Advanced Computer Vision Summative Assignment

May 9, 2024

0.1 1.1

The *yolov9*[3] model was utilized, representing a state-of-the-art advancement despite minimal improvements over *yolov8* (figure 1). To enhance the algorithm's efficiency, the *vid-stride* parameter was modified to align with the video frame rate, resulting in the processing of only one frame per second. This adjustment significantly decreased the algorithm's runtime and increased dataset variability, ensuring a more varied dataset

0.2 1.2

The *'classify_pose'* function in *openposeModified.py* uses specific body parts to identify how a person is positioned in a photo. It checks if the nose is above the neck to tell if the body is facing forward or backward. The function looks at the neck and hips to decide the overall pose and uses the distance between the shoulders to distinguish between full body or just upper body views. This method is straightforward and accurate for classifying poses. OpenPose[1] was chosen over AlphaPose[2] and YOLO Semantic Segmentation[3] because it is faster and almost as accurate, making it better for real-time use.

0.3 1.3

The majority class being "others" complicates the sampling process, particularly due to data sparsity challenges in training GANs. To mitigate this, *openposeModified.py* was adjusted to classify both front and side facial profiles, allowing for the inclusion of some "other" images. Nonetheless, this adaptation led the model to overfit to these specific face profiles during training, affecting its performance in testing. Attempts to implement controls for brightness and restrict samples to only white individuals, reflecting the majority observed class, were unsuccessful. Consequently, the decision was made to focus solely on augmenting the dataset. Following these modifications, the data distribution was as follows (figure 2 to figure 5):

0.4 2.1

A CycleGAN[1] changes images from one style to another without needing paired examples, consisting of generators and discriminators. The generators alter the style of the images, while the discriminators ensure these changes are realistic. The model employs three main types of loss functions: adversarial loss, cycle consistency loss, and identity loss.
Adversarial Loss is used within the GAN framework to compel the generator to produce images that appear real to the discriminator, employing binary cross-entropy to penalize incorrect classifications of images as real or fake.
Cycle Consistency Loss ensures that an image from one domain, when translated to another, can be reverted back to its original domain without losing content. This loss helps preserve important attributes between translations and boosts the model's ability to generalize.
Identity Loss maintains color and composition consistency during translation, ensuring images from the target domain remain unchanged when processed by the generator, preserving essential characteristics.
The total loss function for training the CycleGAN combines these elements:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F) \tag{1}$$

[1] https://towardsdatascience.com/cyclegan-learning-to-translate-images-without-paired-training-data-5b4e93862c8dhttps://towardsdatascience.com/cyclegan-learning-to-translate-images-without-paired-training-data-5b4e93862c8d

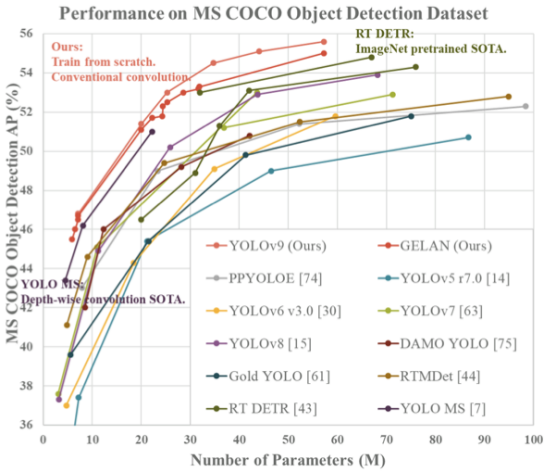

Figure 1: Performance on MS COCO-Object Detection Dataset

Hyper-Parameters	Values
batch_size	24
Activation Function	LeakyRelu
Learning_rate	0.0002
Epoch	100
Loss function	binary cross-entropy
Model Optimizer	Adam(0.0002, beta=0.5,0.999)

Table 1: Hyperparameters for CycleGAN in 2.2

where λ is a parameter weighting the importance of cycle consistency loss.

Directly inputting frames into the CycleGAN proved suboptimal, as the complexity of the data hindered effective style transfer. Figures 6 and 7 show little visible difference in image matrices indicating a failure in achieving style transfer. There wasn't much visible difference between the success and failure case.

0.5 2.2

The dataset was augmented with scenes from the 1996 Italian Mafia movie "Gotti" and gameplay footage from GTA Vice City and GTA 4, aligning with the style and quality of "The Sopranos," "The Godfather," and "Mafia." However, combining video game footage with film introduced challenges, particularly due to the differing texture styles and graphical representations between the two game versions and cinematic footage. GTA Vice City and GTA4 differ in texture quality, shot composition, and cinematography, reflecting their release periods.

The inclusion of subtitles in the video game footage added complexity, causing noise and artifacts during style transfer. Despite these issues, transferring styles from movie to game was generally more successful than game-to-movie, as the detailed cinematic content provided a robust base for simplifying into the less detailed game environments. This suggests an advantage in using high-detail film content for style transfers into simpler digital formats.

