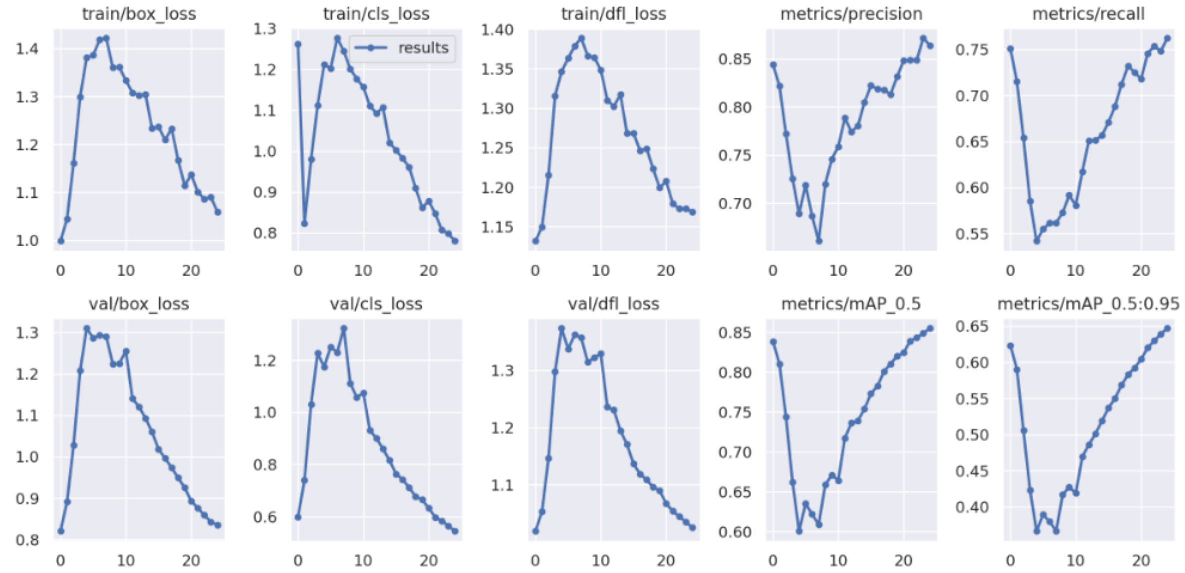
1.1) Training the Human Detection Model

COCO person dataset was trained with the YOLO-v9 model to create a human detection model. This was chosen as it is a state-of-the-art model and COCO has a large subset of accurate bounds. This took 42 minutes to train with a batch size of 16 for 25 epochs, achieving a high accuracy of 86% and recall of 76%. The loss values in *Figure 1* below indicates a significant decrease during training and no signs of overfitting.



The model was run on every 640 frames, on each training set video, taking 1.5 hours, to extract diverse human patches with confidence >82%.

1.2) Data Preparation and Class Selection

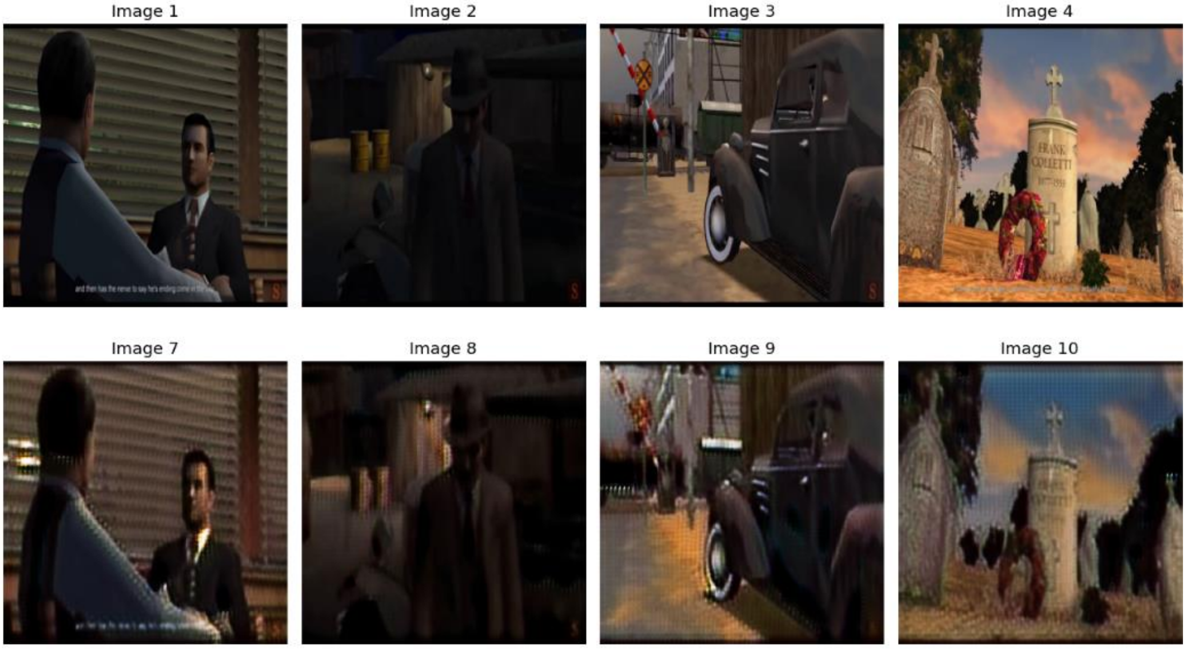
Patches should be clear and consistent to transfer styles appropriately – hence **illumination normalization** was carried out to maximise useful frames for training. This also enhances object detection.

