regulations more rigorously, Track 5 focuses on automatically detecting motorcyclists flouting helmet laws. Leveraging advancements in Computer Vision and Deep Learning, previous studies have demonstrated the potential of automated detection systems in monitoring helmet rule violations efficiently. Such systems promise to enhance road safety and reduce the workload of law enforcement agencies, marking a significant step towards mitigating the number of fatalities associated with motorcycle accidents.

Detecting motorcycle riders faces numerous challenges, particularly in developing countries with crowded roads and diverse traffic conditions. The high vehicle density often results in cluttered scenes with overlapping objects, making it difficult for detection models to discern individual riders amidst the chaos, see Figure 1a. Furthermore, environmental factors such as lighting and weather conditions significantly affect detection accuracy, as shown in Figure 1b. Glare, nighttime conditions, and fog can obscure details in the scene, making it challenging for the model to identify riders reliably. These conditions introduce variability and unpredictability, requiring robust algorithms to adapt to diverse visual environments. Moreover, security cameras positioned at elevated locations with various angles reduce video resolution and diversify objects' proportions and sizes, as demonstrated in Figure 1c.

In tackling the multifaceted challenges of motorcycle rider detection, we employed the Co-DETR [16] model, which emerges as a transformative solution. Its unique approach to collaborative label assignment effectively manages class imbalances within datasets, which is crucial for

(a) Overlapping objects



(b) Glare, nighttime and fog

(c) Various angles

Figure 1. Demonstration of the complex traffic conditions of the datasets

scenarios where certain classes, like instances of motorcycle violations, are underrepresented. By permitting multiple box candidates to correspond with the same ground-truth box during training, Co-DETR enriches its learning process with diverse and informative samples, enhancing its ability to discern individual riders amidst cluttered scenes with overlapping objects. Furthermore, its adaptability to incorporate images captured at various scales during inference addresses environmental factors such as glare, nighttime conditions, and fog, significantly impacting detection accuracy. These enhancements enable Co-DETR to navigate the variability and unpredictability of diverse visual environments, ensuring robust performance in detecting motorcycle riders. Additionally, its capability to accommodate security cameras positioned at elevated locations with varying angles underscores its versatility in handling challenges associated with reduced video resolution and diversified object proportions and sizes. Thus, Co-DETR is a pivotal tool in bolstering motorcycle rider detection amidst the complex and dynamic landscapes of developing countries' traffic environments. Moreover, we employed Weighted Boxes Fusion [12] to integrate results generated by inference models across varying image sizes. This enhanced the model's stability and adaptability to diverse conditions and object proportions. Additionally, we utilized this technique to combine results from models across different checkpoints, aiming to mitigate overfitting.

In the dataset provided by the AI City Challenge 2024, we have observed a severe imbalance among classes, especially concerning classes 2 and 0 for passengers. No examples are available for class 6, as demonstrated in Figure 2. Additionally, the quality, camera angles, and weather conditions cause continuous variations within the same object category throughout the video, significantly impacting the model's performance. We propose two algorithms, the Minority Optimizer and the Virtual Expander, to address these issues.

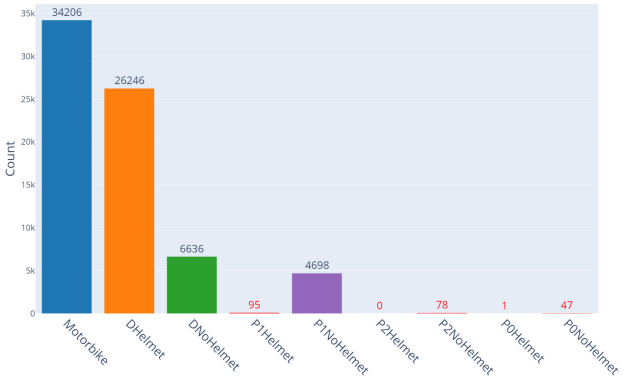


Figure 2. Visualization of classes distribution.

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

- We introduce a framework to tackle the problem of detecting violations of helmet rules for motorcyclists, centered around the Co-DETR model, which includes training strategies aimed at aiding the model in effectively adapting to imbalanced data and various environmental conditions.
- We propose to use two algorithms, the Minority Optimizer and the Virtual Expander, to significantly improve the recall of classes with limited data and prone to confusion without significantly affecting precision, thereby increasing mAP.

