Deep vision with generative adversarial networks to augment and classify tackle images in American youth football

by

Yeling Hu

B.S., Kansas State University, 2021

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
William H. Hsu

Copyright

Abstract

This report presents an application of convolutional neural networks (also known as *convnets* or CNNs) to the video analysis task of detecting risky tackles in American football via classification of image sequences. The solution approach focuses on fine-tuning of pre-trained convnets, extraction of spatial features, and using generative adversarial networks for data augmentation.

American adolescents compete in youth football, one of the riskiest sports in the US with a large proportion of head injuries like concussions, as reported in the Youth Football Surveillance System.[1] To provide a football team and coaches with more convenient and efficient training, deep learning automatically classifies tackle videos that record tackle actions. Inflated 3D Convents (I3D) is used for this task; However, I3D does not have ideal performance when the video data was used to train the model because we lack sufficient data and the label system is complex. Generative adversarial networks (GANs) can efficiently augment data. In this study, the style-based generator, StyleGAN, was used to solve data problems. At the same time, three other GAN models were used on the same data set to horizontally compare StyleGAN's performance to the performance of other GAN models. In the end, StyleGAN performed best. Although the training data took longer with this model, the results were clearer with a higher resolution showing more player detail. The images generated by StyleGAN were more varied than images from other models.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I must express my gratitude to my supervisor, Dr. William H. Hsu, for his selfless help over the past three years. He is patient even when I struggle with English. He is always generous with constructive suggestions and guidance, which are extremely helpful to me. Dr. Hsu always corrects my mistakes kindly. His help has alleviated my fear and confusion in a foreign country.

I would like to extend my deep gratitude to my professors, Dr. Mitchell Neilsen and Dr. Arslan Munir. They have taught me much, and I deeply enjoyed their lectures. I also wish to thank Dr. Scott Dietrich and Dr. DeLoach for taking the time to attend my final defense; I will humbly listen your comments.

Lastly, my thanks also go out to my family. My parents did their best to support my study here, although they did not even go to high school. I am grateful to all those who have helped me, especially Weijiang who brings me happiness, and Ray, who assists me academically.

Chapter 1

Introduction

This introduction describes how scientists train the SATT system to assess athletes. Deep learning-based vision research continues to progress from object detection and scene classification in images to action recognition and analysis in videos; therefore, using video recognition models to detect athlete abnormalities will be explained in detail. This chapter will also explicate the problem of lack of data and data sets imbalance, proposing solutions to each that use generative models.

1.1 Overview

Football is a popular sport in the United States; however, it is a risky sport. Football players are most commonly injured because they use improper form during blocking and tackling, which results in head injuries, usually concussions. In a 2010-2011 emergency room study, nearly 20% of head injuries were directly related to football,[2], and concussions accounted for 9.6% of total injuries reported by the Youth Football Surveillance System.[1]

SATT is a tool developed to reduce improper form in blocking and tackling and thus the occurrence of unsafe tackles. SATT is a standard used to score the performance and quality of the six basic elements of an American football tackle. The six elements of a correct tackle are player control (PC), head-eye and torso position (HET), strike zone (SZ), ascending hit

(AH); maintained leg drive (LD); and final position (FP). Each element is scored using a 4-point sequential scale (0-3) that evaluates the overall tackle quality based on the total score (with a maximum of 18 points). The scores can monitor the position of the head and torso at the contact point during tackle action. If the action is risky, zero (0) points are assigned. If the component exists but is executed inefficiently or ineffectively, one (1) point is awarded. Two (2) points are awarded if the component meets only part of the motion criteria when it appears. If the element meets all criteria, it is awarded three (3) points. A

To detect danger to athletes more conveniently and quickly, deep learning is used to automatically detect and identify information from collected tackling videos. The videos are labeled "risky" or "safe" based on SATT. Aside from video classification, I used several GANs to generate more important frames to enrich the imbalance of data sets.

1.2 Motivation

As stated, SATT not only measures player safety, but also evaluates the effectiveness of tackling. Thus, assessing tackling action using SATT is necessary. With the rapid development of deep learning, researchers have moved beyond detecting static images and have created several methods that can directly process videos. In the Kansas State University Laboratory for Knowledge Discovery in Databases (KSU KDD Lab), the video processing model was applied to tackling videos. Please note, a machine can capture more details than the human eye as well as reduce human bias. At the same time, a machine saves time and energy. Thus, the first goal of the research was to classify tagged tackling video data using I3D.

However, if data is insufficient or the data sets unbalanced, the result of classification and detection will be unsatisfactory. Because we have less risky video data than safe video data, the accuracy and recall rate of the risky data is very low. Therefore, we must research effective methods to address data imbalance and to expand data. Thus, for this research, three questions are posed:

- For data with complex semantics, which GANs model better learns features?

- For data with complex semantics, which GANs model learns more quickly and efficiently?
- For larger data sizes, which GANs model has higher resolution and produces higher quality images?

1.3 Problem Statement

Since Goodfellow et al. introduced GANs in 2014,[3] research on GAN has been in full swing. GANs models have evolved from simple, monotonous prototypes to varied, clear, highly accurate images. DCGAN was an important milestone in GANs research, the first time a convolutions neural network was used in GAN, producing excellent results. DCGAN proposed an important architectural change to solve training instability, mode collapse, and internal co-variate conversion.[4] In the process of improving the generation of high-value and low-variety images, researchers have successively proposed BigGAN, StackGAN, and CycleGAN, among others.[5–7] While improving the quality of generated images, researchers also wanted to differentiate the target subjects of GAN models. The main improvement of BigGAN was the orthogonal normalization of the generator.[5] StackGAN was used for text to image synthesis.[6] CycleGAN was used for different image-to-image translations.[7] In model evolution, InfoGAN, AC-GAN, and styleGAN came subsequently.

Football tackle images for this study are collected manually so as to assure sufficient image quantity and quality. This means the issue of imbalance and scarcity of the data set can be mitigated by using GAN models that generate new examples based on learning from minority examples. Moreover, among the many GAN models, finding a model suitable for training this representative data set would be a significant achievement. This work is not a criticism of any method but takes a step towards better understanding how GANs models can be used.

1.4 Objectives

Football tackling classification data sets often lack essential information, so classifying the correct video architectures is not easy. Most 3-D models result in undesirable performances. Thus, the Inflated 3D CovNet (I3D) was chosen because it can learn the seamless spatio-temporal feature because the optical-flow stream is split from traditional frame learning, and the I3D is a pre-trained model using an image classification model.[8] The first objective of the study is to analyze how much performance improves with these mode training small-scale benchmarks.

In the next section, the research tests which GAN is suitable for what kind of data set, how the fundamental internal network of different models works, and which GAN model best matches our data set. Our research requires a model that can perform the following operations:

1. Building an image data sets by extracting the important frames from videos and labeling them correctly.

2. Training the image data sets on the deep neural network training classifier.

3. Training the image data sets on different GANs network to increase the magnitude of the image data set.

4. Training newly generated data on the same classifier and comparing it with the baseline of the original data sets.

Chapter 2

Background

This chapter surveys extant architectures and training methods for deep learning neural networks. It first introduces standard convnets and the innovation of 3-D convolution that allows three-dimensional features to be extracted for solid object recognition. Next, it describes additional advances in spatiotemporal feature extraction. It then describes the need for generative adversarial networks (GANs) for data synthesis in data-poor domains, and finally discusses the type of GANs applied in this work.

2.1 Deep Learning for Classifying Video

In processing images, only static images are convolved, so a 2-D convolutional network suffices. In video interpretation, however, retaining timing information is necessary to learn spatiotemporal features at the same time. If 2DCNN is used to process videos, the motion information encoded between consecutive multiple frames will not be considered. Andrej proposed Slow Fusion as the first significant achievement of deep learning in the video field. Slow Fusion extracts the features of each frame and fuses all features as a basic concept.[9]

2.1.1 3DCNN

Because convnets were shown to be very effective for object detection, object classification, and scene classification from images, they were soon applied to motion recognition in videos. They can extract features from space and time dimensions and then perform 3-D convolution to capture the motion feature from multiple consecutive frames. A 3DCNN is based on the 3-D convolution feature extractor. This architecture can generate multi-channel information from consecutive video frames and then separate convolution and down-sampling operations on each channel. Finally, the information of all channels is combined to get the final feature description.[10;11]

