# Machine Learning: Video Object Detection

Kieran Chan

## 1 Project Overview

The primary goal of this project was to develop a system to identify sports balls in video footage, with a particular focus on improving the detection of footballs. This was achieved by creating pseudo annotations, which provided a foundation for training the model. The object detection pipeline was then iteratively refined through data augmentation and continuous improvement of the dataset. By evaluating the model's performance and identifying areas for enhancement, the project aimed to achieve greater accuracy and robustness in detecting footballs in various conditions and scenarios.

## 2 Data Collection

Data for this project was sourced from the Sports-1M dataset. However, locating videos containing footballs was challenging due to a significant number of the videos in the dataset were unavailable. The videos featured a diverse range of content, including both old football gameplay and freestyles, but the main criteria for selection was the consistent presence of footballs within the footage. This selection allowed for a more targeted approach in training the object detection model, ensuring that the dataset contained relevant examples for detecting footballs in various contexts.

## 3 Pseudo-Annotations

### 3.1 Tools Used

For this project, the YOLO library was used to detect footballs in the selected video footage. YOLO was chosen for its high accuracy and fast processing speed. Its ability to perform object detection on large datasets made it ideal for the task. Additionally, the project incorporated the OpenCV library for visualising and annotating the detected footballs within the video frames.

### 3.2 Process

The YOLOv8m model was selected due to its balanced performance for object detection, a compromise between accuracy and processing speed compared to other versions of YOLOv8. To ensure efficient data processing, every 10th video frame up to the 1000th frame was extracted, resulting in a representative sample of the footage. The YOLOv8m model was applied to these frames to identify bounding boxes around the sports balls. The model predicted the bounding box coordinates in the YOLO format, allowing for consistent annotation of the frames. If no football was detected in a given frame, the corresponding bounding box information was left empty, providing clear annotations that indicated the absence of a detected object. All frames were resized to a resolution of 640x480 pixels to maintain consistency across the dataset.

### 3.3 Limitations

Despite the YOLOv8m model's strong performance, there were instances where it failed to identify sports balls in certain frames and in some cases, non-ball objects were misclassified as balls. To

address these issues and improve the accuracy of the annotations, LabelImg was used as an additional tool for manual image annotation. Every frame where no ball was detected by the model was carefully reviewed and for frames with false negatives, bounding boxes were manually drawn around the sports balls. This process ensured the annotations were as accurate as possible, providing high-quality data for model retraining and enhancing the overall detection performance."

# 4 First Iteration

## 4.1 Object Detection Pipeline

### 4.1.1 Training Process

The initial phase of training used the YOLOv8m model for object detection, employing a dataset of 200 images split into training and validation subsets, with 70% of the data allocated for training and 30% for validation. The model was trained on Google Colab, using its GPU resources to accelerate computations. The default AdamW optimiser was used to manage the learning rate and update parameters efficiently. Due to time constraints associated with the Google Colab platform, the training process was limited to 10 epochs with an input image size of 640 pixels. This configuration balanced computational feasibility with the need to achieve baseline performance for further iterations.

## 4.2 Evaluation Procedure

After training the YOLOv8m model, a normalised confusion matrix was employed to assess the model's performance across the different classes. The confusion matrix provided a clear representation of the true positives, false positives, false negatives and true negatives, allowing for a detailed evaluation of the model's accuracy. As shown in Figure 1, the results from the first iteration indicated a 76% true positive rate for detecting footballs, demonstrating a solid initial performance. This metric reflects the proportion of correctly detected footballs relative to the total number of actual footballs in the dataset.