The results of the AI City Challenge 2024 Track 5 final leaderboard results indicate that the proposed framework achieves first place with a score of 0.4860.

2. Related Works

2.1. Pipeline

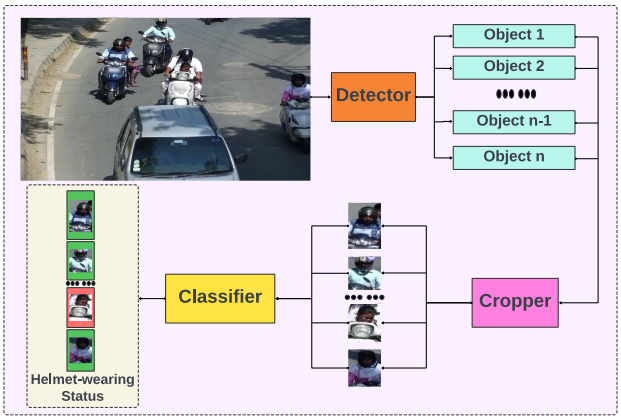


Figure 3. Multiple stages pipeline. These pipelines often include two modules: Detector and Classifier

In the study by Espinosa et al. [7], various methodologies for motorcycle detection are explored comprehensively. Traditional approaches, as discussed in the works of Silva et al. [11], Dahiya et al. [4], and Talaulikar et al. [13], typically follow a sequential pipeline. Initially, these methods utilize motion segmentation techniques such as optical flow and background subtraction to identify moving objects. Subsequently, handcrafted feature descriptors like local binary pattern (LBP) and histogram of oriented gradient (HOG) are employed to extract features specific to motorcycles, followed by the application of binary classifiers like support vector machines (SVM) and K-nearest neighbors (KNN) for classification.

However, these traditional approaches suffer from limitations. The multi-stage operation often hinders real-time processing capability. Moreover, accurately discerning helmet usage becomes challenging, especially in scenarios involving multiple motorcycle riders, particularly when one rider obscures another wearing a helmet. External factors such as road congestion and camera instability can also significantly impede the effectiveness of motion segmentation-based methodologies. This paper proposes leveraging modern object detection methods to simplify the processing pipeline while maintaining or improving performance.

2.2. Object Detection

Object detection is a pivotal task within computer vision, necessitating the localization and classification of objects into their respective categories. Prior to the emergence of Transformer models [2, 5, 9] in this domain, the RCNN family [8] and YOLO family [6], which utilize CNN-based methods, have been instrumental in advancing object detection. The RCNN series, including Fast RCNN and Faster RCNN, incrementally improved the efficiency and accuracy of object detection by enhancing feature extraction and streamlining the detection process. On the other hand, the YOLO (You Only Look Once) series revolutionized the field by enabling real-time object detection, emphasizing speed and efficiency without significantly compromising accuracy. These families laid the groundwork for understanding and processing visual data, setting high standards for accuracy and efficiency in object detection tasks. Additionally, the advent of Transformer models has introduced new possibilities, with some object detection systems adopting Transformer architectures to achieve notable effectiveness. DETR (Detection Transformer) [2] emerged as the first model to integrate Transformer architecture, aiming to reduce the reliance on numerous manually designed components in object detection, showing promising performance. However, DETR encounters challenges like slow convergence and limited spatial resolution of features due to Transformer attention modules' constraints in processing image features. To mitigate these issues, Deformable DETR [15] was proposed, significantly enhancing DETR's efficiency, particularly for small objects and in datasets with imbalance, by focusing its attention modules on a select group of crucial sampling points.

Furthermore, another research on Co-DETR has identified a critical limitation in DETR: assigning too few queries as positive samples, resulting from one-to-one set matching, leads to insufficient supervision of the encoder's output. This inadequacy detrimentally affects the learning of discriminative features by the encoder and, inversely, the learning of attention by the decoder. Addressing this, Co-DETR introduces multiple parallel auxiliary heads and tailored positive queries, which improve the encoder's learning capacity and the overall training efficiency without adding extra parameters or computational demands. This approach also obviates the need for manually crafted non-maximum suppression at the inference stage, presenting a viable and efficient alternative for object detection endeavors. Simultaneously, Co-DETR has also achieved 66.0% AP on COCO test-dev and 67.9% AP on LVIS val, outperforming previous methods by clear margins with much fewer model sizes. This paper utilizes Co-DETR as a principal strategy to tackle the challenge of detecting helmet rule violations among motorcyclists, offering a sophisticated yet accessible approach to enhance road safety through cutting-edge computer vision technology.