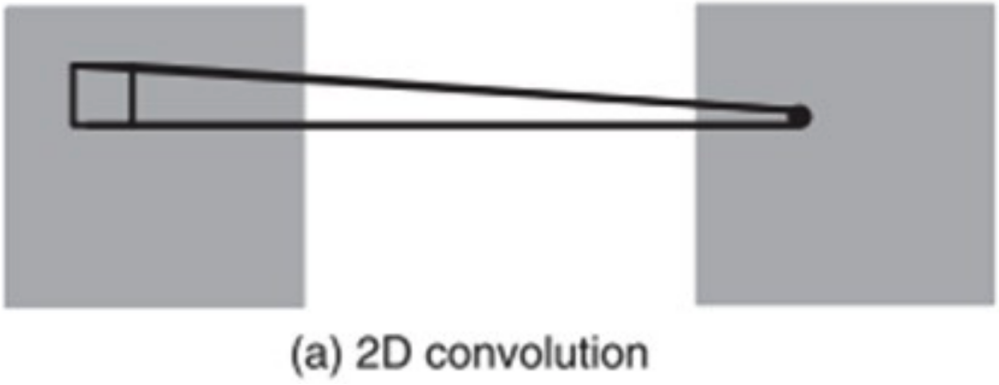

Figure 2.1: *2-D*[10]

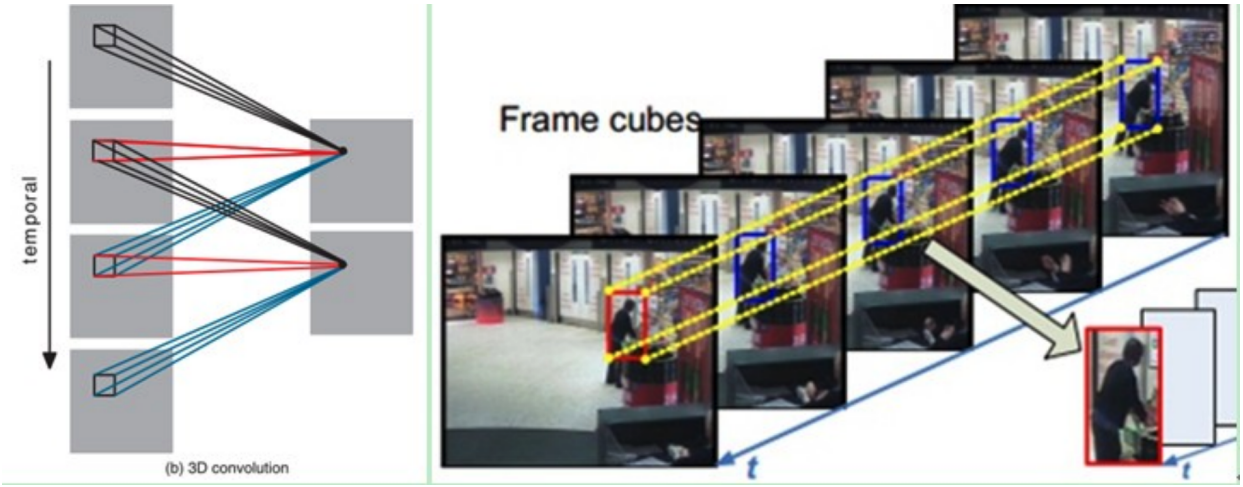

Figure 2.2: *3DCNN*[10]

3-D convolution forms a cube by stacking multiple consecutive frames and then using a

3-D convolution kernel in the cube. In this structure, each feature map in the convolutional layer is connected to multiple adjacent consecutive frames in the previous layer, thus capturing motion information. For example, in Figure 2.2, the value of a certain position of a convolution map is obtained by convolving the local perception of the same position of three consecutive frames of the previous layer.[10]

2.1.2 Inflated 3-D ConvNet

In Figure 2.3, the evolution of I3D is outlined, making the advantages of I3D more easily understood.

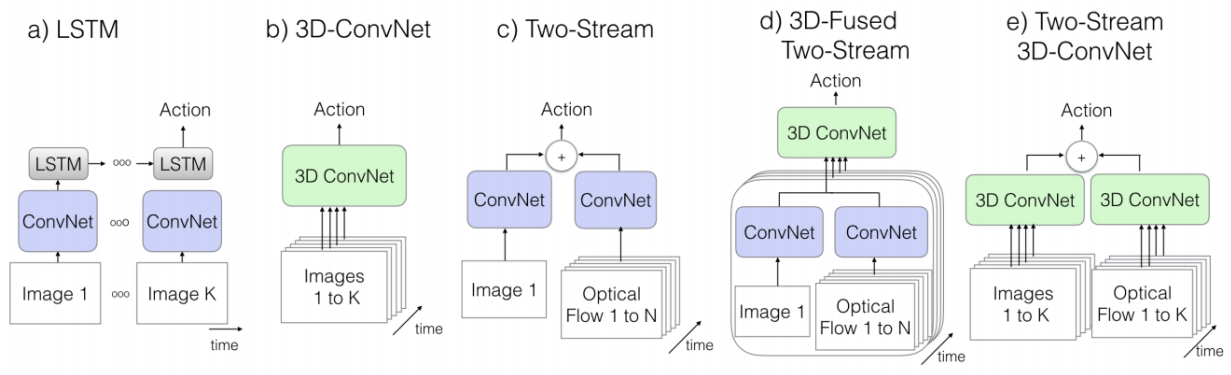

Figure 2.3: *I3D*[8]

Since 3DCNN emerged, two-stream networks have also been proposed. This method divides the entire model into Spatial Stream Convnet, which uses a single frame to capture features, and Temporal Stream ConvNet, which uses multiple computed optical flow frames. The features extracted from the two networks are fused at the end.[12] The features can be fused by adding a CNN network after the two-branch model, which improves accuracy and achieves end-to-end training.[13]

The most critical and significant step of optimizing the model in I3D is next change. First, researchers used ImageNet on the pre-trained frame while also using 3-D convnets to extract the temporal feature of the RGB stream. Finally, the optical-flow stream improves network performance. To turn the temporal dimension from N*N filter into N*N*N, researchers processed N*N filter N times with a pre-trained 2-D convnett.[8]

2.2 The Generative Adversarial Network (GAN)

The basic principle of GAN is actually very simple. In Figure 2.4 are two networks: G (Generator) and D (Discriminator). In the training process, the goal in generating network G is to generate real pictures as frequently as possible to deceive discriminating network D. The goal of network D is to identify the pictures generated by G from the real pictures. Thus, G and D constitute a dynamic "adversarial process." [3;14]

Generator G continuously strengthens its own capabilities to generate samples more and more similar to the real sample. That is, discriminator D increasingly fails to distinguish if the sample is real. At the same time, discriminator D also improves its ability to identify images. The above process continues until the discriminator cannot distinguish whether the received sample is real or generated.[3;14]

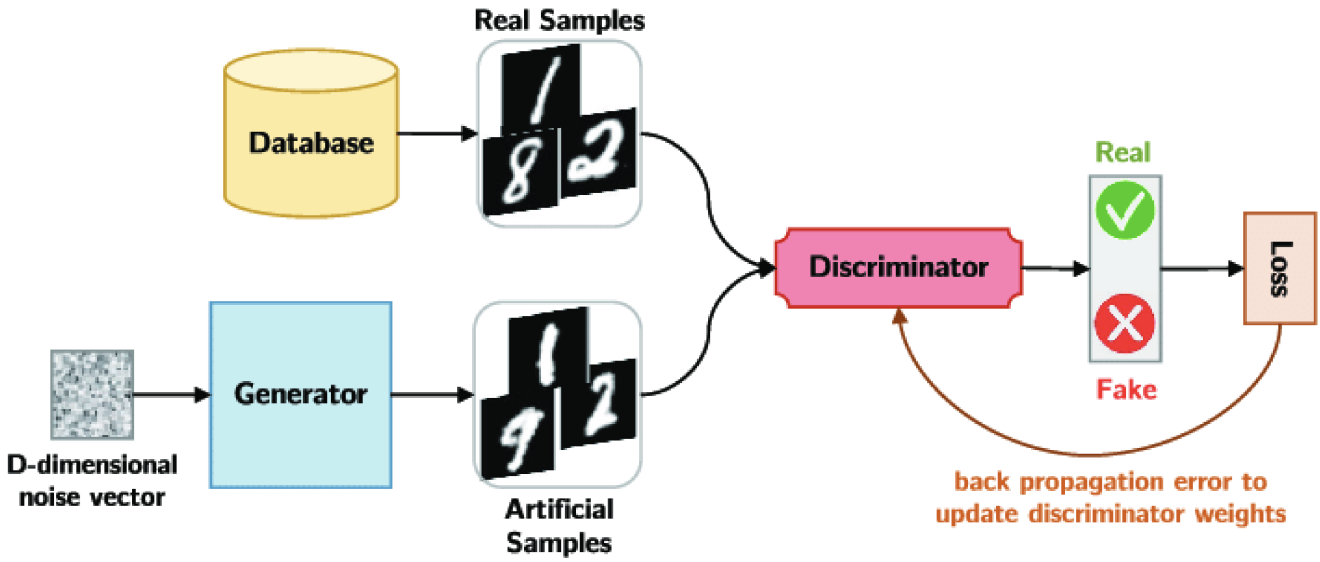

Figure 2.4: *Structure of GAN*[15]

The mathematical formula is shown below:

$$\min_G \max_D V(D,G) = \min_G \max_D [E_{x\sim p_{data}(x))}[logD(x)] + E_{z\sim p_z(z))}[log(1-D(G(z)))]] \quad (2.1)$$

To explain:

- The formula comprises x, representing the real picture; z, representing the noise input to the G network; and G(z), representing the picture generated by the G network.