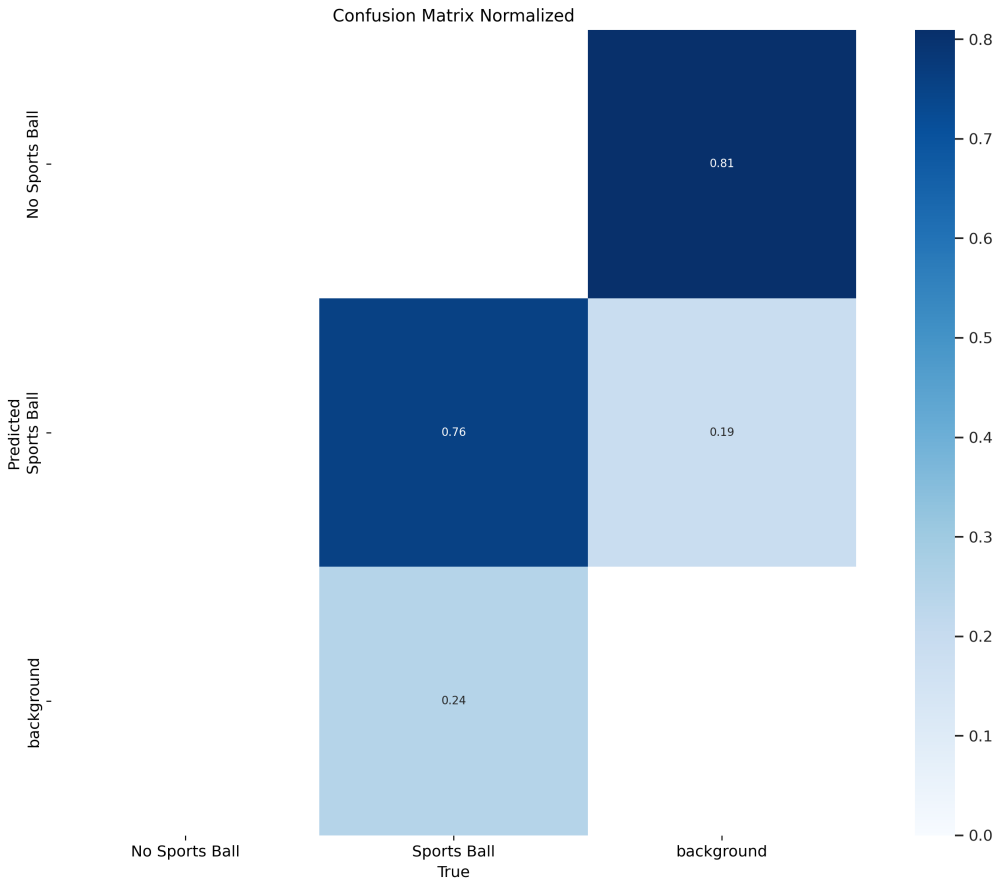


Figure 1: First iteration normalised confusion matrix training

During the training of the YOLOv8m model, several losses were calculated, including those for bounding boxes, classes and Distribution Focal Loss (DFL). For ten epochs, the error between the predicted and actual bounding box coordinates and class labels consistently decreased, indicating improved performance in terms of both localisation and classification. As the loss function decreased, there was a corresponding increase in the model's confidence in detecting objects. However, while the losses for the training dataset were decreasing, they had not yet fully converged. On the other hand, the error for the validation dataset began to increase, suggesting potential overfitting.

In addition to the loss function, several key performance metrics were calculated, including precision, recall and mean Average Precision at IoU 50 (mAP50 and mAP40-95). These metrics revealed a positive trend, with true positive bounding box detections and object identifications both showing steady increases. Furthermore, the model demonstrated an improvement in the accuracy of its bounding box predictions, as the overlap between predicted and actual objects became more precise. Despite these improvements, the results displayed inconsistency, as evidenced by abrupt spikes and dips in the performance metrics, indicating that the model was still struggling with certain variations in the data.