It's important to note that when comparing the two models, the left model is the CycleGAN from 2.1 and the right model incorporates methods from 1.1 and 1.3 into the CycleGAN. The focus was to style transfer humans so the local methods purely enhanced the style transfer of facial features. An oversight being not enough data for the background.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recog-*

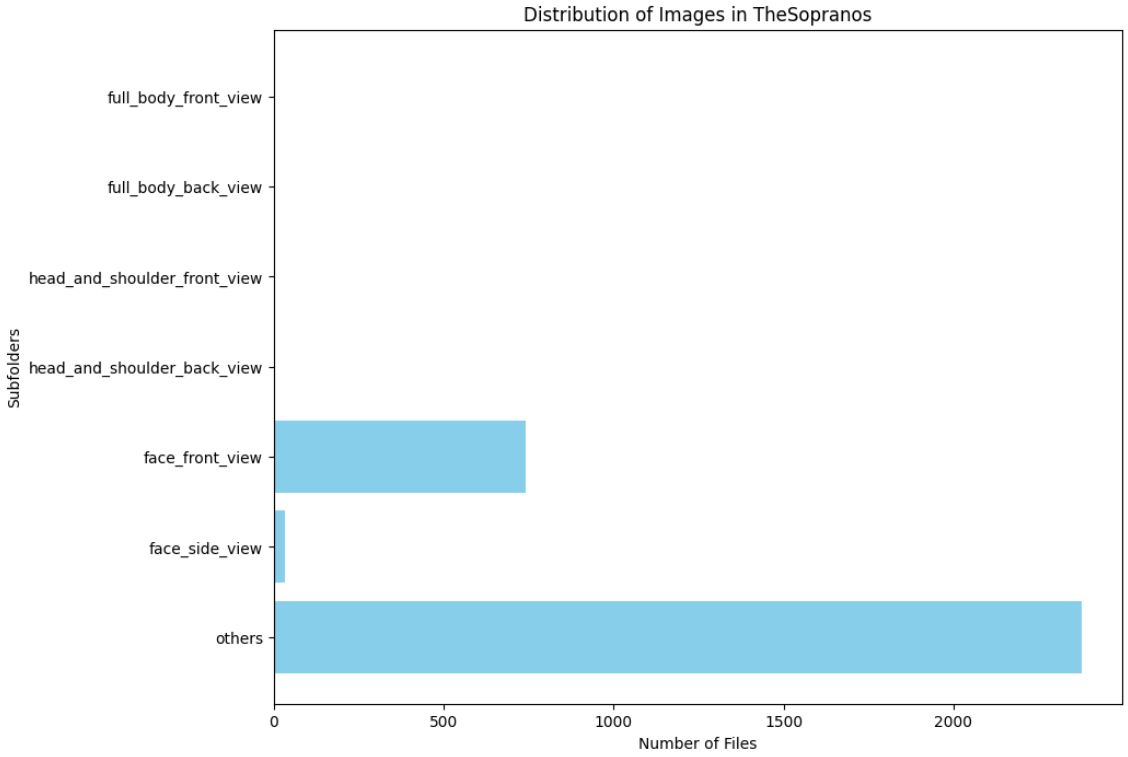

Figure 2: Distribution of Image Classes for The Sopranos

nition, pages 7291–7299, 2017.

[2] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[3] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2024.