2.3. Ensemble

Object detection methods in computer vision often yield numerous predictions within a single frame, aiming to maximize object identification while minimizing omissions.

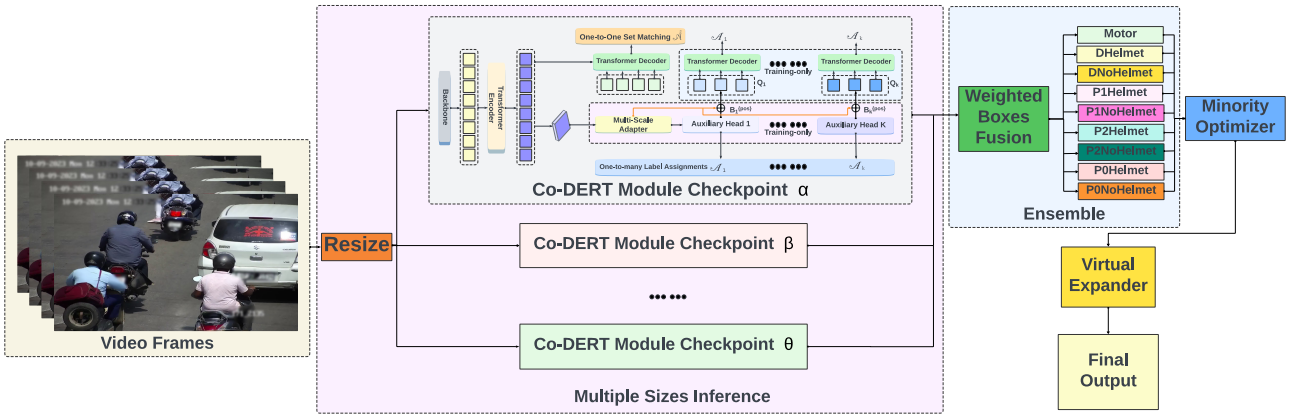


Figure 4. Framework Architecture Overview: This framework integrates four main components: Object Detection, Ensemble, Minority Optimizer, and Virtual Expander Algorithms. It begins with the Co-DETR model analyzing video frames across different conditions. Next, the Ensemble Module combines results from several inference rounds, with the Minority Optimizer Algorithm enhancing recall for infrequent classes. The Virtual Expander Algorithm then reduces False Negatives by generating extra bounding boxes at a specific confidence threshold, thus fine-tuning Precision.

However, this approach generates redundant bounding boxes, potentially undermining the model's accuracy. To address redundancy, Non-Maximum Suppression (NMS) [10] is commonly employed. NMS operates by sorting detection boxes based on their confidence scores, selecting the box with the highest confidence score, and filtering out others with significant overlap. Although effective, NMS relies on fixed thresholds and may need help to seamlessly integrate results from multiple models due to its reliance solely on confidence scores.

In contrast, ensemble methods, instrumental in non-real-time applications, amalgamate predictions from diverse models to enhance overall accuracy. A novel approach, Weighted Boxes Fusion (WBF), is proposed to address the challenges posed by traditional NMS techniques. Unlike NMS, which merely discards redundant predictions, WBF leverages confidence scores from all bounding boxes to construct average boxes, thereby substantially improving the quality of combined predictions. In this study, we employ WBF to merge results from multiple models, enabling adaptation to varying conditions such as diverse camera angles and weather conditions.

3. System Architecture

3.1. Overview

The general architecture of our framework, as shown in Figure 4, revolves around four principal components: the Object Detection Module, the Ensemble Module, the Minority Optimizer Algorithm, and the Virtual Expander Algorithm. The Co-DETR model processes video frames, accommodating various checkpoints and image sizes. Following this, the Ensemble Module aggregates the outcomes from multiple inference iterations and then uses the Minority Optimizer Algorithm to boost the recall for rare classes. Lastly, the Virtual Expander Algorithm actively minimizes the volume of False Negative instances. It achieves this by creating additional bounding boxes at a carefully determined confidence threshold, effectively controlling the impact on overall Precision.