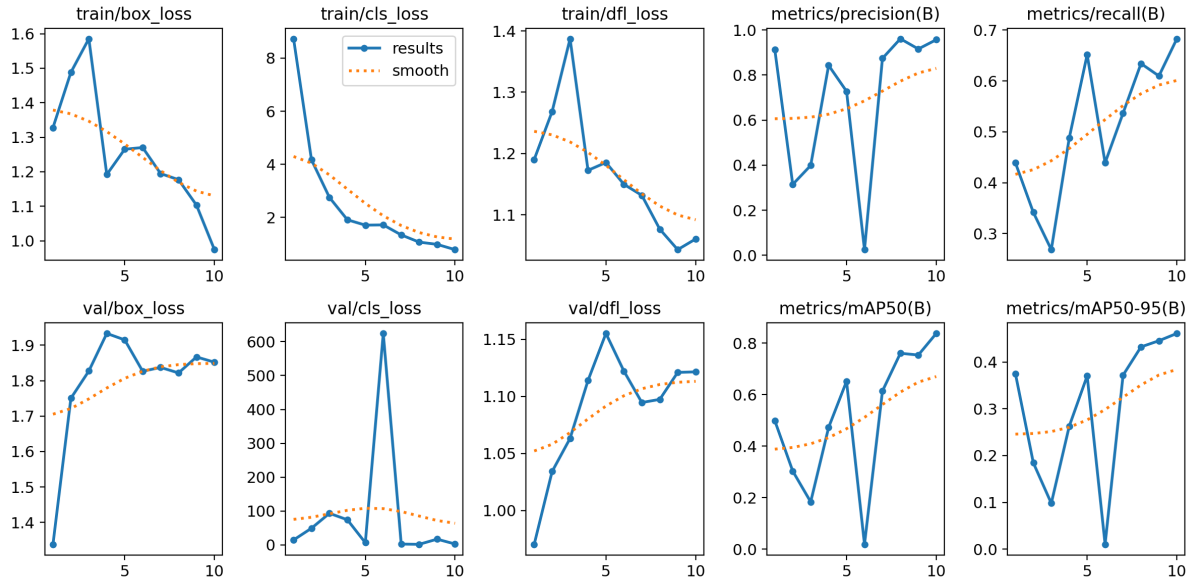


Figure 2: First iteration Training and Validation performance


Figure 3: First iteration predictions made by the model for images in validation

### 4.3 Iterative Improvement

Despite the improvements observed during training, there is a clear need to enhance the model's performance in detecting sports ball classes. The training data losses did not fully converge, suggesting that the model has not yet reached optimal performance. Additionally, as the error for the validation dataset increased, it became evident that overfitting was occurring, where the model was learning to perform well on the training data but struggled on unseen data. This highlights the need to address overfitting.

## 5 Second Iteration

### 5.1 Refinements for second iteration

To address the issue of the training data failing to converge, several refinements were made to the training process. The number of epochs was increased from 10 to 15, allowing the model more time to learn from the data and better capture the underlying patterns. Additionally, the step size was increased by a magnitude of 1, which helped accelerate the learning process by enabling more significant updates to the model's weights during training. These changes were to improve the convergence rate, ensuring that the model could achieve better optimisation and more stable performance over training.