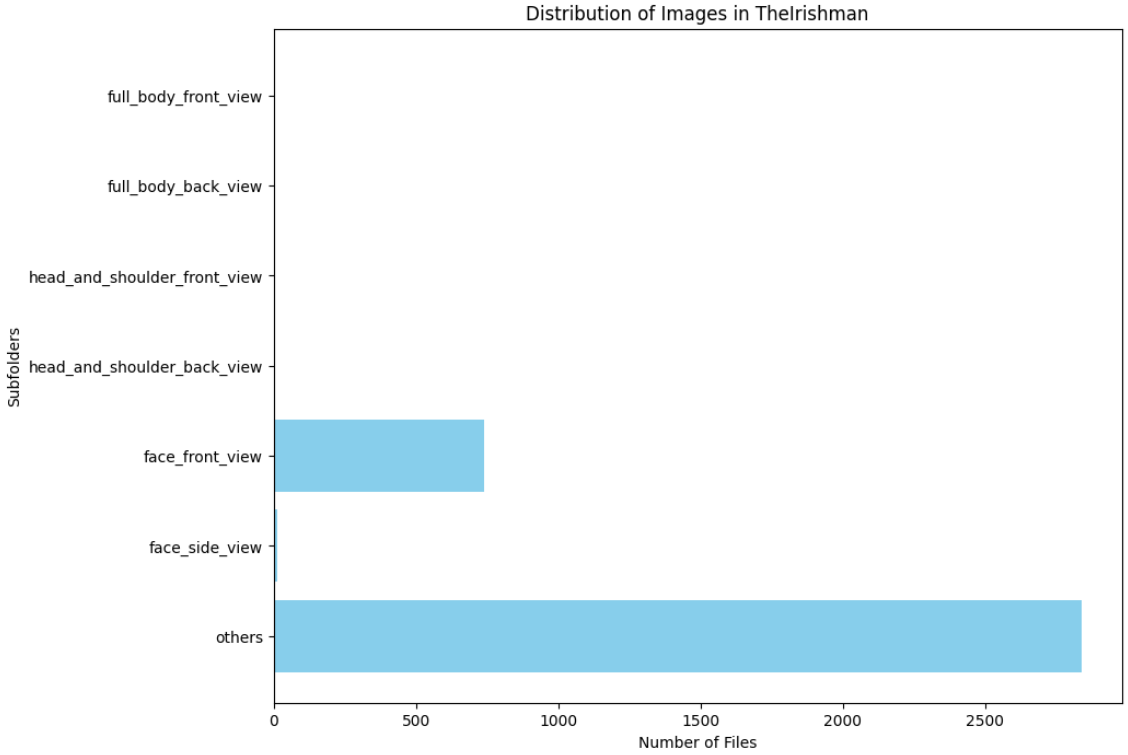

Figure 3: Distribution of Image Classes for The Irishmen

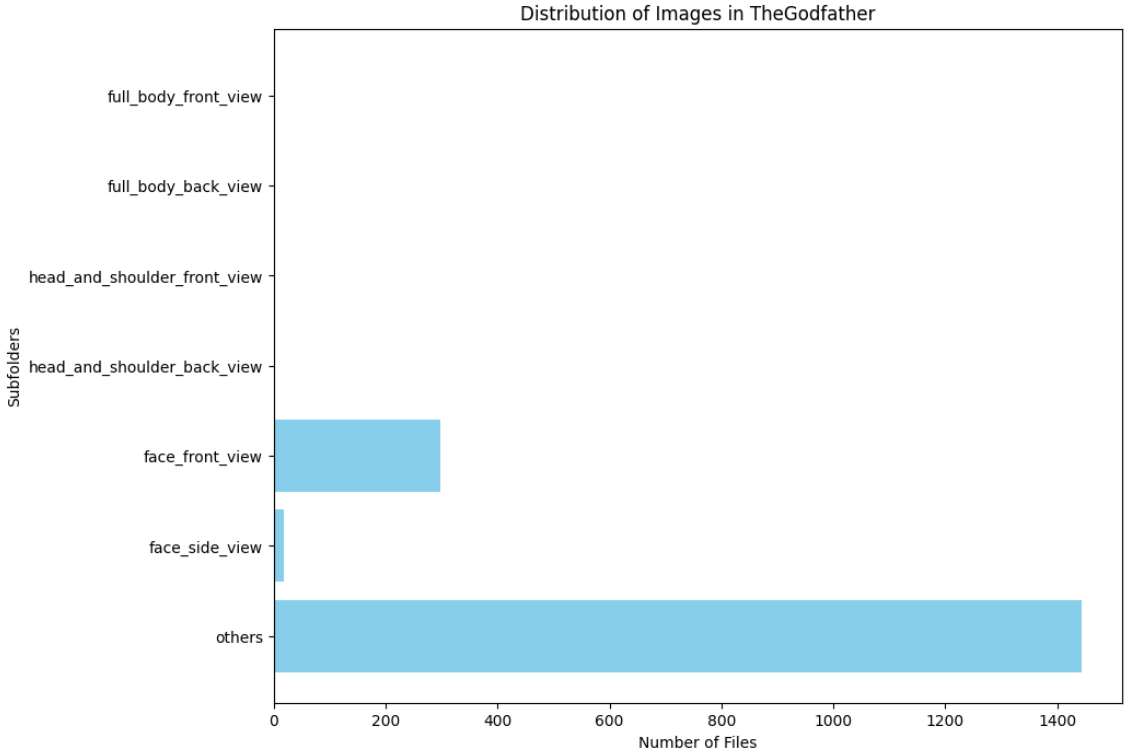

Figure 4: Distribution of Image Classes for The Godfather