- D(x) represents the probability that the D network judges whether the real picture is real (because x is real; for D, the closer this value is to 1 the better). D(G(z)) is the probability that the D network will judge the picture generated by G.

- Purpose of G: As mentioned, D(G(z)) is the probability that the D network will judge the picture generated by G real, while G hopefully generates a picture closer to the real image. In other words, G wants D(G(z)) to be as large as possible and V(D, G) to become smaller at the same time. Therefore, we see that the front mark of the formula is min_G.

- Purpose of D: As the ability of D to recognize a picture as real becomes stronger, D(x) should become larger while D(G(z)) becomes smaller. That means V(D,G) will become larger. Therefore, the formula for D is to maximize (max_D)

Chapter 3

Methodology

3.1 Tackle Video Classification: I3D

Section 2.1.2 describes the basic structure of I3D. This chapter will outline some of the parameter settings of I3D in this research. As in Carreira's research[8], the I3D used ImageNet to pre-train the frames. In the headbone of I3D, clip length was set at 32, frame interval set at 2, and the number of clips set at 1, which means that every time the machine collects 1 clip to train, there are 32 frames in the clip. Thus, the input shape is [1, 32, 224, 224, 3]. The loss function is common "CrossEntropyLoss". For the backbone of model, we chose ResNet50, which passes 4 blocks, each block with either a 3, 4, 6, or 3 bottle neck.[16] For evaluation, accuracy was computed from top 1 to top 5 and mean_class_accuracy was computed for each iteration.

3.2 Data Augmentation Techniques

For data augmentation,we used the following 4 GANs models, each with different styles and characteristics:

3.2.1 DCGAN

The full name of DCGAN is deep convolutional Generative Adversarial Network. It is an unsupervised representation learning network, as the name implies. DCGAN replaces the fully connected neural network in the original GAN with a convolutional neural network in the generator and discriminator feature extraction layer, using the DCEloss function and the Adam optimizer.[4;17]

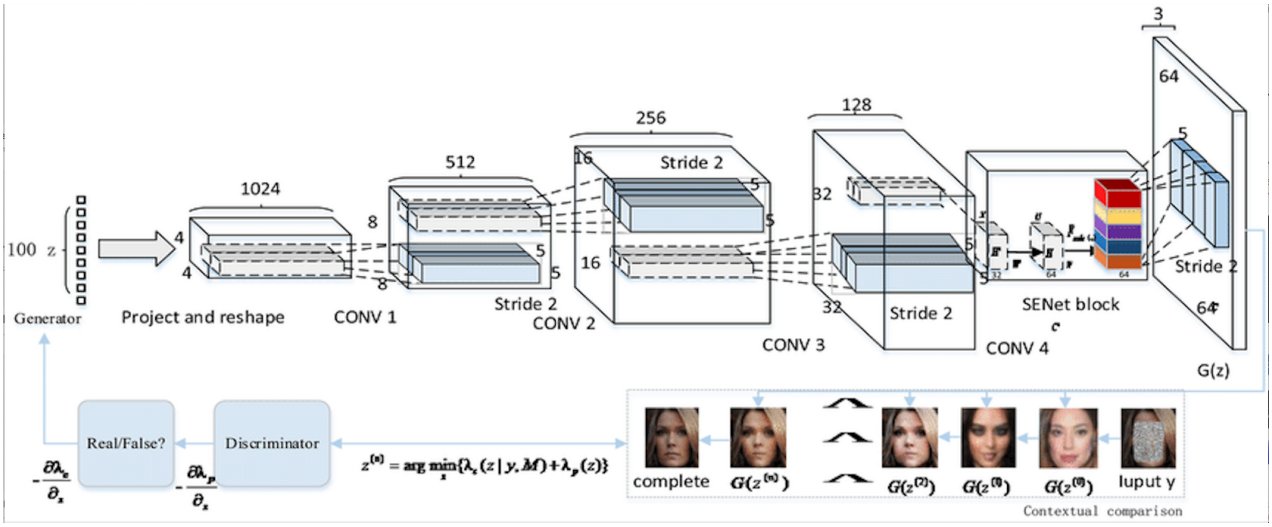

Figure 3.1: *Generator structure of DCGAN*[18]

DCGAN differs from the traditional GAN in the following ways: network:[4;17]

- Both the generator and discriminator of DCGAN abandon the pooling layer of CNN; the discriminator retains the overall architecture of CNN, and the generator replaces the convolutional layer with a fractional-strided convolution or convolution transpose. Please see Figure 3.1.
- The Batch Normalization (BN) layer is used after each layer in the discriminator and generator.
- The fully connected layer is removed, replaced by the global pooling layer.
- The output layer of the generator uses Tanh activation function, and the other layers use RELU.

- All layers of the discriminator use LeakyReLU activation function.

3.2.2 LSGAN

LSGAN's full name is least square generative adversarial networks. The generator network and the discriminator network of LSGAN use convolution and deconvolution like DCGANs, and they use the Adam optimizer, but they do not use the fully connected neural network. The main difference between LSGAN and a traditional GAN is that the cross-entropy loss function is replaced with a least-squares loss function. The image quality generated by a traditional GAN is not ideal, and the training process is unstable. Traditional GANs use sigmoid cross entropy as the loss function of the discriminator. The small square loss function used by LSGANs penalizes samples that are far from the decision boundary, and the gradient of these samples is the direction of the gradient descent.[19]

3.2.3 WGAN-GP

WGAN-GP is a improved model of Wasserstein GANs. It uses convolution and deconvolution in the generator and discriminator network structure the same as DCGAN, and does not use a fully connected neural network. WGAN-GP improves on WGAN. In one article,[20;21] WGAN-GP exemplifies the problem with WGAN; that is, WGAN directly uses weight clipping when faced with Lipschitz constraints. WGAN checks whether the absolute value of all parameters of the discriminator exceeds a threshold every time the parameters of the discriminator are updated. Thus, the discriminator cannot discriminate between two samples with few differences by ensuring that all parameters of the discriminator are bound during the training process, which means a WGAN indirectly recognizes the Lipschitz restriction. In actual training, the discriminator loss should enlarge the score difference between true and false samples as much as possible, but weight clipping independently limits the value range of each network parameter. In this case, the optimal strategy is to make all parameters as extreme as possible, either the maximum value or the minimum value, so the parameters of the discriminator are almost all concentrated on the maximum and minimum. In sum-

mary, WGAN-GP differs from WGAN in three main ways: 1) weight clipping is replaced by gradient penalty, 2) Gaussian noise increases on the generated image; 3) the optimizer uses Adam to replace RMSProp.[21]

3.2.4 StyleGAN

First, we explain PG-GAN (PROGAN) because it is an important breakthrough in the structure of GANs models. Then we will introduce StyleGAN.

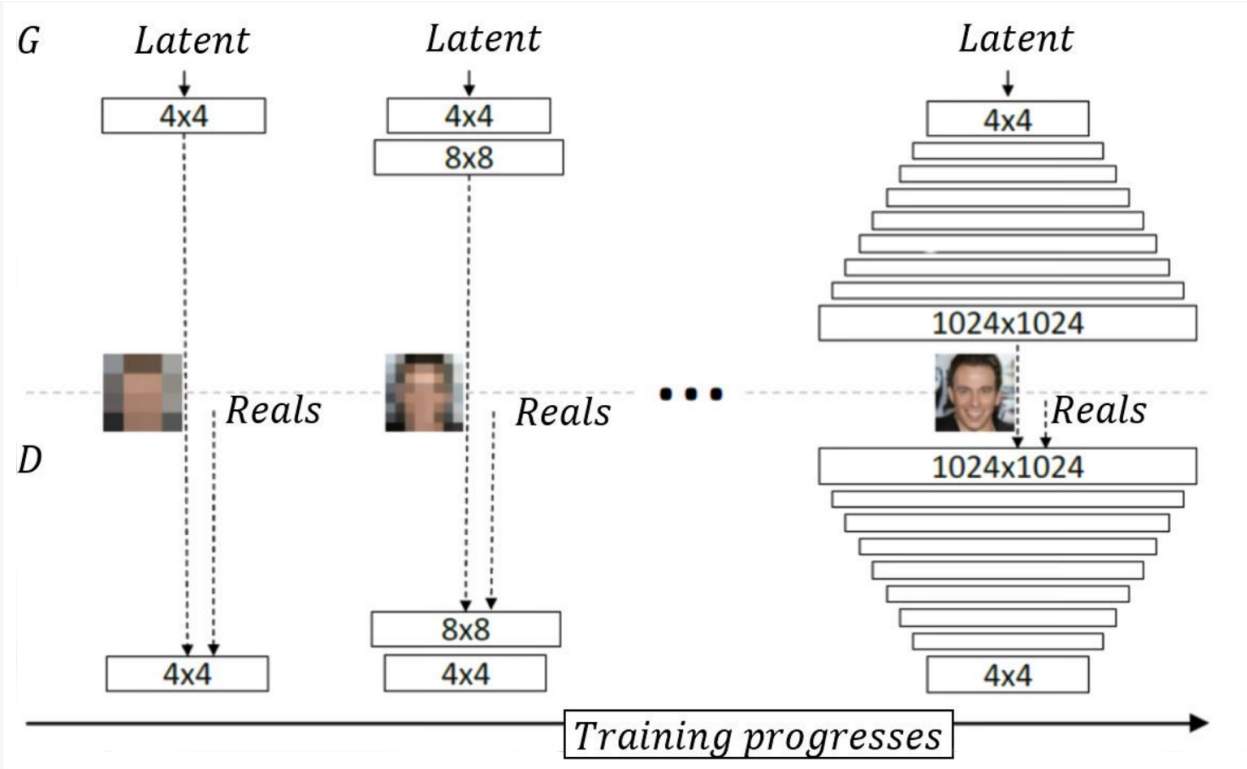

Figure 3.2: *Generator structure of PG-GAN*[22]