### 5.2 Evaluation for second iteration

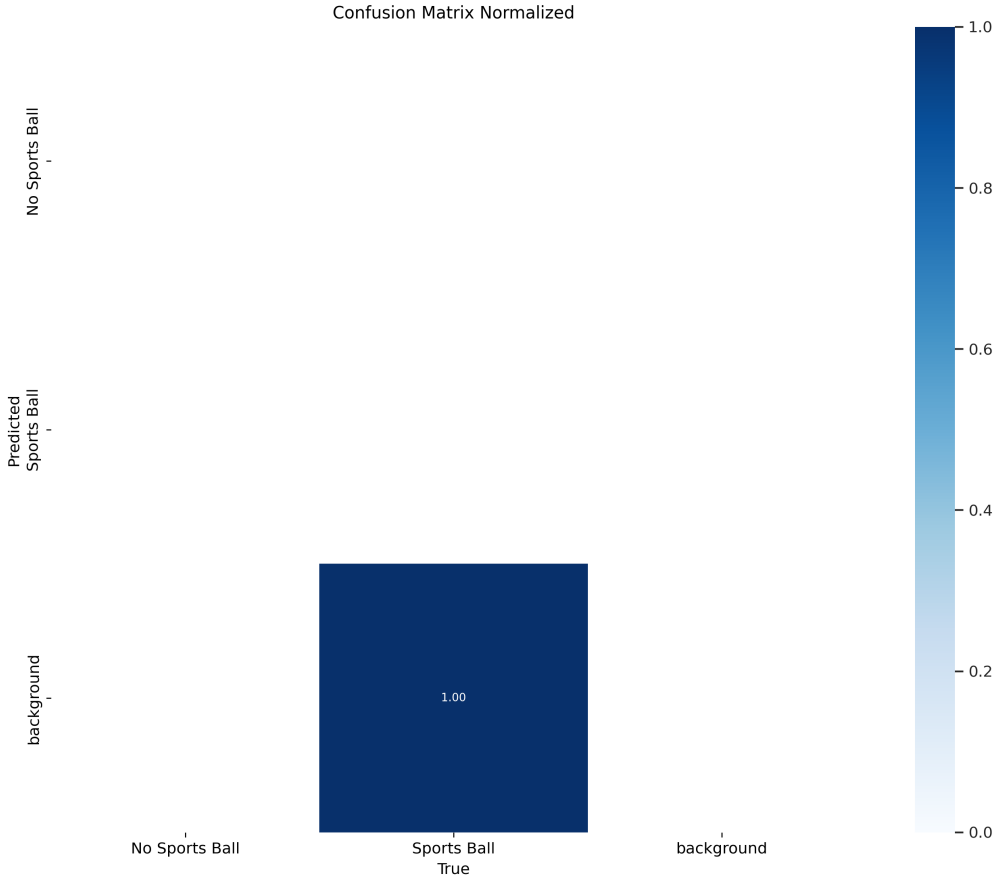


Figure 4: Second iteration normalised confusion matrix training.

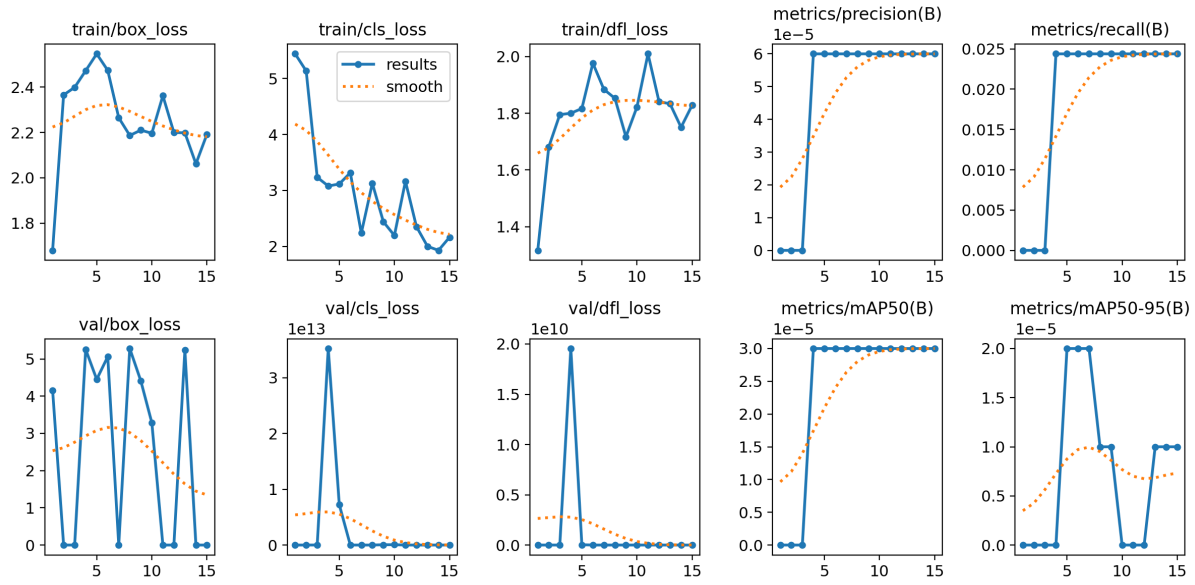


Figure 5: Second iteration Training and Validation performance


Figure 6: Second iteration predictions made by the model for images in validation