The state-of-the-art convolutional vision transformer (CVT), adapting vision transformer (ViT) was implemented with the CIFAR-100 dataset – with baby, boy, girl, man and woman classes – to aptly select and capture the underrepresented classes for future style transfer with a balanced dataset. This achieved mixed accuracy results on the test set between 15% and 50% with the state-of-the-art model. This was implemented with calculated early stopping at 8 epochs, to prevent overfitting.

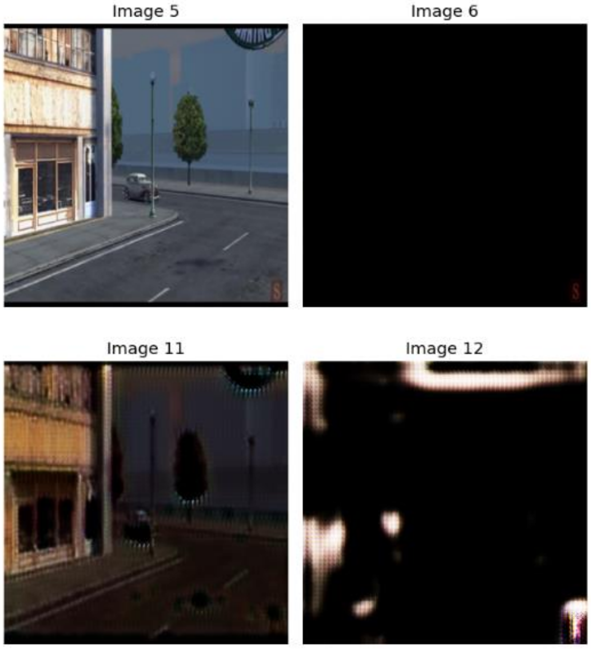
2.1) Training for Style Transfer Between Game and Movie Frames

CycleGAN was implemented with frames at intervals of 500 from the training set, split into a game and movie set. This took 52 minutes to train with 12 epochs, 4 delay epochs and a batch size of 2 (maximising GPU usage) to successfully transfer styles forwards and backwards. The epochs were chosen to prevent overfitting and incorrect convergence.

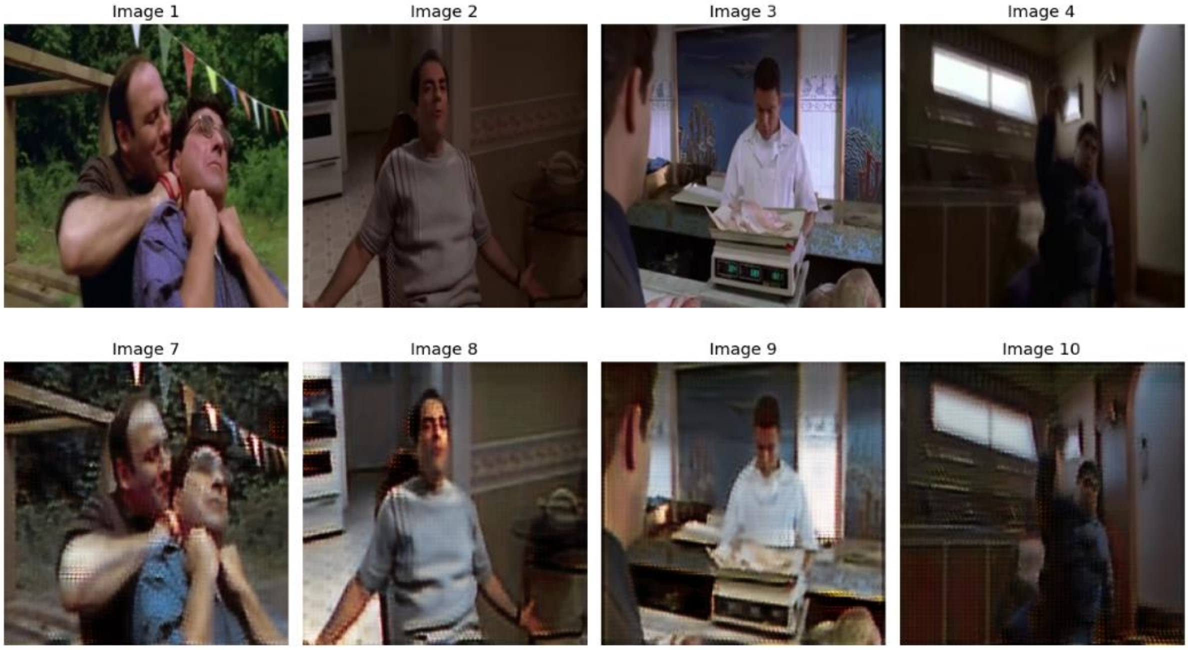
The training results appear to be standard with losses higher for initial epochs. These general decrease, starting at 2.07 and 1.24 and reducing to 0.384 and 0.195 for the generator and discriminator respectively, indicating that each learns well.