PG-GAN can generate samples of 1024 pixels. Obviously, building a mapping network G from latent code to 1024x1024 pixels samples using pure GAN is difficult. The procedural training method uses not one step, first trying to generate low-resolution or low-quality images, and then continuously increase the resolution or details for generated images based on layers that are incrementally added to G and D. The structure of PG-GAN is shown in Figure 3.2. This way of changing the network and procedurally generated image is similar to

human intuition and easy to understand. However, how PG-GAN can achieve such amazing effects is inseparable from some of its capabilities and detailed processing.[22;23]

Next, we discuss the basic structure of StyleGAN (Figure 3.3), its advantages, and how it achieves style transfer on generated images. StyleGAN has 3 neural networks: G_mapping; G_synthesis and discriminator D; G_mapping and G_synthesis, all of which constitute the main generator.[24]

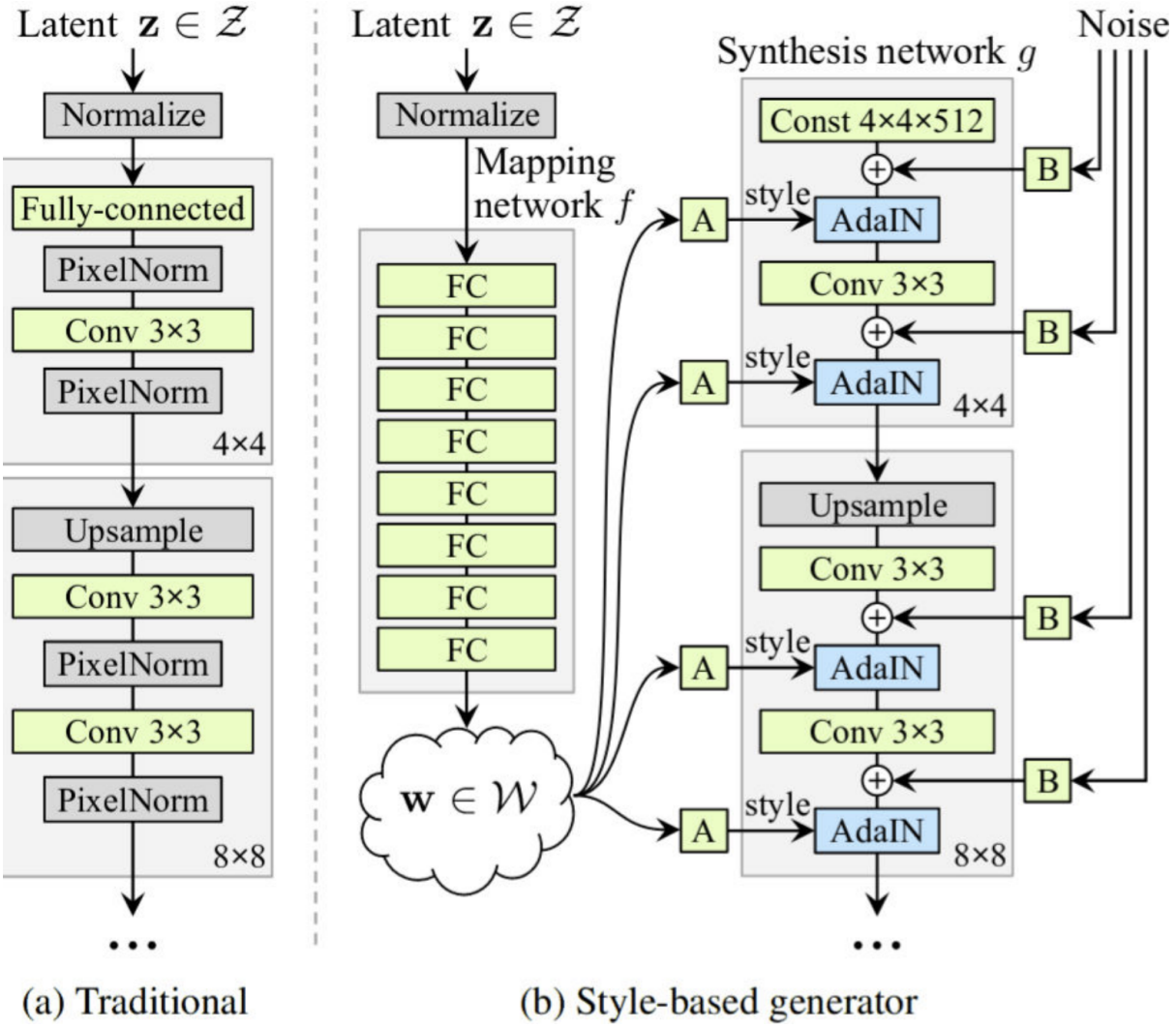

Figure 3.3: *Generator Structure of StyleGAN*[24]

Removing Traditional Input and Mapping Network: Traditional models use latent code as the initial input of the generator, so no more steps for latent code are included in the following computation. Thus, the latent code will have a weaker effect when the layer

becomes deeper. In StyleGAN, latent Z is input into G_mapping, which is a fully connected network, and then output W of the mapping network is mapped into each layer of the synthesis network. The generator starts with a learnable constant input, and the hidden code adjusts the style of the image in each convolutional layer, thereby directly controlling the intensity of image features at different scales.[24]

Style-based (AdaIN): The math function of AdaIN:

$$AdIN(x_i, y) = y_{s,i}\frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i} \tag{3.1}$$

First, each feature map xi (feature map) is normalized independently $(x_i - u(x_i))/sigma(x_i)$. Each value in the feature map is subtracted from the mean value of the feature map and then divided by the variance. Second, a learnable affine transformation A (fully connected layer) transforms w into the translation and scaling factor of AdaIN in style $y = (y_{s,i}\,, y_{b,i}\,)$. Finally, for each feature map, the translation and scaling factors learned in style perform scale and translation transformation. This method results not only in a capture of advanced attributes (e.g., face, pose), but also randomly changes style (e.g., freckles, hair).[24]

Blur: A large number of blurring operations, known as blur, are used in StyleGAN's generator and discriminator network. Blur uses a convolution kernel to perform convolution operations on each channel. This differs from normal convolution. StyleGAN uses the leaky_relu function.[24]

Chapter 4

Experimental Design

4.1 Tackle Video Classification: I3D

I3D uses two pre-trained (2-D) models through ImageNet and inflates the 2-D network into a 3-D one by extruding it into a temporal dimension. Of its constituent 2-D models, one is thus used for RGB data processing and one for optical flow data processing. It can extract spatial and temporal features from video data for action recognition and capture motion information in space and time dimensions.

4.1.1 Data Set Preparation

The data set was collected by Dr. Scott Dietrich, who recorded 190 short videos of athletes completing the tackle action. The action of tackle includes a series of motions: running towards an object, carrying an object, throwing an object. The environment in the data is complex, and the semantics are redundant.

To keep the model from failing to grasp the main objective, two methods based on SATT were used to label the data.

- First, because the strike zone is the most dangerous part of the whole tackle action, players need to rush towards the object and lift it. In computing linear regression of 6 scores, Strike Zone is revealed as the SATT score component most relevant to the

overall total. Thus, Strike Zone is the only indicator of labeling, thus simplifying the model's task. In this way, 57 videos were labeled risky and 133 videos were labeled safe.

- Second, to follow the original evaluation criteria of the SATT model, the data was labeled based on the total SATT (scores over 10 were considered safe). We had 115 safe data samples and 75 risky data samples.

The proportion of training and testing is 7:3.

4.1.2 Data Set Problem

The tackling videos posed some unique challenges for deep learning, as with most sample data sets taken from real life. First, duration differed greatly from video to video. Some were around 10 seconds long, and some were a minute long. In addition, the videos had some noise with people other than players walking around. In addition, due to the angles, an athlete's movements could be obscured and misunderstood. For example, when players make contact with the front of an object, the movement will be labeled incorrect if the player's head is tilted at a specific angle.