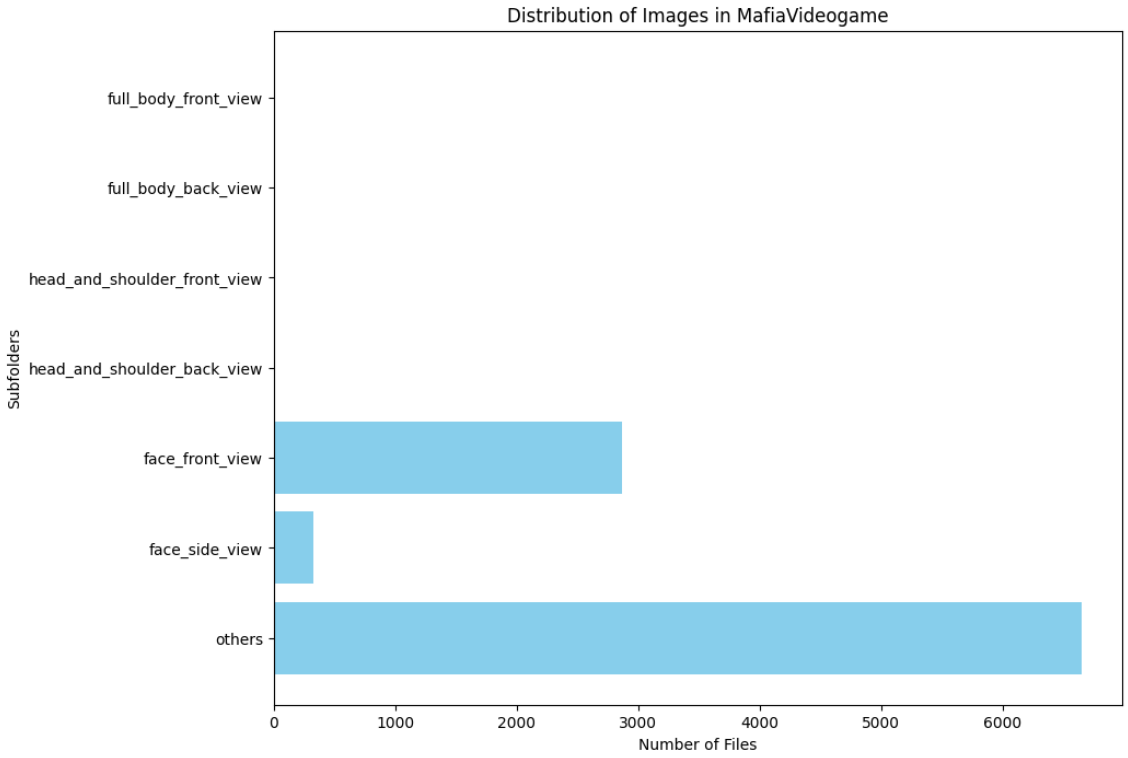

Figure 5: Distribution of Image Classes for Mafia Videogame

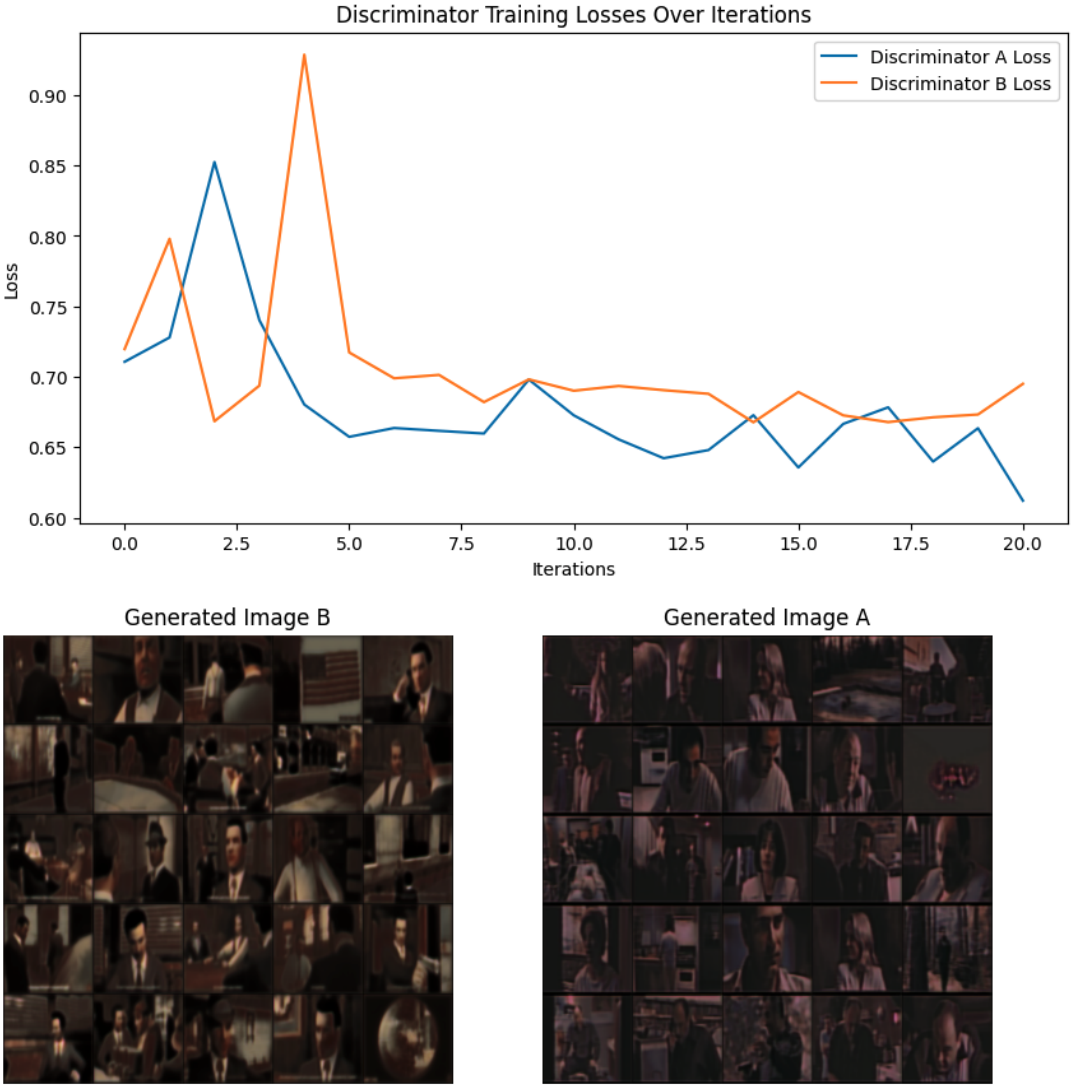

Figure 6: Discriminator Loss, first epoch generations for 2.1