This iteration yielded poor results, with no observable improvements in model performance. Despite increasing the number of epochs and step size to facilitate better convergence, these adjustments did not lead to any meaningful enhancements. The training and validation metrics remained stagnant,

suggesting that the modifications did not effectively address the underlying issues in the training process.

### 5.3 Further improvement from second iteration

This outcome highlights the need for a more comprehensive approach to refine the training strategy, potentially involving adjustments to the learning rate, optimisation techniques or augmenting the dataset.

## 6 Third Iteration

### 6.1 Refinements for the third iteration

While the training strategy remained unchanged, an additional 100 images were added and an augmentation was introduced to enhance the diversity of the dataset. Each image was duplicated and flipped horizontally, doubling the number of images in the training set. This augmentation aimed to introduce more variability in the data, enabling the model to train better to different orientations of the sports balls. By diversifying the dataset, the goal was to improve the model's robustness and accuracy in detecting footballs under various conditions.

### 6.2 Evaluation for third iteration

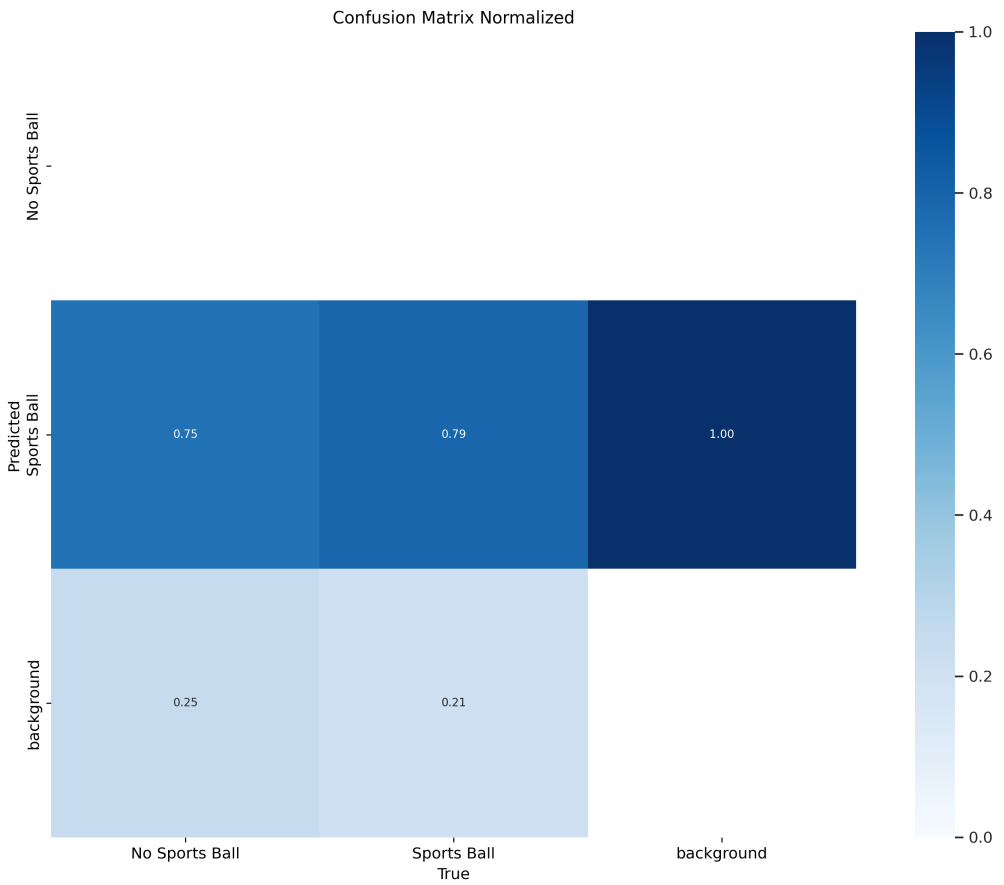


Figure 7: Third iteration normalised confusion matrix training.

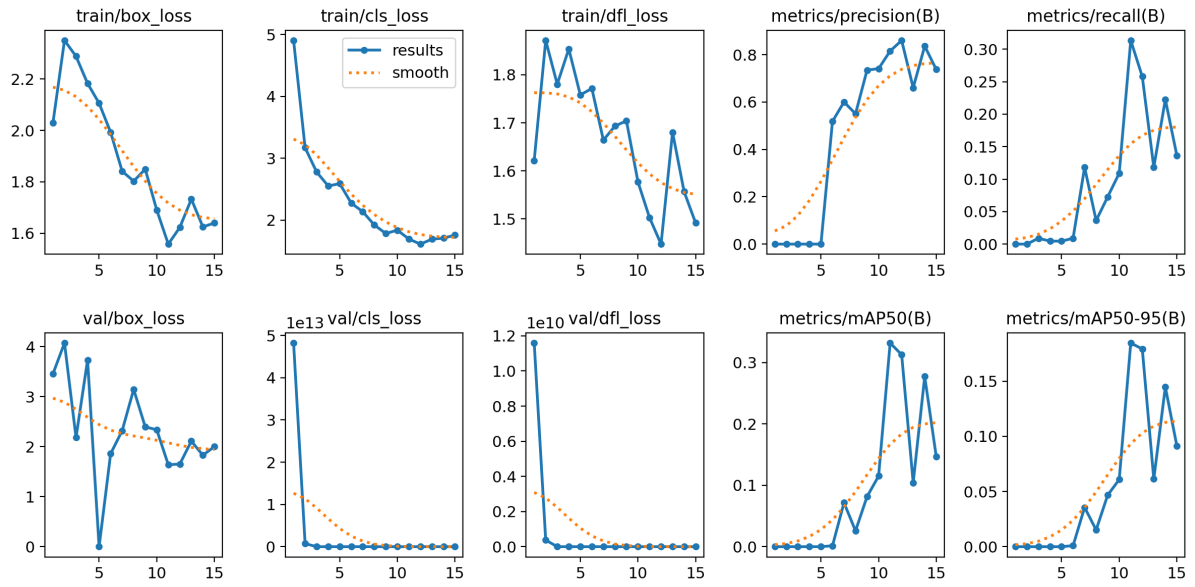


Figure 8: Third iteration Training and Validation performance


Figure 9: Third iteration predictions made by the model for images in validation

The normalised confusion matrix indicated that when the model predicted the presence of a sports ball, there was a true detection rate of 79%. However, this metric was not entirely reliable, as the model also displayed high confidence in detecting sports balls even in frames where only background

elements were present and no ball was visible. This suggests that the model may be misidentifying background objects as sports balls, pointing to issues with its ability to differentiate between relevant and irrelevant features in the image.

Despite increasing the number of epochs to 15, the errors and loss function followed a similar trend to the first iteration, where they gradually decreased. However, as with the first iteration, the training and validation datasets failed to converge fully, indicating that the model had not reached an optimal state.

Other performance metrics, such as true positive bounding box detections and object identifications, showed a positive trend, with steady increases for training. While these metrics were improving, they did not reach high levels, suggesting that the model was struggling to achieve high accuracy in detecting the sports balls. Furthermore, the results continued to display inconsistency, as evidenced by abrupt spikes and dips in the performance metrics, indicating that the model was facing challenges with certain variations in the data, such as changes in background or ball orientation.

### 6.3 Improving from the third iteration

The results demonstrate that a further iteration is necessary to refine the training strategy, potentially involving adjustments to the learning rate or optimisation techniques.

## 7 Fourth Iteration

### 7.1 Refinements for the fourth iteration

The dataset remained unchanged, as augmenting the images had positive effects on the model's performance. The number of epochs was maintained at 15 to ensure sufficient training time, but the training strategy was refined by reducing the magnitude of the step size by 1. This adjustment was made in response to the high box loss of 1.6 and class loss of 2 observed in the previous iteration. By decreasing the step size, the aim was to allow the model to take smaller, more precise steps during optimisation, potentially leading it to converge to a better local minimum and improving overall model accuracy. The decision to keep the number of epochs at 15 was to ensure that training could persist long enough to observe meaningful improvements in model performance without sacrificing the training duration.

## 7.2 Evaluation for fourth iteration

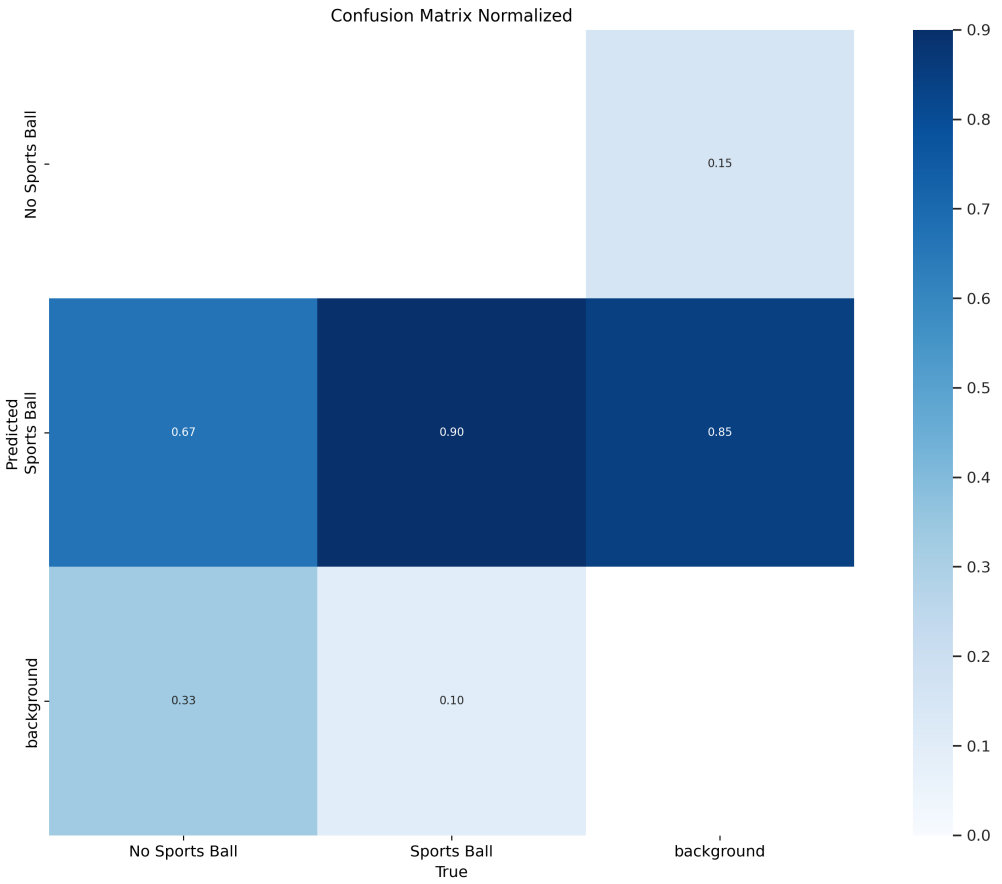


Figure 10: Fourth iteration normalised confusion matrix training.

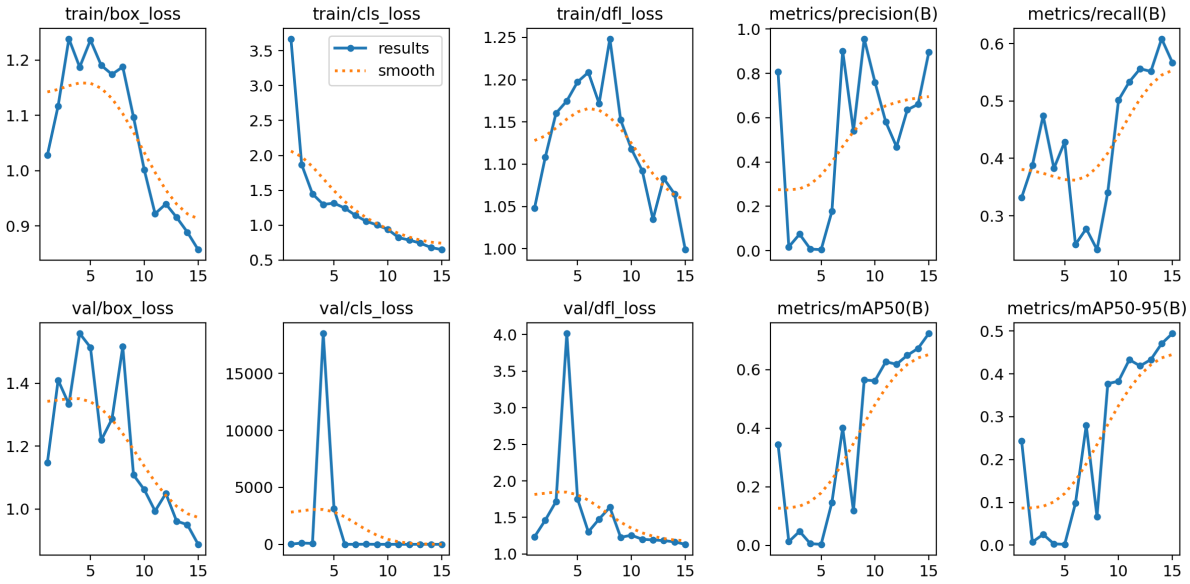


Figure 11: Fourth iteration Training and Validation performance


Figure 12: Fourth iteration predictions made by the model for images in validation

The normalised confusion matrix revealed an improvement in the model's performance, showing a true detection rate of 90% when predicting the presence of a sports ball. However, similar to the third iteration, this was not reliable. The model still showed high confidence in detecting sports balls, even in frames where only background elements were present and no ball was visible.

Decreasing the step size by a magnitude of 1 led to a significant improvement in both the training box and class loss, with both values dropping below 1. This refinement helped to stabilise the model's learning process and reduced the large loss values observed in previous iterations. However, despite this improvement, the losses still failed to converge fully, indicating that the model had not yet reached an optimal state.

Other performance metrics, such as precision, recall and mAP50, showed a positive trend, with steady increases observed during training. These improvements suggest that the model was learning to detect sports balls more accurately over time. However, the results continued to exhibit inconsistency, as evidenced by abrupt spikes and dips in the performance metrics. This ongoing instability highlights that the model still struggles with certain variations in the data and further refinements are necessary to achieve more stable and reliable performance.

## 8 Challenges and Improvements

To further improve the model's accuracy and maintain clean and accurate data, images where the ball was incorrectly identified as a false positive need to be reviewed. Specifically, the bounding boxes for these false positives should be removed to ensure that the dataset remains clean. This process will help eliminate any noise in the dataset that could affect model performance.

Another crucial step is to find a more up-to-date dataset that better represents the diversity of

current scenarios. An updated dataset could provide more varied examples, helping the model predict better and improve detection accuracy.

Finally, training the model with larger epochs will allow for more extensive learning, potentially leading to better convergence and improved accuracy. A longer training period will also give the model more opportunity to refine its predictions and reduce inconsistencies in the results."