4.2 GANs

Regardless of the labeling method used, the precision and accuracy of video classification are thus far poor overall. This result may improve if the model can focus on important frames, but only a few important frames occur in each video; longer videos had more important frames than short ones. In addition, the total data set is limited, and the data is imbalanced. Thus, the most reasonable method is to expand the number of important frames using GANs.

Figure 4.1: *Three data sets of GAN*

4.2.1 Data Set Selection

As shown in Figure 4.1, three movements have been designated for GANs research. The three movements not only can detect whether the athlete's state is safe, but also whether players tackle efficiently. In addition, the movements are diverse, including the position of the athlete's torso and the relationship between the athlete and the reference object. The bias of different models can be reduced by using a variety of data sets. GAN models often require many samples. We used approximately 30 images for each data set.

- In the first action, the player rushes toward the object. If head and torso are up, the player is safe.
- The second action shows a player making contact with the object. Players who initiate contact with the head instead of the shoulder are at risk.
- The third action shows the player throwing the object. Players who hold the object tightly are safe.

4.2.2 Data Set Preparation

The laboratory computer has limited computing power, so the images must be resized from 1024*1024 to 64*64. StyleGAN, however, can process up to 128*128. Blindly shrinking

images would be unwise because some details would be lost, and it could be difficult for the model to learn details. In this respect, StyleGAN is better than other GANs.

During training, differently labeled images of the same data sets (data set 1 labeled Risky and data set 1 labeled Safe) are combined and both treated as original data. In this way, we could test whether GAN models could capture specific details used to label images. In addition, we could check GAN models for the ability to generate differently labeled images. In other words, the user can check the diversity of models.

4.2.3 Model Setting

The original internal network structure and loss function of all models is used. For StyleGAN, the batch size change with the different layers: $4:32, 8:32, 16:32, 32:16, 64:8, 128:4$. For other the three GANs, the batch size is 4.

4.2.4 Evaluation

Because the test results were complex, during data collection, only data set 3 was used to calculate various indicators. Data set 3 was chosen because it is a representative data set, and its performance was better than the other two data sets for every GAN model. Furthermore, we ignored DCGAN's data calculation because the results were vague.

Observing Detail with the Naked Eye: The easiest and fastest evaluation method is to observe the generated images with the naked eye and compare images generated by different models in the same time iteration. Execution time is another indicator for judging GAN models.

Inspection Details by Image Classifier: An image classification model can objectively judge the clarity of the generated image. First, the model is trained using the labeled original images from resnet50. Then, testing the images generated on the resnet50 model

can represent clarity. The generated images are clearer, and testing accuracy is higher.

SSIM: To compare two images for structure similarity required three aspects: luminance I (x, y), contrast c (x, y), and structure s(x, y). The final similarity between x and y is a function of these three aspects: S(x,y) = f(l(x,y),c(x,y),s(x,y)).[25]

- Calculation of Luminance:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

- Calculation of Contrast:

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

- Calculation of Structure:

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x + \sigma_y + C_3}$$

FID: The Frechet Inception Distance Score (FID) is used to calculate the distance between the feature vector of the real image and the generated image. FID measures the similarity of two groups of images using their statistical similarity to the computer vision features of the original image. This visual feature is calculated using the Inception v3 image classification model. A lower score means the two sets of images are more similar, or the more similar the statistics of the two the lower the score; the FID score in the best case is 0.0, which means that the two sets of images are the same. FID scores are used to evaluate the quality of images generated by generative adversarial networks, and lower scores correlate with higher quality images.[26–28]

$$FID(x,g) = \|\mu_x - \mu_g\| + Tr(\Sigma_x + \Sigma_g - 2\sqrt{\Sigma_x\Sigma_g}) \tag{4.1}$$

In general, the FID sends the samples from the generator and the samples from the discriminator to the classifier (Inception Net-V3 or other CNN). Then, the FID extracts the abstract features of the middle layer of the classifier and assumes the abstract features conform to the multivariate Gaussian distribution. Finally, FID estimates the mean u_g

and variance sum_g of Gaussian distribution of generated samples, as well as the training sample u_data and variance sum_data and calculates the Freche distance of the two Gaussian distributions. This value is the FID.[26–28]

Chapter 5

Results

In this chapter, the results of the research will be presented. The two parts comprise the results for I3D and the GANs.

5.1 I3D

Because of the inadequacy of the data sets, the accuracy peak came at the third epoch. After the third epoch, although the accuracy of the top 1 improved, the mean class accuracy did not, and the resulting value was approximately 0.5. Whatever the label used, the I3D performance was not ideal. For the first label, Strike Zone, the accuracy increased from 0.4986 to 0.70175. For the second label, based on total SATT score, the accuracy increased from 0.4521 to 0.6892 (see Table 5.1).

Table 5.1: *Accuracy of I3D*

	Labeling by Strike Zone	Labeling by Total SATT
1	0.4986	0.4521
2	0.54386	0.5287
3	0.70175	0.6892

5.2 GANs

Among four GANs, a style-based generator has special advanced structure, it modifies the input for each level without changing other layers, thus controlling the visual features represented by the level. Hereby, the result of StyleGAN is significantly greater than the result of other three GANs.

5.2.1 Observed Detail with the Naked Eye

DCGAN:

For Data Sets 1 and 3, the athlete's torso generated in the images can be identified after 200 epochs, but the images are blurred. For Data Set 3, the model showed relatively better results after 800 epochs.

Figure 5.1: *Results of DCGAN*

In Figure 5.1, each result from the three data sets have learned the features of the original images, even though the images are blurred. The athlete's general actions and the reference object can be identified, but details like face and hands are obscured.

LSGAN:

Many noise points show up in the results of LSGAN before 370 epochs, so the generated images are very blurry. After 800 epochs, the results are clearer.

In Figure 5.2, the images generated by LSGAN are blurry, and most of the semantics in the images have been lost. The naked eye cannot identify the athlete's body parts. However,

Figure 5.2: *Results of LSGAN*

the background of generated images is similar to the background of original images.

WGAN-GP:

For Data Sets 1 and 2, the results of WGAN-GP is definitely better than the results of either LSGAN or DCGAN. In the picture, you can see not only the athlete's torso and limbs, but also the details of the movement. The leftmost picture is labeled risky because it is apparent that the player's head and eyes are angled down. The results show less noise. The results from Data Set 2 are generally blurrier than Data Set 1.

The results for Data Set 3 are ambiguous; it is difficult to see where the player is. The results of WGAN-GP are shown in Figure 5.3.

Figure 5.3: *Results of WGAN-GP*

StyleGAN:

Of the four GANs, StyleGAN performed the best. Although it took a long time to train, StyleGAN generated the clearest images with the most detailed content. The samples it generated are more realistic that for the other GANs, and the images are more diverse and interesting. In the results, different backgrounds, different poses of athletes, different colors of clothes and helmets, and even different patterns on the cloth appeared.

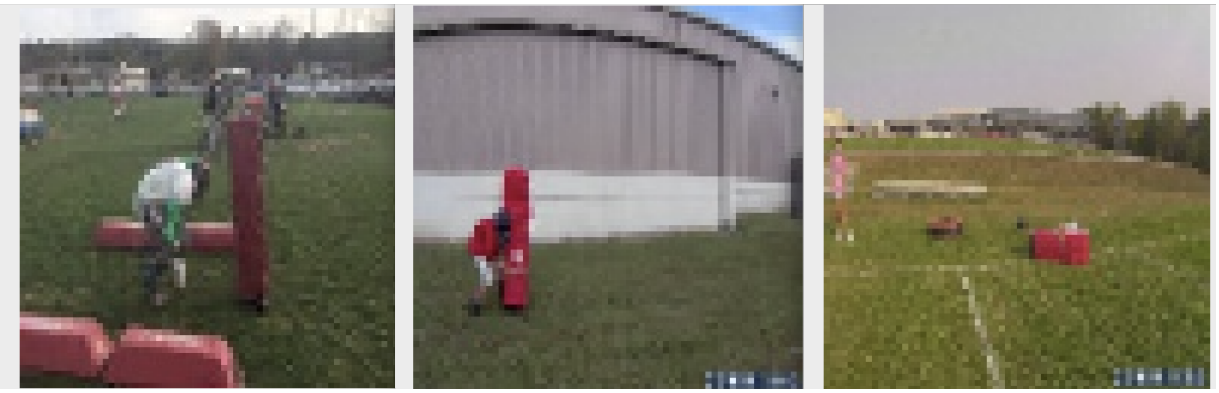

Figure 5.4: *Results of StyleGAN*

The results generated by StyleGAN can capture the details to be labeled. In the Figure 5.4, the label for the first image is risky because the player's head and eyes are down, but the second image is labeled safe because the player uses his shoulder to contact the object and the third image is labeled safe because the player's arm holds object.