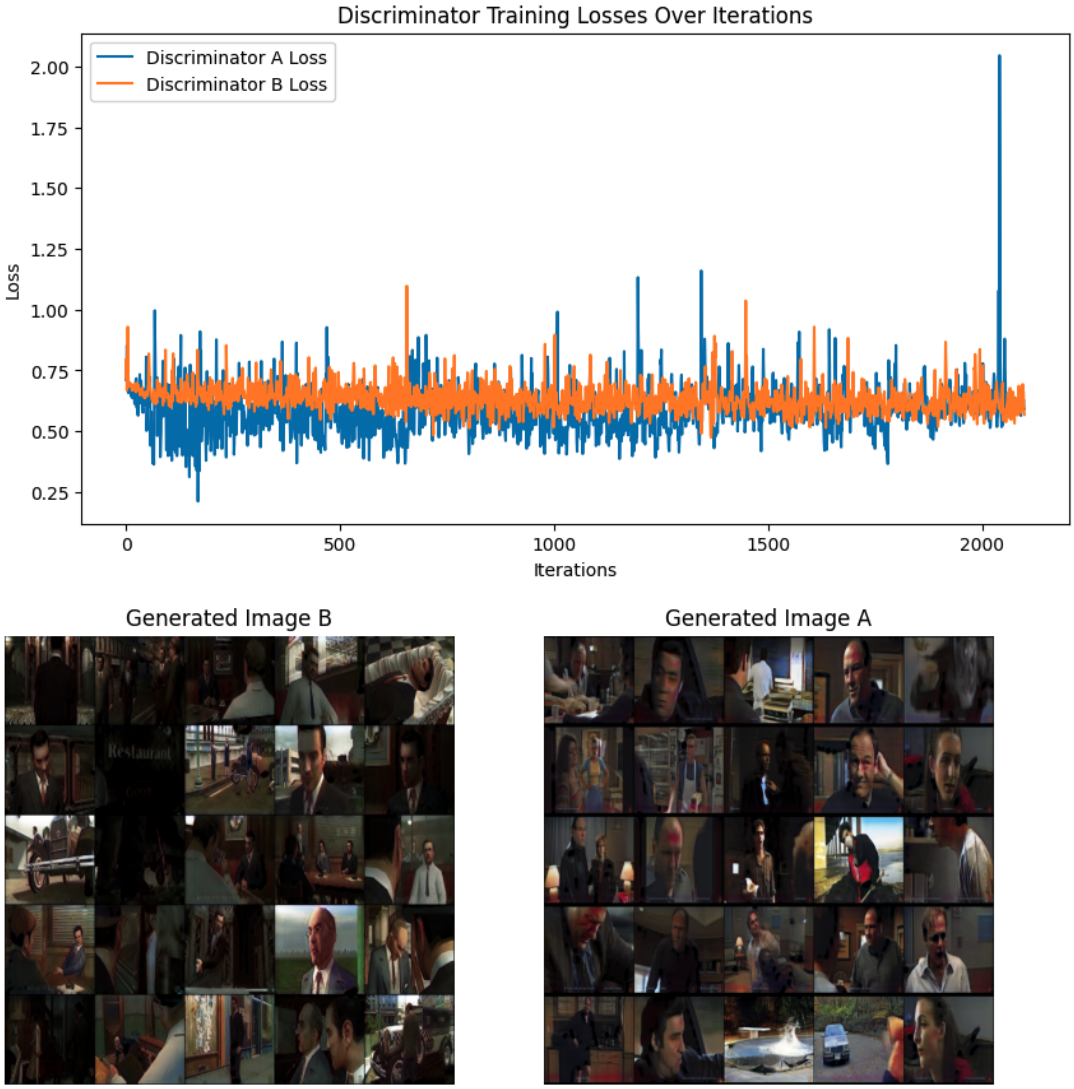

Figure 7: Discriminator Loss, last epoch generations for 2.1

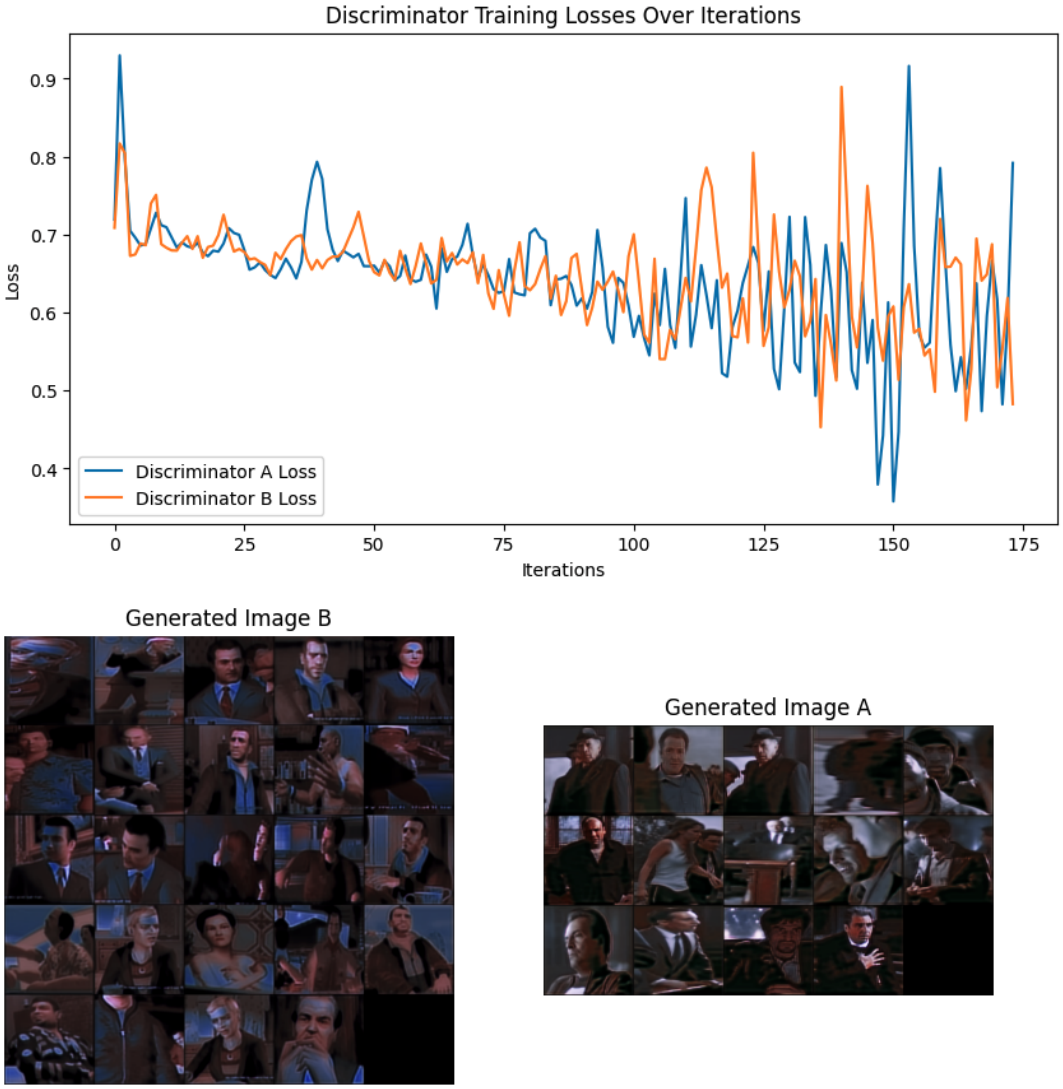