The successful game-to-movie style in *Figure 2* above (where the top rows show the original image) demonstrates the fine attention to detail in the model – through translating facial and object features well. Additionally, the success in the illumination normalization is particularly clear in image 8.



The weaker images in *Figure 3* show that at times distant objects and flat colours build noise around them.



For movie-to-game in *Figure 4*, the flat and brighter game textures are replicated well, particularly in image 8 and 10. Colours are transferred very well.

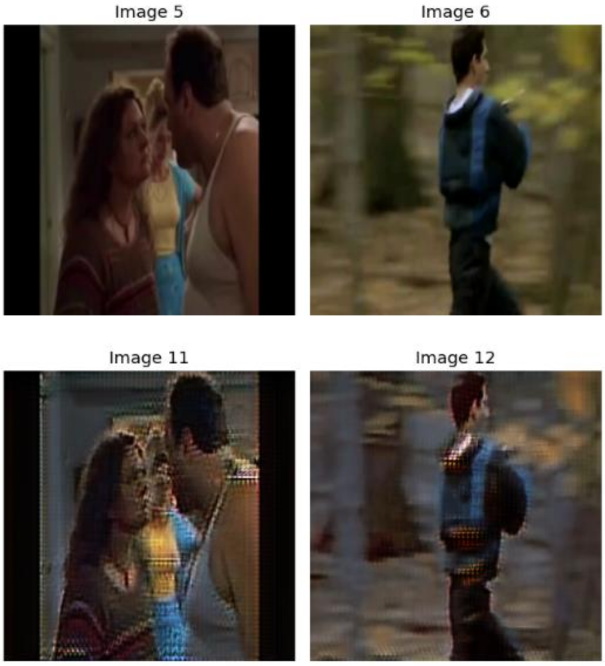


Figure 5 shows that in some cases, distant and also moving objects build noise and sometimes too much detail is maintained.

Batch LPIPS was used to determine the similarity between fake and real pairs. Scores ranged between 0.326 and 0.468, indicating variety yet not too much diversity from the original image – as style transfers should.

2.2) Proposed Improved Video Creation

To successfully create an improved video, the following process was attempted:

1) Obtain a video from the standard CycleGAN style transfer
2) Train a **CycleGAN style transfer specific to people patches** (could have been enhanced with foreground people from 1.2.
3) Detect person patches for every frame on the test image using 1.1.
4) Run the person patches through the new CycleGAN patch model
5) Add these patches into to the frames from 1) at the correct position

This method was implement, however step 5 did not run correctly. As a results, the back-up method of directly using the test video on the CycleGAN patch model was implemented.

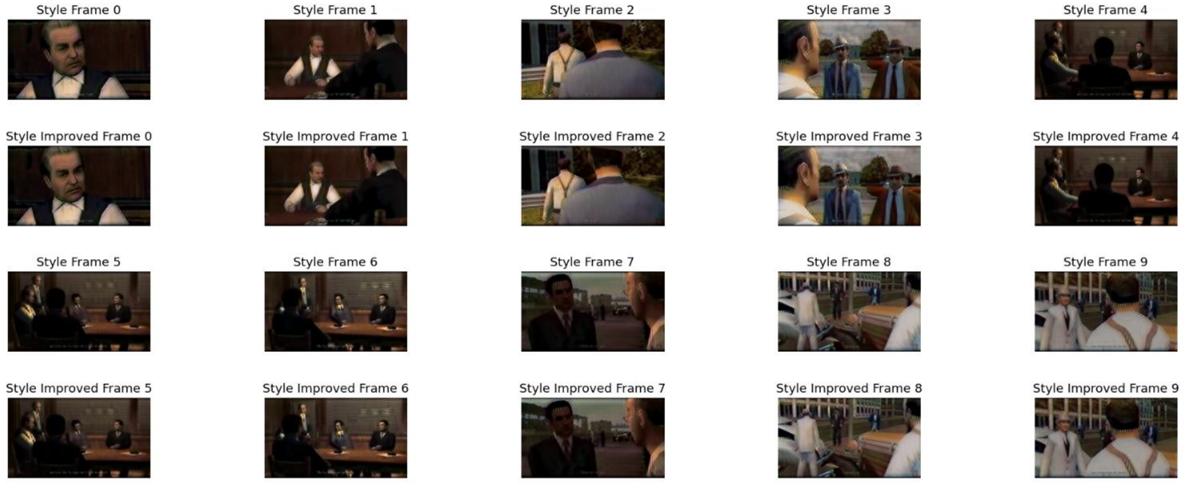


Figure 6 above shows the original model from 2.1 (rows 1 and 3) and the improved model from 2.2 (rows 2 and 4). This demonstrates only very minor changes, specifically reflected in the detail of close-up people. To improve this further, the detailed method from previously could be implemented.