5.2.2 Inspection Details by Image Classifier

Image classifiers are better at objectively detecting the clarity of generated images than human eyes. Thus, this detection method was used in the model trained on the original data to test data generated by the different models. The results are in Table 5.2.

The data in this table is unconvincing because the accuracy and F1 score of StyleGAN should be the highest based on its clear images, but such unreasonable results occur for several reasons. On one hand, the test data of LSGAN and WGAN-GP were limited because

Table 5.2: *Results of Different GANs*

	ACCURACY	F1
Original data	0.818	0.811
LSGAN data	0.7619	0.7826
WGAN-GP	0.8071	0.79
StyleGAN	0.65	0.78

their batch size was 4 with only 4 images generated for each training. On the other hand, their generated images are not clear enough to be properly labeled.

To obtain data with representative clarity, we need another feasible test method. We suggest, as part of the first step, the generated images should be added to the original data and training the model again. The next step would be to check whether expanding the data sets improves the accuracy of the new data sets.

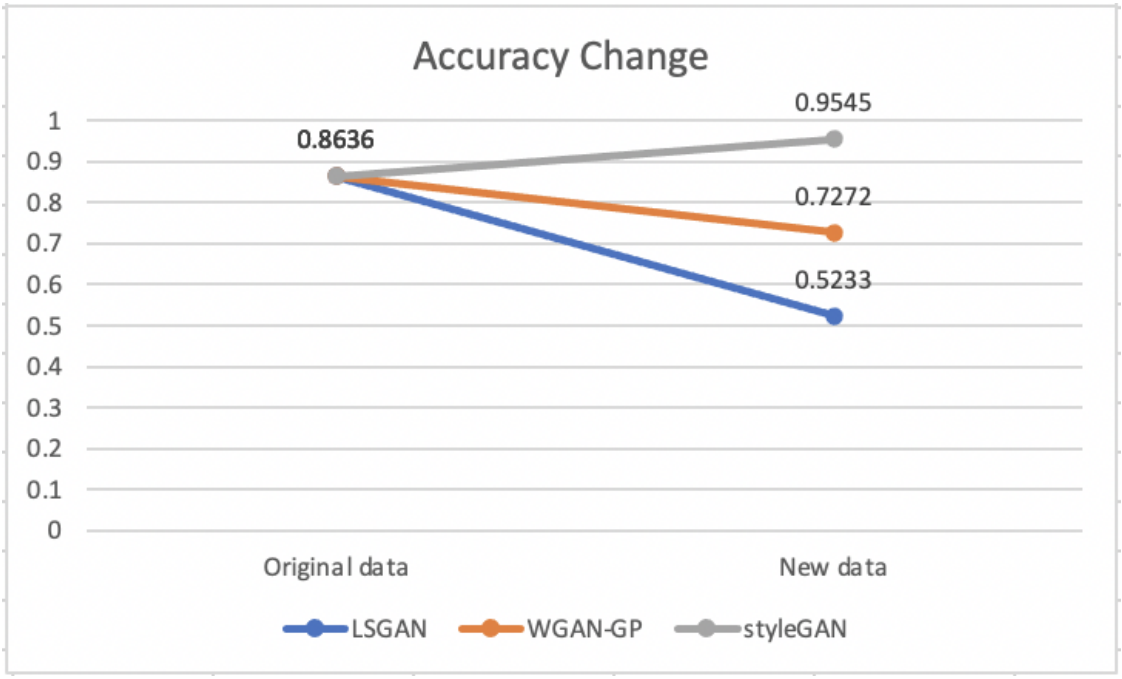

Figure 5.5: *Accuracy Change*

This change improved the accuracy of the new dataset; adding StyleGAN generated images improved accuracy from 0.8636 to 0.9545 (see Figure 5.5). In addition, the F1 score of new data sets improved from 0.823 to 0.947 with added images generated by StyleGAN (see Figure 5.6). However, the performance of other three GANs was worse after adding

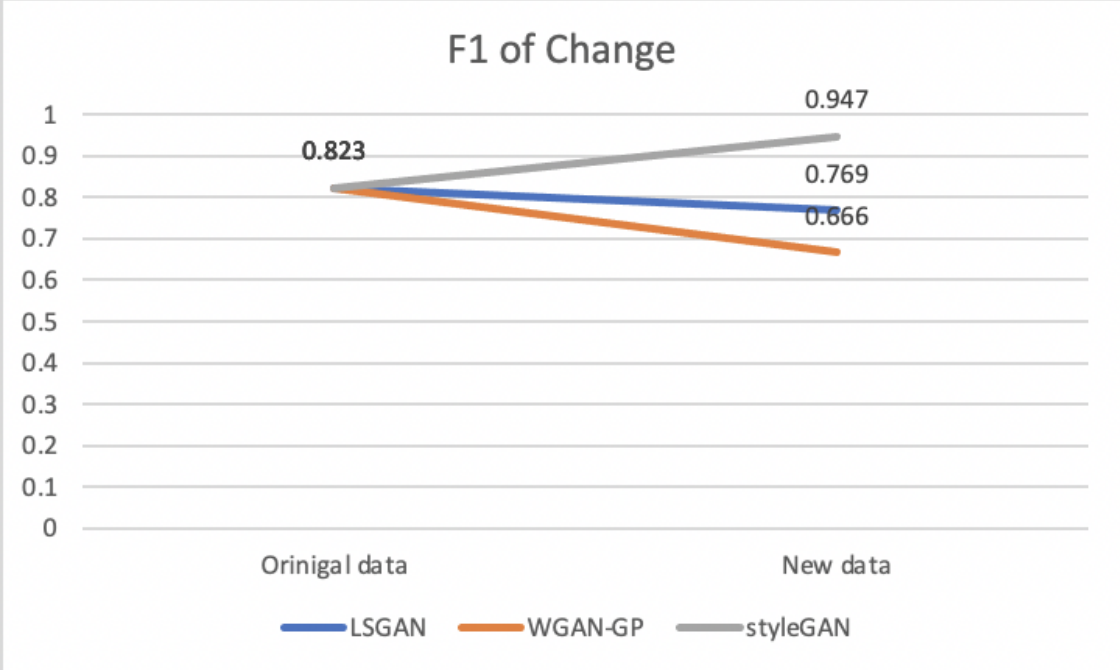

Figure 5.6: *F1 Change*

images generated by each GAN. In conclusion, the images generated by StyleGAN are clear enough to be identified by image classifier, and models learned the most important feature in the images, which could then be labeled correctly. The performance of the model improved by expanding the data set. However, adding images with unclear label features will reduce the performance of the model.

5.2.3 SSIM

Figure 5.7 shows the maximum SSIM value of all three GANs, which are all lower than 0.2, indicating that the images are all generated by GANs, not by changing the original images. Of the three scores, the highest SSIM score was for StyleGAN, partly because the images generated by StyleGAN are clear. The images from StyleGAN also can keep most of the features of the original images.

Because SSIM is a calculation value based on each pixel, to reduce deviation, the mean SSIM value was compared. Figure 5.8 shows the highest mean SSIM score is still StyleGAN at 0.1799.

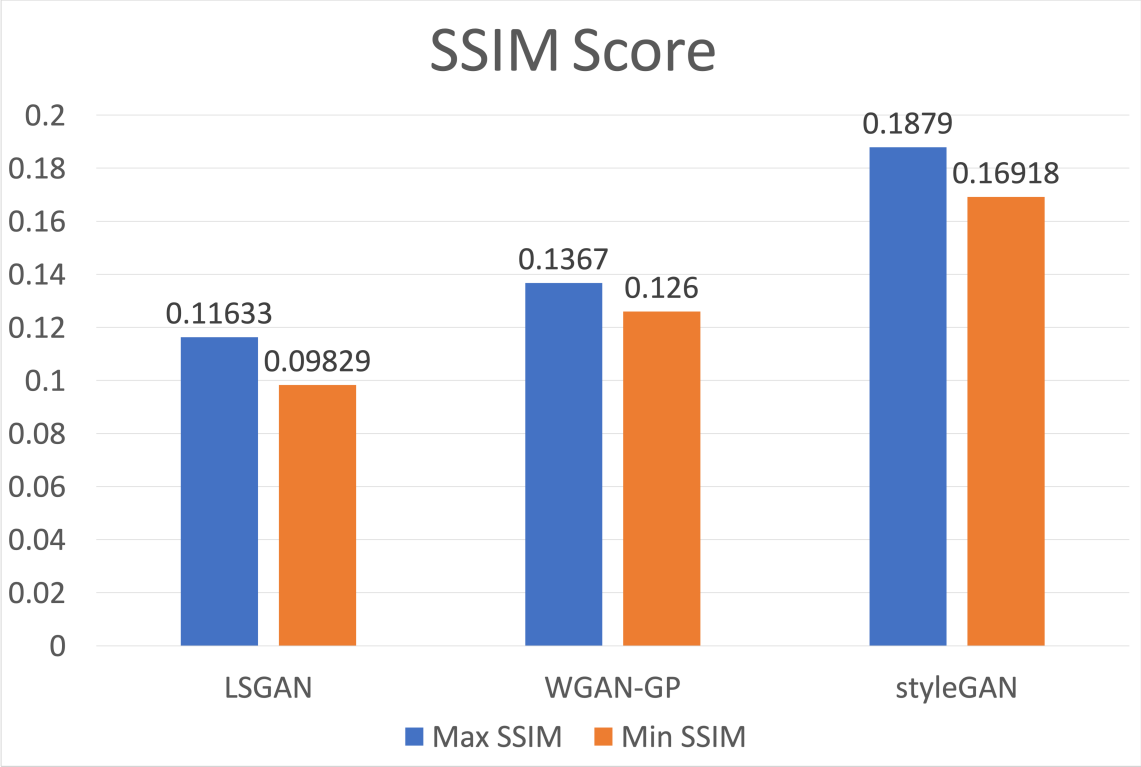

Figure 5.7: *Min SSIM scores and Max SSIM scores*

5.2.4 FID

For common distributions (such as Gaussian distribution), when the distribution type is determined, as long as the mean and variance are known, the distribution can be determined. We assume that the generated image and the real image also obey a similar distribution. If the mean and variance between them are similar, the generated image is likely to appear more real. A lower FID means the generated distribution is closer to the real image distribution.