3.2. Detection Module

In this study, we utilized the Co-DETR model as the primary tool for object detection. This decision was motivated by its status as a cutting-edge object detection model and its suitability for addressing the specific challenges inherent in our problem domain.

Our initial concern was the pronounced class imbalance within the dataset demonstrated in Figure 2 Ċo-DETR introduces a novel collaborative one-to-many label assignment approach to confront this issue. This strategy effectively manages hyper-imbalanced data by permitting multiple box candidates to be associated with the same ground-truth box during training, thereby enriching the model's learning process with more varied and informative samples. This approach proves particularly beneficial in scenarios where certain classes, such as instances of motorcycle violations, are underrepresented in the dataset.

Moreover, our approach of combining multiple checkpoints and image sizes during inference extends beyond addressing inherent dataset discrepancies; it also offers advantages in coping with challenging weather conditions. Envi-

ronmental factors like rain, fog, and other phenomena can substantially alter the visual appearance of objects, impeding their detectability. Our method enhances the model's resilience to such environmental changes by incorporating images captured at various scales. This diversity in input scales enables the Co-DETR model to generalize more effectively across diverse visibility conditions, ensuring robust object detection irrespective of weather-related visual impairments. This adaptation is crucial for applications necessitating high accuracy in real-world, variable conditions, underscoring the model's enhanced capability to maintain performance even in less-than-ideal environmental scenarios.

3.3. Ensemble Module

3.3.1 Weighted Box Fusion

In tackling the multifaceted challenges inherent in helmet violation detection within diverse and complex environments, our ensemble module is pivotal in refining detection accuracy while effectively mitigating false negatives and positives. We employed the Weighted Box Fusion (WBF) method to address this issue. Unlike conventional techniques such as NMS and soft-NMS [1], which tend to discard specific predictions, WBF leverages the confidence scores associated with all proposed bounding boxes to construct average boxes. This approach markedly enhances the quality of the combined predicted rectangles. To elaborate, when presented with a set B of predicted bounding boxes within a frame

$$\begin{aligned} B &= \{box_1, box_2, \ldots, box_n\} \\ conf_i &\geq conf_j \quad \forall i < j \end{aligned} \tag{1}$$

the WBF method organizes these boxes into clusters, forming clusters denoted as L.

$$L = \{cluster_1, cluster_2, \ldots, cluster_m\} \tag{2}$$

where $cluster_i$ is a list containing all bounding boxes in that cluster. Each cluster is then represented by bounding boxes contained in F.

$$F = \{box_1, box_2, \ldots, box_m\} \tag{3}$$

where box_i is represented for $cluster_i$.
For each cluster, the representative bounding box $\{x1, y1, x2, y2, c\}$ is calculated using the following formula:

$$\begin{aligned} C &= \frac{\sum_{i=1}^{T} C_i}{T} \\ X_{1,2} &= \frac{\sum_{i=1}^{T} C_i * X_{1,2_i}}{\sum_{i=1}^{T} C_i} \\ Y_{1,2} &= \frac{\sum_{i=1}^{T} C_i * Y_{1,2_i}}{\sum_{i=1}^{T} C_i} \end{aligned} \tag{4}$$

where T is the number of all bounding boxes in that cluster.

After processing all boxes in B, the confidence scores in the F list are rescaled by multiplying them by the number of boxes in a cluster and then dividing by the number of inference sizes. This rescaling strategy accounts for variations in the number of predictions across clusters. If a cluster contains a low number of boxes, indicating limited prediction support from models, the confidence scores need adjustment to reflect this uncertainty. There are two methods to achieve this rescaling:

$$C = C * \frac{\min(T, N)}{N}, \tag{5}$$

or

$$C = C * \frac{T}{N} \tag{6}$$

This comprehensive approach ensures robustness and accuracy in synthesizing detection results, particularly in scenarios where bounding boxes overlap extensively. An illustrative example showcasing the difference in output results between WBF and NMS is presented in Figure 5.

