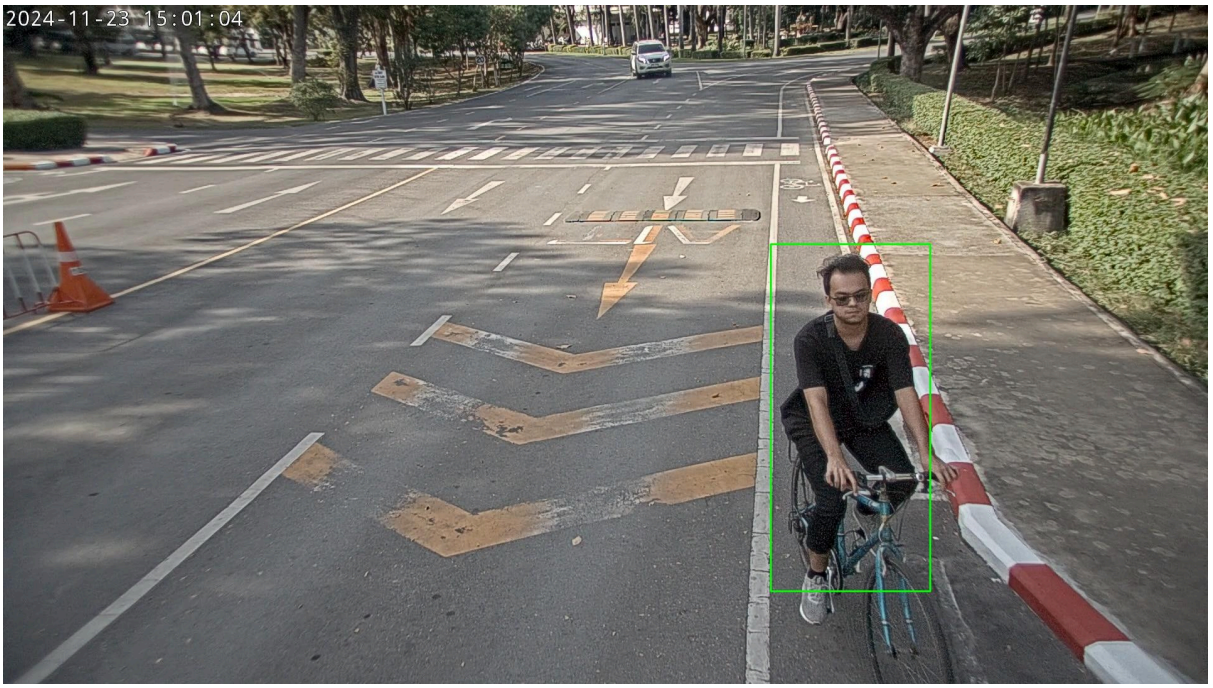
Report

In this experiment, I aimed to evaluate and compare the performance of YOLOv5, YOLOv8, and YOLOv11 models on an object detection task using a provided dataset containing 1,100 images. Initially, the dataset was utilized without addressing class imbalance, and each model was trained for 50 epochs with a batch size of 16. The primary metric for evaluation was the mean Average Precision (mAP) score, which guided the selection of the best-performing model among the three. After this phase, it became evident that class imbalance was a significant limitation in improving the models' overall performance. To partially address this issue, I augmented the dataset with additional images, increasing its size to 2,309 images. However, the augmentation was limited and did not completely resolve the class imbalance. Considering advancements in model architecture, I decided to discontinue experiments with YOLOv5, as it is generally considered outdated, and instead focused on training YOLOv8 and YOLOv11 models, which offer superior performance and efficiency.

For the expanded dataset, I trained YOLOv8 and YOLOv11 models in two variations each, with batch sizes of 16 and 32, over 1,000 epochs. These experiments were computationally intensive, fully utilizing an NVIDIA GPU with 11GB of VRAM. In parallel, recognizing the importance of a diverse and balanced dataset, I conducted additional data collection specifically for the "person" class using cameras located at the gates of the Asian Institute of Technology (AIT). This effort resulted in a dataset of 2,933 labeled person images, which will be shared during the final submission. These newly collected images were labeled carefully to ensure high-quality annotations, aiming to improve the model's accuracy for this critical class.

Example of collected data:



The results of these experiments led to the identification of two best-performing models: q1_balanced_yolov11_b32_epch1000_02, trained on the partially balanced dataset, and q1_yolov8_best_res_b16_epch1000. The first model benefits from the inclusion of the additional person class data and the larger batch size, which allows for more stable gradient updates. The second model, trained with a smaller batch size, shows robust generalization. Both models demonstrate competitive mAP scores and precision across key classes. During the final evaluation, I will test these models extensively on unseen data and select the one that performs best for submission. This comprehensive approach highlights the iterative process of model improvement and the critical role of dataset balancing and augmentation in achieving optimal results.