The FID of StyleGAN is the lowest at only 260 (Figure 5.9). This means the images generated by StyleGANs are the most realistic of those for the three GANs.

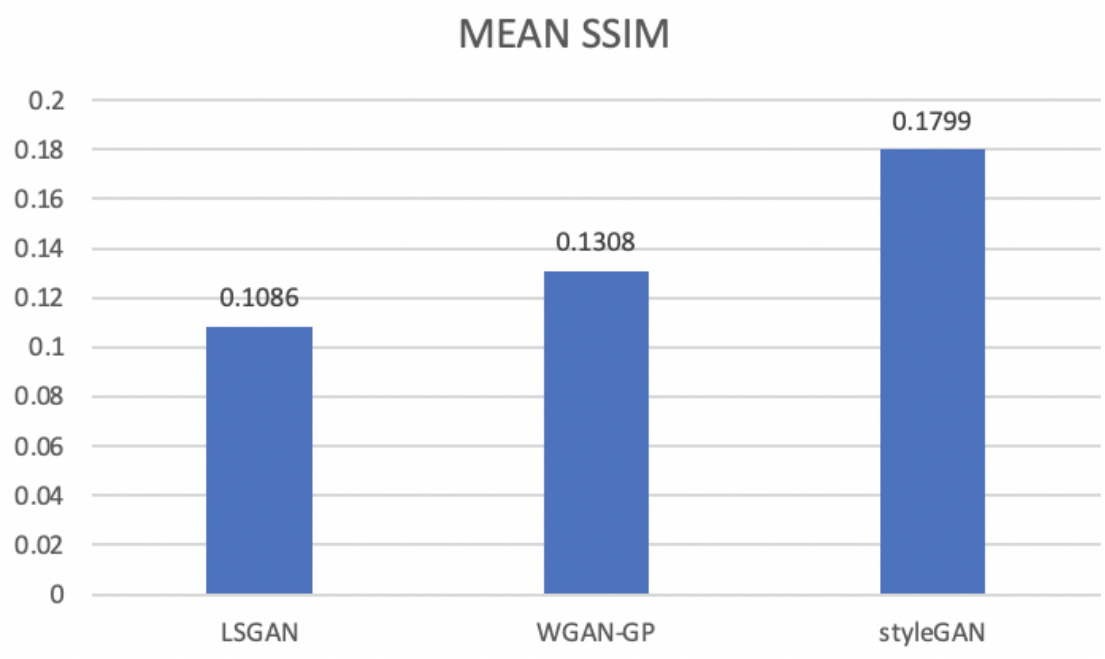

Figure 5.8: *Mean SSIM Scores*

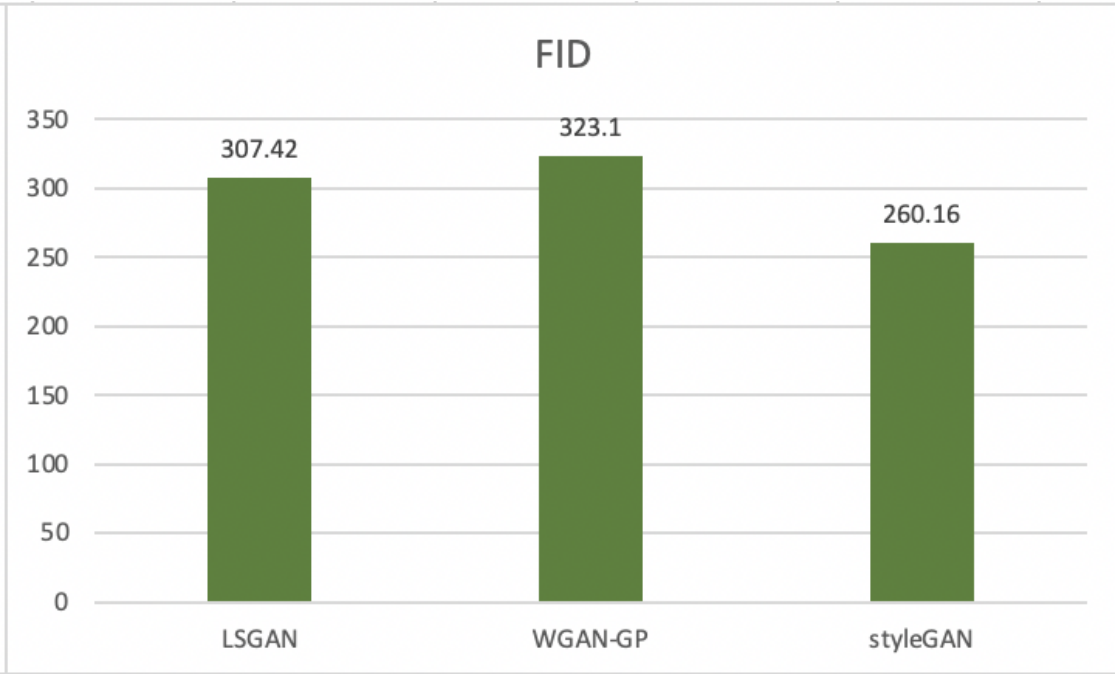

Figure 5.9: *FID Scores*

Chapter 6

Conclusions and Future Work

6.1 Summary and Conclusions

Research on athlete safety is constantly evolving, particularly in research focusing on finding ways to avoid head injury and prevent potential adverse contact coming at the start of development. The technical accuracy of screening player tackles may be an important step in keeping athletes safe. Therefore, because tackling is a source of many injuries, accurately identifying and correcting risky tackle techniques is imperative. Though Schussler SATT assessment comprehensively monitors an athlete's body parts to detect the quality of a tackle, limited research exists on deep learning focusing on tackling behaviors.[29]

In this experiment, deep learning used to model football safety was researched. The main task was to identify risky and safe tackle actions. I3D, a resend video classification model, was selected as the model for supervised learning. Both labeling methods are based on SATT.

At the conclusion of the experiment, it was clear that the I3D does not perform well in handling tackle videos, no matter what kind of labeling method was used. The model did not improve after three epochs of training, but a possible explanation of this unsatisfactory result is the limited data quantity. However, this is an easily solved problem. We need more data sets. The more difficult problem of the I3D is the complexity of the video and the

excess of semantics in the video. Unless the video focuses specifically on capturing a series of actions, a single label only can be attached to a video. This label is a threshold for the total score as based on 6 actions. Because of this complexity and because the models could not learn many semantics with a small data set, we can understand less than ideal results. This problem can be addressed by creating a much larger data set that covers all possible samples of each total score.

Manipulating the video data is difficult, so one possible solution is to focus on important frames. This can solve the problem of inadequate data and ensure the model can grasp key points. The first step is to expand the data derived from important frames and solve the data imbalance. Therefore, in the second part of the experiment, we used GANs to generate further images. DCGAN, LSGAN, WGAN-GP, and StyleGAN were chosen to expand three different data sets.

By comparing these expanded data sets, see 4.2.4, whether with the naked eye or with a classifier, we concluded the images generated by StyleGAN are clearer than the generated images of other models with less noise. The accuracy of image classification improved from 0.86 to 0.95 when StyleGAN-generated images were added. In addition, the FID score of StyleGANs was the lowest, which means that the images generated by StyleGAN are more realistic. The SSIM score of the StyleGAN was highest of those for the three types of GANs I applied but still less than 0.2, which means the images generated by StyleGAN are clearer and keep the features of the original image. Lastly, StyleGAN1 recognizes style migration on the dataset. The images it generated were diverse, no longer limited to realizing generating images with different labels, but generating a collection of different clothes and different helmets that do not appear in the original images.