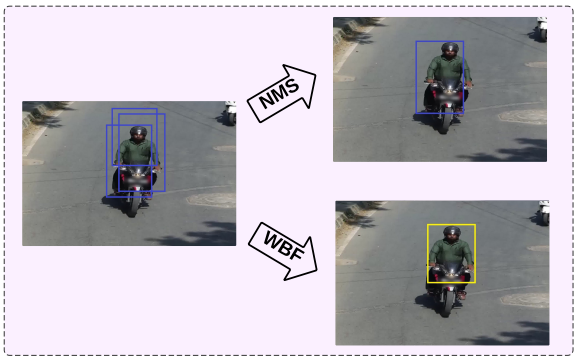


Figure 5. Example for Weighted Boxes Fusion and NMS

3.3.2 Models Ensembling.

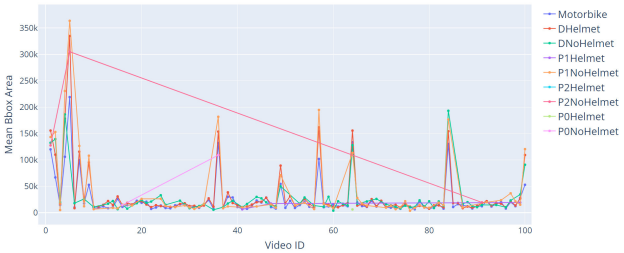


Figure 6. Distributions of area of bounding box in each videos

Multiple Scales Adapting We encountered substantial variations in camera angles and zoom levels, leading to considerable disparities in the sizes of bounding boxes assigned to objects of the same class, demonstrated in Figure 6. To mitigate this challenge, we adopted a diverse training regimen for Co-DETR, encompassing a wide range of image scales. Subsequently, we integrated multiple image sizes during the inference stage to mitigate these discrepancies. The results of this experiment demonstrated a marked improvement in the model's performance.

Table 1. Class-wise Average Precision (AP) and Mean AP (MAP) on the validation dataset

Class	AP
Motorbike	0.9849
DHelmet	0.9895
DNoHelmet	0.9915
P1Helmet	0.9958
P1NoHelmet	0.9780
P2NoHelmet	0.9455
P0NoHelmet	1.0
Mean Average Precision	0.9836

Overfit Cutting After training and evaluating the model on the validation dataset, we achieved a significantly high mean Average Precision (mAP) score of 0.9836, as detailed in Table 1. Subsequently, upon conducting thorough analysis and observations, we identified instances where models at earlier epochs exhibited superior accuracy compared to those at later epochs, as evidenced in Figure 7. To enhance model stability and mitigate overfitting on the test dataset, we adopted the WBF technique elucidated in Section 3.3.1. This entailed amalgamating outcomes from models across different epochs.

3.4. Minority Optimizer

In this module, we focus on mitigating the occurrence of False Positive Samples. To achieve this, we identify an optimal confidence threshold that effectively filters out bounding boxes falling below it. However, a significant challenge arises due to the pronounced imbalance within our dataset (Figure 2). Consequently, our strategy revolves around prioritizing thresholds based on classes that are deemed rare, ensuring that Recall remains robust for these specific classes.

Our approach begins by pinpointing classes that exhibit significant imbalances, characterized by fewer samples, amounting to less than α compared to the class with the highest sample count. Subsequently, we embark on the quest to determine a confidence threshold that is sufficiently diminutive to retain the integrity of these classes.

Figure 7. Example difference outcomes of the checkpoint model at epochs 10 and 15, corresponding to the images above and below, respectively. The areas highlighted in blue signify the disparity between the results; specifically, the checkpoint at epoch 10 detects more 2 true positive samples than the checkpoint at epoch 15.

Nevertheless, critical consideration surfaces due to the rescaling confidence scores by WBF. This rescaling may result in extremely diminutive confidence scores. To avert the potential pitfall of minimal thresholds, which could inadvertently inflate the count of False Positive boxes, we restrict by setting the minimum confident threshold to ρ.
To gain a deeper understanding of the algorithm's workings, the step-by-step process is outlined in Algorithm 1.

3.5. Virtual Expander

Figure 8. Demonstration of inconsistent of object's class

Based on our observations, one of the recurring issues encountered during the inference process is the inconsis-

Algorithm 1 Minority Optimizer

procedure MINORITYOPTIMIZER(ρ, classes)
  $n_{\text{max\_class}} \leftarrow$ number of samples in max_class
  $mean\_samples \leftarrow total\_samples/9$
  $\alpha \leftarrow n_{\text{max\_class}}/mean\_samples$
  $rare\_classes \leftarrow []$
  for each $class$ **in** $classes$ **do**
    $n_{\text{class}} \leftarrow$ number of samples in $class$
    if $n_{\text{class}} < n_{\text{max\_class}} \times \alpha$ **then**
      append $class$ to $rare\_classes$
    end if
  end for
  $min\_thresh \leftarrow 1.0$
  for each $class$ **in** $rare\_classes$ **do**
    for each $sample$ **in** $class$ samples **do**
      if confident of $sample < min\_thresh$ **then**
        $min\_thresh \leftarrow$ confident of $sample$
      end if
    end for
  end for
  return $\max(min\_thresh, \rho)$
end procedure

Figure 9. Demonstration of noise in cropping image.

tency between the presence and absence of helmets worn by the same object in the video, demonstrated in Figure 8. To address this issue, we experimented with additional training of classifier models [3] and employed a voting strategy to mitigate the problem. However, due to numerous noise factors such as multiple individuals within a single bounding box and fog during the cropping of bounding boxes, demonstrated in Figure 9, which serve as inputs to the classifier model, we did not achieve the expected results. Consequently, we propose a trade-off strategy between precision and recall to generate "virtual" bounding boxes with appropriate confident scores to optimize a portion of the recall score in cases where the detector fails to classify the object class.

4. Experiments

4.1. Dataset

The AI City Challenge 2024 dataset comprises 100 training videos, each lasting 20 seconds and recorded at a frame rate of 10 frames per second (fps). These videos have a resolution of 1920×1080 pixels. Annotations within each video consist of ground truth bounding boxes detailing motorcycles, their riders, and their helmet-wearing status. Each annotated frame includes bounding box annotations for motorcycles and up to four riders per motorcycle (i.e., driver, passenger 1, passenger 2, passenger 0). Each rider is individually identified based on whether or not they are wearing a helmet. The challenge's objective is to develop an algorithm that accurately identifies motorcycles and their riders, discerning their helmet status.

4.2. Implementation Details

Table 2. Overview of Image Scales Employed for Co-DETR Model Training and Inference

Phase	Image Scales
Training	(480x2048), (512x2048), (544x2048), (576x2048), (608x2048), (640x2048), (672x2048), (704x2048), (736x2048), (768x2048), (800x2048), (832x2048), (864x2048), (896x2048), (928x2048), (960x2048), (992x2048), (1024x2048), (1056x2048), (1088x2048), (1120x2048), (1152x2048), (1184x2048), (1216x2048), (1248x2048), (1280x2048), (1312x2048), (1344x2048), (1376x2048), (1408x2048), (1440x2048), (1472x2048), (1504x2048), (1536x2048)
Inference	(640x640), (1280x1280), (2048x1280)

We selected the Co-DETR detector with Co-DINO pretraining and a Swin Transformer Large (Swin-L) backbone due to its previously demonstrated high performance in related tasks. The training phase spanned 16 epochs, exposing the model to images resized to a predefined set of dimensions to ensure learning across various scales and aspect ratios, as detailed in Table 2. At inference time, specif-

ically on epoch 11, the team evaluated the model's performance across three resolutions: 640x640, 1280x1280, and 2048x1280, and on epoch 15 across two resolutions: 640x640 and 1280x1280.

To further refine the detection accuracy and reliability, we employ an ensemble method utilizing Weighted Boxes Fusion (WBF) with an Intersection Over Union (IOU) threshold of 0.7. This advanced ensemble technique allows for the integration of detection results across multiple scales, effectively reducing false positives and enhancing the precision of the bounding boxes.

4.3. Experiments Results

4.3.1 Evaluation Metric

For the AI City Challenge 2024 Track 5, teams utilized the mean Average Precision (mAP) to evaluate performance, calculating it across all frames within the test videos. The mAP measures the mean of the average precision values from the Precision-Recall curve for each object class, following the PASCAL VOC 2012 competition's calculation methodology. This metric assesses a team's ability to detect objects across various categories within the test dataset comprehensively.

4.3.2 Ablation Study

The team presents the contribution of each module to the final results in Table 3. Starting with a CO-DERT model trained to detect all nine specified classes, the team observed substantial performance improvement, from an initial metric of 0.4104 to 0.4365, using multiple sizes inference and the Weighted Boxes Fusion (WBF) ensemble technique. Adding the Minority Optimizer module further boosted performance to 0.4830, placing the team at the top of the challenge leaderboard. The implementation of the Virtual Expander module provided a slight improvement in the results.

Table 3. Analysis of Module Contributions to Mean Average Precision (mAP) Enhancement in Detecting Violation of Helmet Rule for Motorcyclists problem

Module	**mAP**
CO-DERT Model (baseline)	0.4104
+ WBF	0.4365
+ Minority Optimizer	0.4830
+ Virtual Expander	0.4860

4.3.3 Comparison with other teams.

We evaluated our solution using the Track 5 evaluation system. As indicated in Table 4, our solution achieved a mean Average Precision (mAP) of 0.4860, securing the **first place** on the public leaderboard of the challenge, surpassing over 40 participating teams.

Table 4. The challenge leaderboard summary. Our proposed solution achieved results that surpassed those of other teams. (VE is Virtual Expander Module)

Team ID	Team Name	Score (↑)
155	TeleAI	0.4675
9	VNPT AI	0.4792
76	CMSR_PANDA	0.4824
99	Helios (**Ours** - Without VE)	**0.4830**
99	Helios (**Ours** - With VE)	**0.4860**

4.3.4 Discussion

With the inclusion of the Virtual Expander module in our system, which has demonstrated exceptional performance in securing first place in the AI City Challenge 2024 Track 5 competition, it is imperative to acknowledge certain limitations. Despite our success, the integration of the Virtual Expander module does introduce additional computational demands, as evidenced by its marginal improvement in mean Average Precision (mAP) from 0.4830 to 0.4860. This increase in computational requirements could potentially hinder the scalability and cost-effectiveness of our solution, particularly in resource-constrained environments or scenarios where efficient utilization of computational resources is paramount. Furthermore, although our system remains at the top of the leaderboard both with and without the Virtual Expander module, its relatively modest improvement suggests that further optimization may be necessary to justify its inclusion in terms of resource expenditure.

5. Conclusion

In summary, deploying Co-DETR alongside the novel Minority Optimizer and Virtual Expander algorithm marks a significant leap forward in intelligent traffic management, specifically in identifying helmetless motorcyclists. Our system adeptly navigates the complexities of varying environmental conditions and the challenge of uneven data distribution, establishing a new benchmark for object detection in urban settings. Validated by our leading performance in the AI City Challenge 2024 Track 5.

Acknowledgment

This research is funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number D1-2024-04

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 5

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

[3] Po-Yung Chou, Cheng-Hung Lin, and Wen-Chung Kao. A novel plug-in module for fine-grained visual classification. *arXiv preprint arXiv:2202.03822*, 2022. 7

[4] Kunal Dahiya, Dinesh Singh, and C Krishna Mohan. Automatic detection of bike-riders without helmet using surveillance videos in real-time. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3046–3051. IEEE, 2016. 3

[5] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1601–1610, 2021. 3

[6] Juan Du. Understanding of object detection based on cnn family and yolo. In *Journal of Physics: Conference Series*, page 012029. IOP Publishing, 2018. 3

[7] Jorge E Espinosa, Sergio A Velastín, and John W Branch. Detection of motorcycles in urban traffic using video analysis: A review. *IEEE Transactions on Intelligent Transportation Systems*, 22(10):6115–6130, 2020. 3

[8] O Hmidani and EM Ismaili Alaoui. A comprehensive survey of the r-cnn family for object detection. In *2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet)*, pages 1–6. IEEE, 2022. 3

[9] Teli Ma, Mingyuan Mao, Honghui Zheng, Peng Gao, Xiaodi Wang, Shumin Han, Errui Ding, Baochang Zhang, and David Doermann. Oriented object detection with transformer. *arXiv preprint arXiv:2106.03146*, 2021. 3

[10] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, pages 850–855. IEEE, 2006. 4

[11] Romuere Silva, Kelson Aires, Thiago Santos, Kalyf Abdala, Rodrigo Veras, and Andre Soares. Automatic detection of motorcyclists without helmet. In *2013 XXXIX Latin american computing conference (CLEI)*, pages 1–7. IEEE, 2013. 3

[12] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107: 104117, 2021. 2

[13] Abhijeet S Talaulikar, Sanjay Sanathanan, and Chirag N Modi. An enhanced approach for detecting helmet on motorcyclists using image processing and machine learning techniques. In *Advanced Computing and Communication Technologies: Proceedings of the 11th ICACCT 2018*, pages 109–119. Springer, 2019. 3

[14] 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. 1

[15] 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

[16] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023. 1