Figure 8: Discriminator Loss, first epoch generations for 2.2

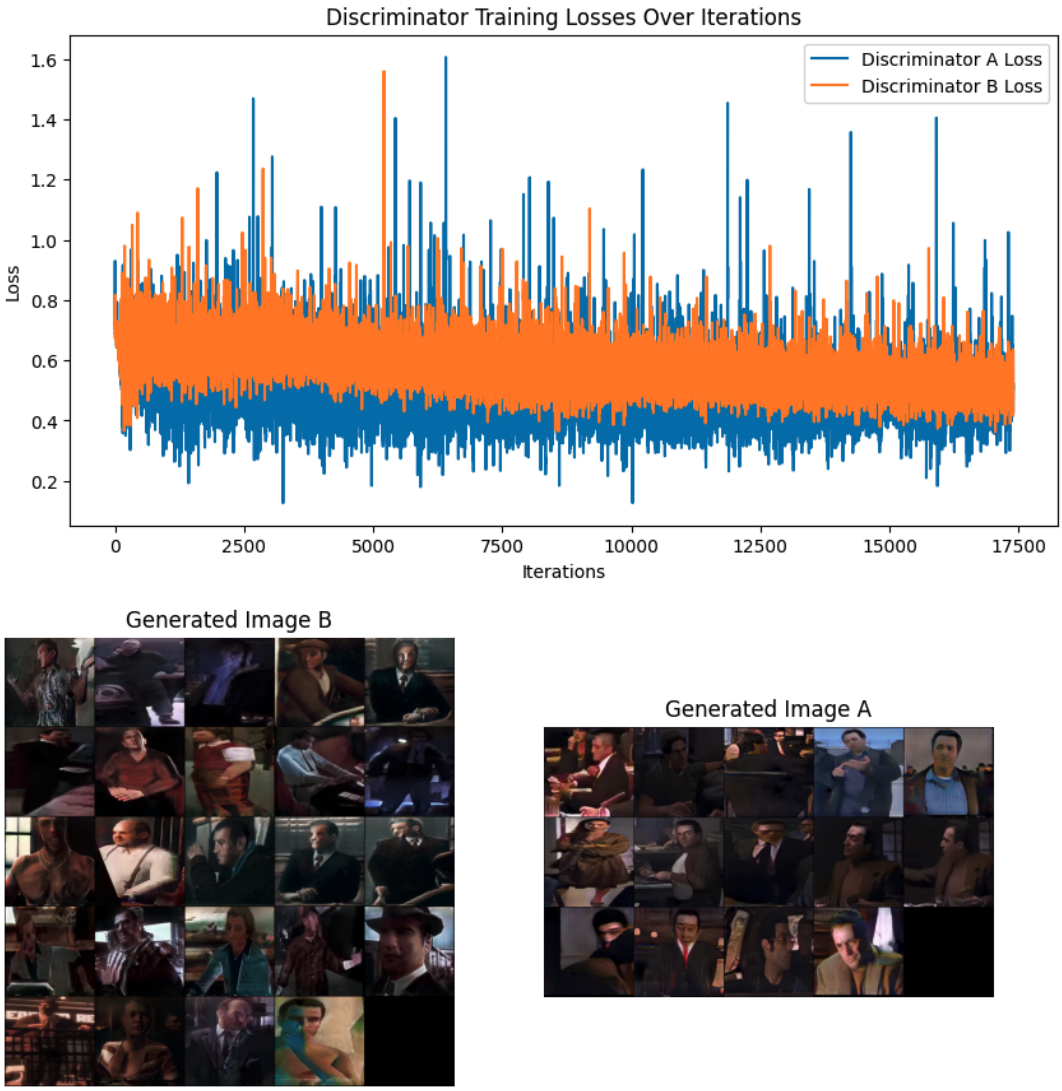

Figure 9: Discriminator Loss, last epoch generations for 2.2

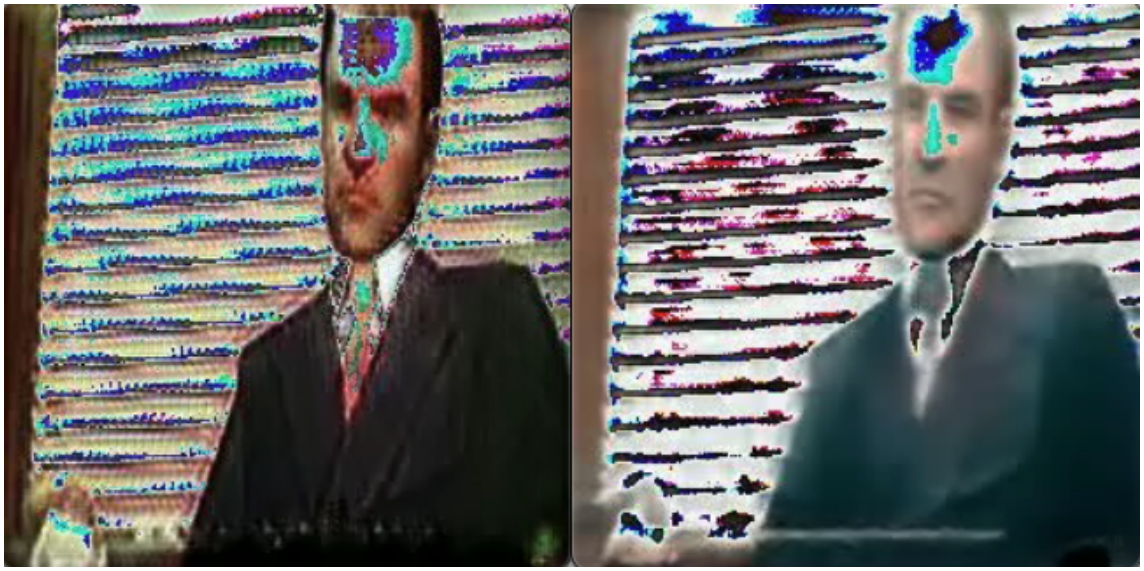

Figure 10: Style transfer alters appearance of character

Figure 11:

Figure 12: softer tones decrease detail which creates confusion which makes the characters on the right look more human.

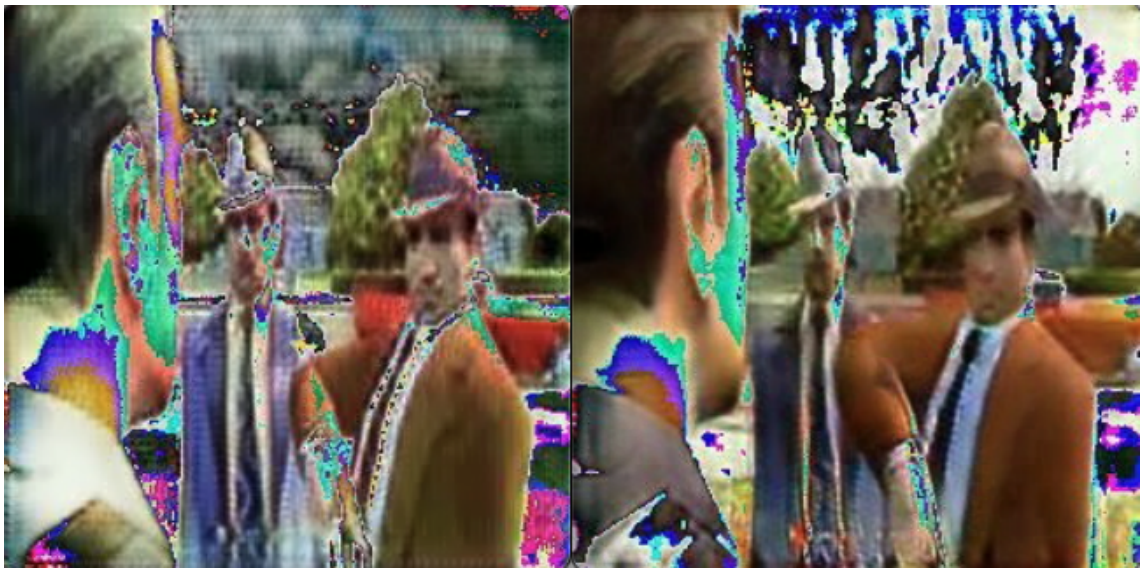

Figure 13: The background appears to be more realistic

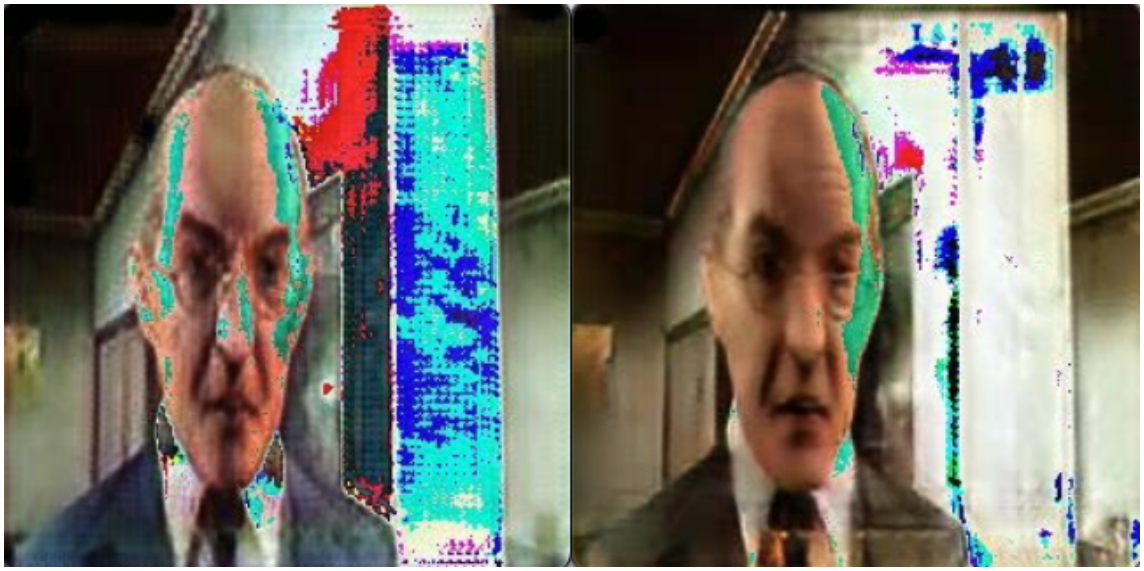

Figure 14:

Figure 15: More pronounced facial features from incorporating side profiles into training

Figure 16: softer tones decrease detail which creates obscurity as to whether or not the characters on the right are more human.

Figure 17: Partial resemblance shows that the game character's style transferred to Chris from the Sopranos

Figure 18: Partial Resemblance also shows how the NPC on the right became Tony Soprano (slightly).

Figure 19: Softer fabric on the blazer, background composition has changed