Overall, StyleGAN performed excellently on the tackle data. StyleGAN is more suited to this dataset than other GANs because of the large image size and the complexity of the image semantics. The generated data can be adequately prepared for further experiments in the future. The image size of StyleGAN, 128*128, dominates in machine recognition.

6.2 Future Work

In this experiment, the video classifier I3D performed poorly on the tackle data set. However, its performance should improve if long videos can be cut to 6 judgment standards and semantically clear videos that retain only one judgement standard for each small video. Thus, the next task is to work on short videos with I3D. If I work on this way, another important question would be raise up: automatically video-trimming. Trimming video to short videos would be increase the user's workload and destroy the automatic function of deep learning.

Another possibility for future research is to apply generated images to video recognition. Adding the generated images without destroying the sequence of video frames is an important prerequisite because the advantage of video recognition that can learn temporal features must be preserved. There is another way to use generated images that changing video classification question into image classification question. The first step is to choose the important images from frames and augment images data set by adding generated images, then classify and detect those important image based on SATT criteria.

As the result shown in SSIM result 5.2.3, the SSIM scores of StyleGAN result are higher than other SSIM scores of other models whatever Min_SSIM, Max_SSIM, or Mean_SSIM. Dr. Hsu remarked that for this application, 'the idea that SSIM could be "too high" suggests that an orthogonal diversity score might be useful'. He suggested using multi-objective optimization. In the future, I will work on it and understand the problem.

For the result of video classification, Dr. Munir mentioned that Resnet 50 and multilayer LSTM got 87% accuracy on some football tasks and he sent me some references about other soccer events classifications. Many effective methods that segment videos and identify the major soccer or football events showed in papers. Hidden Markov models (HMMs) are used to segment long video into small semantic units,[30] which inspired me on trimming tackle videos. C3D is used to learn spacial and temporal features of soccer events, and its result has high efficiency.[31] Learning the advantages of those papers can improve the performance of tackle video classification.

Bibliography

[1] T. Dompier, R. Lynall, E. Wasserman, K. Campbell, and Z. Kerr. Comparison of concussion injury mechanisms and contact types between youth, high school, and college football players. 51(6):S202.

[2] B. J. Smart, R. S. Haring, A. O. Asemota, J. W. Scott, J. K. Canner, and J. K. Neijim. Tackling causes and costs of ed presentation for american football injuries: A population-level study. 37(7):1198–1204, 2016. https://doi.org/10.1016/j.ajem.2016.02.057.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[4] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019.

[6] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017.

[7] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.

[9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[10] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. doi: 10.1109/TPAMI.2012.59.

[11] Dhananjay Kumar, Priyanka T, Aishwarya Murugesh, and Ved P. Kafle. Visual action recognition using deep learning in video surveillance systems. In *2020 ITU Kaleidoscope: Industry-Driven Digital Transformation (ITU K)*, pages 1–8, 2020. doi: 10.23919/ITUK50268.2020.9303222.

[12] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos, 2014.

[13] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[15] Octavio Loyola-González. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113, 10 2019. doi: 10.1109/ACCESS.2019.2949286.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[17] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

[18] Fangyan Zhang, Xin Wang, Tongfeng Sun, and Xinzheng Xu. Se-dcgan: a new method of semantic image restoration. *Cognitive Computation*, 05 2021. doi: 10.1007/s12559-021-09877-y.

[19] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks, 2017.

[20] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[21] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017.

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018.

[23] Zhaoyu Zhang, Mengyan Li, and Jun Yu. D2pggan: Two discriminators used in progressive growing of gans. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3177–3181, 2019. doi: 10.1109/ICASSP.2019.8683262.

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[25] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

[26] D.C Dowson and B.V Landau. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982. ISSN 0047-259X. doi: https://doi.org/10.1016/0047-259X(82)90077-X. URL https://www.sciencedirect.com/science/article/pii/0047259X8290077X.

[27] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them, 2020.

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[29] Eric Schussler, R. Jagacinski, Susan E White, A. Chaudhari, J. Buford, and J. Oñate. Inter-rater agreement and validity of a tackling performance assessment scale in youth american football. *International journal of sports physical therapy*, 13 2:238–246, 2018.

[30] Mostafa Tavassolipour, Mahmood Karimian, and Shohreh Kasaei. Event detection and summarization in soccer videos using bayesian network and copula. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(2):291–304, 2014. doi: 10.1109/TCSVT.2013.2243640.

[31] Muhammad Zeeshan Khan, Summra Saleem, Muhammad A. Hassan, and Muhammad Usman Ghanni Khan. Learning deep c3d features for soccer video event detection. In *2018 14th International Conference on Emerging Technologies (ICET)*, pages 1–6, 2018. doi: 10.1109/ICET.2018.8603644.

Appendix A

SATT

1) PC - Player Control - *Does CoG remain within the Base of support?* — Evaluated moments just prior to contact

Did not occur - 0	Poor - 1	Average - 2	Excellent - 3
NOT under control	Player is NOT under control - *reaches or leans*	Player is mostly under control	Player is under control
Feet leave the ground	Feet are NOT in balance	Feet are NOT balanced	Feet are balanced
Player launches into target	*CoG outside of BoS*	*CoG near or within of BoS*	*CoG is low and in middle of BoS*

2) HET - Head Eye Torso Position - *Does Angle of torso remain even/above angle of the shin?* — Evaluated just prior to initial contact

a. Head & Eyes in Neutral; b. Torso aimed at target; c. Hips stay low

Did not occur - 0	Poor - 1	Average - 2	Excellent - 3
Head is pointed DOWN (crown is exposed)	Head looking down, player bending at the waist	Head is neutral and chest remains UP	Head & Eyes are UP
is UNABLE to achieve breakdown position	Angle of torso is Greater than the angle of the shin	Angle of torso is even with the angle of the shin	Chest UP points at target
primarily frm the WAIST; little to no hip/knee Flxn	Chest points toward the ground (Shoulders fwd)		Angle of torso is less than the angle of shin (taller)
Angle of torso is Greater than the angle of the shin			**Hips are LOW: but on the rise**

3) Strike Zone (SZ) - *"Head remain CLEAR of target?"* - Makes contact with FRONT of chest / top of the shldr *[ ] or Horizontal Tackle* — Evaluated AT POINT of contact

Did not occur - 0	Poor - 1	Average - 2	Excellent - 3
Initiates contact with head, minimal shoulder strike	Head makes incidental contact	Head remains clear/side	Head remains clear
Arms tucked, leading with the shoulder	Does not initiate contact with chest/shldr	Makes contact with TOP of SHOULDER	Makes contact with front of CHEST
	Arms are out of position	Arms are NOT engaged or ready to fire	Arms are cocked back and ready to fire
Horizontal Tackle (HT) - Head position			
Head is not to the side	Head to Front/Leading Side	Head is to Side, but makes contact with opponent	Head to Back/Near Side
initiates the contact			(Spirals player to ground)

4) Ascending Hit? "Raise CoG of the target?" *a. Do hips forcefully extend, THEN b. Do arms "Rip" upwards:* *[ ] or Horizontal Tackle* — Evaluated moments AFTER contact

Did not occur - 0	Poor - 1	Average - 2	Excellent - 3
Little to no Hip Explosion	Partial, Ineffective, poorly timed Hip Explosion	Hips are Low/short but do not rise UP into target	Hips start out low and explode UP and through
Arms do not effectively fire, or extend	Hips remain HIGH/tall throughout;	Hips remain the on the same level	Target rises up off of the ground
	Incomplete Arm Rip - Arms fire late, no squeeze		
Horizontal Tackle (HT) - Head position			Head to Back/Near Side
Does NOT Secure	Partially Secures Torso or Hips	Head is to Side, but makes contact with opponent	(Spirals player to ground)

5) Leg Drive; (LD): "Drive for Five" *Do Feet remain churning after contact?* 5x? *[ ] or Horizontal Tackle* — Evaluated 2-3sec AFTER contact

Did not occur - 0	Poor - 1	Average - 2	Excellent - 3
Legs do NOT drive through contact	Legs drive but stop on contact	Legs Drive for one or two steps	Legs continue to drive all the way through contact
Feet leave the ground			Able to make at least 5 quick steps
Horizontal Tackle (HT) - Head position			Head to Back/Near Side
Does NOT spiral opponent	Partially Spirals Opponent to Ground	Head is to Side, but makes contact with opponent	(Spirals player to ground)

6) Finish Position; (FP): Lands in a position of dominance? Chest on target, hips high, toes in grass, hands grip the jersey? *[ ] or Horizontal Tackle* — Evaluated 2-3sec

Did not occur - 0	Poor - 1	Average - 2	Excellent - 3
Feet leave the ground	Feet not engaged on ground	Are on the ground	Toes locked on the ground
Lands or rolls off to one side	Has body weight on opponent briefly	Has body weight on opponent	Chest is on-top puts body weight into target
Does not Grip jersey	Does not Grip jersey	Does not Grip jersey/ grips with only one hand	Both hands Grip jersey

Figure A.1: