applied sciences

MDPI

Article

Generation of Pedestrian Crossing Scenarios Using Ped-Cross Generative Adversarial Network

James Spooner [1,*,†], Vasile Palade [2,*], Madeline Cheah [3], Stratis Kanarachos [4] and Alireza Daneshkhah [2]

[1] Centre for Connected and Automated Automotive Research (CCAAR), Institute for Future Transport and Cities, Coventry University, Coventry CV1 5FB, UK
[2] Research Centre for Data Science, Coventry University, Coventry CV1 5FB, UK; ac5916@coventry.ac.uk
[3] Horizon Scanning, HORIBA MIRA Ltd., Watling Street, Nuneaton CV10 0TU, UK; madeline.cheah@horiba-mira.com
[4] Faculty of Engineering and Computing, Coventry University, Coventry CV1 5FB, UK; ab8522@coventry.ac.uk
* Correspondence: spoonerj@uni.coventry.ac.uk (J.S.); ab5839@coventry.ac.uk (V.P.)
† Current address: Engineering and Computing Building, Coventry University, Gulson Road, Coventry CV1 5FB, UK.

Abstract: The safety of vulnerable road users is of paramount importance as transport moves towards fully automated driving. The richness of real-world data required for testing autonomous vehicles is limited and furthermore, available data do not present a fair representation of different scenarios and rare events. Before deploying autonomous vehicles publicly, their abilities must reach a safety threshold, not least with regards to vulnerable road users, such as pedestrians. In this paper, we present a novel Generative Adversarial Networks named the Ped-Cross GAN. Ped-Cross GAN is able to generate crossing sequences of pedestrians in the form of human pose sequences. The Ped-Cross GAN is trained with the Pedestrian Scenario dataset. The novel Pedestrian Scenario dataset, derived from existing datasets, enables training on richer pedestrian scenarios. We demonstrate an example of its use through training and testing the Ped-Cross GAN. The results show that the Ped-Cross GAN is able to generate new crossing scenarios that are of the same distribution from those contained in the Pedestrian Scenario dataset. Having a method with these capabilities is important for the future of transport, as it will allow for the adequate testing of Connected and Autonomous Vehicles on how they correctly perceive the intention of pedestrians crossing the street, ultimately leading to fewer pedestrian casualties on our roads.

Keywords: CAV; automotive; autonomous; pedestrian; dataset; human pose; GAN; machine learning

check for updates

Citation: Spooner, J.; Palade, V.; Cheah, M.; Kanarachos, S.; Daneshkhah, A. Generation of Pedestrian Crossing Scenarios Using Ped-Cross Generative Adversarial Network. *Appl. Sci.* **2021**, *11*, 471. https://doi.org/10.3390/app11020471

Received: 30 November 2020
Accepted: 27 December 2020
Published: 6 January 2021

1. Introduction

According to the World Health Organisation, approximately 1.35 million people die each year due to road traffic crashes with more than half of these deaths being among vulnerable road users (including pedestrians) [1]. With the advent of connected and autonomous vehicles (CAV) being introduced on roads, the issue of pedestrian safety has never been more critical.

National Highway Traffic Safety Administration (NHTSA) (2015) report that, of just over 4 million vehicle crashes in the USA between 2005 and 2007, 94% of the collisions had a critical reason of the collision assigned to the driver [2]. Knowing this, the development and eventual adoption of CAVs is expected to have a major impact on the number of pedestrian casualties on the road and at least one study has indicated that pedestrian fatalities could be reduced by 30% to 90% based on CAV sensor technology alone [3]. However, to be able to achieve widespread and public deployment, the CAV needs to be sufficiently trained using datasets that reflect the reality of how pedestrians act and interact on the road.

The progress yet to be made with regard to pedestrian safety in CAVs is highlighted by the recent death of a pedestrian in an Uber autonomous driving trial. This created a significant negative impact on public perception and acceptance of such technology [4–6]. This is proof that there is more work to be done in the area of CAVs and pedestrian protection in particular.

Using pedestrian datasets captured from a naturalistic driving video is not unique and has been used for machine vision tasks for many years. However, for CAVs to be able to perform as well as (if not better than) a human driver, they need to not only identify the presence of a pedestrian, but also need be able to judge a pedestrian's actions. To do this, the richness of data that can be acquired from real world driving scenarios is needed. This has traditionally been done using dash board mounted cameras where pedestrians have been assigned ground truth bounding boxes. The dataset has then been published in a video format with ground truth pedestrian labels [7]. The issue with datasets like these, is that the driving video does not cover all scenarios, such as rare and unexpected events. The dataset may contain such an event, but that would be entirely coincidental.

In this research, video data captured from three naturalistic driving datasets are used to form the Pedestrian Scenario dataset. The video footage from this dataset is simplified by using human pose estimation. Human pose estimation is an active area of research, which has the aim of simplifying a human into a set of keypoints. When this has happened, you are left with a simplified representation of a human with all the rich data corresponding to the main points and shape of the body. Human pose estimation is used for numerous research activities as outlined in Section 2.3, however for the task at hand in this research, it enables video data to be simplified into a form where the rich pedestrian data is extracted and then used in a novel Generative Adversarial Network (GAN) to generate new pedestrian scenarios.

We present two contributions in this paper. Firstly, a novel Generative Adversarial Network (GAN) is introduced for the purposes of generating entirely new pedestrian crossing scenarios, based on those learned from a dataset introduced in this paper. This GAN, named Ped-Cross GAN, when combined with the pedestrian scenario dataset is capable of generating pedestrian crossing scenarios, not simply based on the primary movement characteristic, but is able to generate scenarios based on the behavior, speed, and age of the prospective pedestrian. This work can have far reaching implications for the training, testing, and validation of future autonomous vehicles. This is due to the ability to generate a plethora of pedestrian scenarios either by generating a random combination of parameters or by allowing engineers to tightly define and test a specific scenario. This has the advantage of an engineer being able to generate and test a specific rare case scenario or to remove any kind of hard coded bias entirely if conducting random testing.

The second contribution is the development of a novel dataset for pedestrian scenarios, derived from existing datasets, but with extensive relabeling to ensure that data about pedestrian movements and actions is present. We also provide pedestrian sequences not only in image format, but also human pose format. This change in format significantly reduces the amount of data required for training and testing systems on CAVs, while also maintaining the richness of scenario data captured from the original images.

The remainder of this paper is structured as follows: Section 2 outlines the related work to this research. This is then followed by Section 3 where the Pedestrian Scenario Dataset is outlined and introduced. Generative Adversarial Networks are introduced in Section 4, while the Ped-Cross GAN is presented in Section 5. Section 6 highlights the results from Ped-Cross GAN, and the discussion is in Section 7. Finally, the Conclusions and Future work are presented in Section 8.

2. Related Work

2.1. Pedestrian Deaths and CAVs

When considering world crash statistics, it is clear that the fatalities of pedestrians make up a large proportion of all road deaths. Table 1 shows the global distribution of road traffic deaths and it can be seen that in the more developed regions of the world, pedestrian deaths account for between 22% and 27% of all road deaths, with 22% as the global average [8]. Further to this, recent trends in road deaths show that pedestrian fatalities are decreasing at a slower rate when compared to all other types of road users [9].

Table 1. Proportion of pedestrian deaths per world region [8].

World Region	Proportion
Africa	39%
Eastern Mediterranean	27%
Europe	26%
Western Pacific	23%
The Americas	22%
South-East Asia	13%
World	22%

The United States has seen a rise in pedestrian fatalities, with a 9% rise in deaths between 2015 and 2016 [10]. 2016 saw the overall proportion of pedestrian deaths rise to 16% as a national average in the US compared to 11% in 2007, however, when considering highly populated urban areas, this proportion is significantly higher, as seen in Table 2. The European Commission (2015) also comment that 69% of all pedestrian fatalities occur in urban areas [9].

Table 2. Proportion of pedestrian deaths from total road deaths in selected US cities—2016 [10].

City, State	Proportion	Total Road Deaths
New York, NY	59.6%	230
San Francisco, CA	50.0%	28
Boston, MA	48.1%	27
Fresno, CA	46.2%	13
San Diego, CA	43.8%	96
Philadelphia, PA	42.6%	101
Los Angeles, CA	41.3%	315

The Reported Road Casualties Great Britain reported that 25% of all road fatalities in 2016 were pedestrians and that 18% of all road accidents in urban areas included a pedestrian. It is also reported that the vast majority of pedestrian to vehicle interactions occur in urban area; unsurprisingly, four wheeled motor vehicles are the most frequently involved vehicle [11].

A total number of 448 pedestrians lost their lives on UK roads in 2016 and interestingly, 61.6% of these fatalities were in non-occluded scenarios, meaning that the pedestrian was in full view at the time of collision. A total of 12.5% of pedestrian deaths occured on a pedestrian crossing facility at the time of collision. When considering all severities of injury to pedestrians on crossings, the most frequently involved crossing type is a Pelican crossing (34.3%), closely followed by a light controlled junction (31.6%), and Zebra crossing (28.6%). With this information, it is clear that collecting relevant scenarios is crucially important [11].

All of these statistics can offer valuable insight into scenarios, junctions, and situations which will be of higher interest for automated vehicles when training, testing, and validating their pedestrian safety. This information gives a macroscopic view of the work currently being done to reduce pedestrian fatalities on the road. With pedestrian deaths reducing at a slower rate, and in some areas actually increasing, it is clear that pedestrian safety is an issue that needs to remain at the forefront of research. This allows developers and engineers to better concentrate on areas of interest where interactions are more likely to occur and to train their systems to have a better knowledge of these scenarios.

2.2. Pedestrian Datasets

Pedestrian datasets have been an area of interest for more than 15 years. One of the first examples was published by Dalal and Triggs (2004) [12]. Their INRIA dataset contained 1805 64 × 128 images of people in various orientations. However, this dataset contained only individual images of people, and not sequences/videos, and therefore is not capable of understanding the context of a pedestrian's movements, such as intention to cross or looking at the ego vehicle.

The datasets used in this work are the Caltech pedestrian dataset [7], the Joint Attention for Autonomous Driving (JAAD) [13], and the Daimler pedestrian dataset [14]. Caltech was one of the earliest comprehensive pedestrian datasets aimed to improve the detection of pedestrians. Subsequently, it has been used extensively as a benchmark for machine vision tasks and pedestrian detection [15–17]. It comprises approximately 10 h of 640 × 480 30 fps driving footage. A total of 250,000 frames were annotated, with more than 350,000 bounding boxes and 2300 unique pedestrians.

JAAD is a relatively new pedestrian dataset and is different from others as it goes a step beyond simply labeling the pedestrian [13]. This dataset labels the behavior of the pedestrian, such as 'Looking', and their speed. JAAD has been used to instantaneously predict if a pedestrian is about to cross the road [18]. However, what the dataset lacks is specific information regarding the movement action, such as their movement direction or their speed. What is presented in Section 3.2, shows that this research solves this issue. The JAAD dataset contains 346 videos between 5 and 10 s in length, their ground truth contains approximately 82,000 frames and 2200 unique pedestrians, resulting in 337,000 bounding boxes.

The final dataset used was the Daimler pedestrian dataset which was collected from moving or stationary vehicles [14]. The 68 video clips contain 4 key movement types, which were crossing, stopping, starting to walk, and bending-in. Of these 68 video clips, there were more than 12,000 images containing pedestrians.

2.3. Human Pose Estimation

Another area of related research to both the dataset presented in this paper as well as pedestrian detection is that of human pose estimation. Human pose estimation takes an image or video, identifies any humans in the image, and overlays a skeleton type structure over the image. This skeleton type structure is built up by estimating keypoints on the body. Human pose estimation is an active area of research, with several different methods published in recent years. One of the most well-used pose estimators is known as OpenPose [19], which was one of the first pose estimators to be able to make pose predictions in real time. OpenPose is able to score an average precision on keypoint predictions of 84.9% at 50% confidence.

The pose estimator used in this research is known as Alpha Pose [20]. This pose estimator was selected for its improved results versus other state-of-the-art pose estimators. When compared to Openpose, Alphapose is able to score an average precision of 89.2% at a 50% confidence, therefore scoring higher than OpenPose.

Human pose estimation conventionally works in a two step method. Firstly, an image is processed to scan for humans through a human proposal network. This network is usually a pretrained human identification network. Following this, the regions of the image where it is thought a human might be present are passed to a second part of the network. This is where the human pose estimation occurs. If the confidence in the human pose is too low, then the human pose is not output, and the region suggested is therefore considered to be a false positive.

Following the proposal of the human, the image goes through 3 stages in order to extract the human pose: A Semantic Spacial Transformer Network (SSTN), a Single Person Pose Estimator (SPPE), and a Spacial De-Transformer Network (SDTN). In summary, the SSTN identifies regions on the human which could be body parts, the SPPE uses a Recurrent Convolutional Neural Network to estimate the human pose in the segment of the image, while the SDTN enables the translation for the human pose to be remapped on the image. Further details can be found in their paper [20].

The human pose estimation network is trained using annotated human pose datasets. In the case of this paper, Alpha Pose was trained on the COCO dataset [21] and MPII Human Pose dataset [22].

Human pose estimation is not without its drawbacks and limitations, and such limitations will trickle down into the results presented in this research. Alpha Pose reports that their human pose estimator can struggle to correctly estimate the pose of two humans who are overlapping. This will obviously have connotations in this research when pedestrians are observed to be crossing each other's path. We outline methods of how we mitigate this in Section 3.1.2.

Pose estimation is especially relevant due to how greatly it reduces the size and dimensionality of the data given to it. For example, a pose estimator will take an image of a person of size 200 × 180 × 3 data points, and predict keypoints, reducing the data sample size from 108,000 to just 34 (17 keypoints in x and y). In the context of training CAVs and any type of machine learning, this will greatly reduce the time taken to train, the complexity of the model (due to having fewer dimensions/features), and computer hardware required, while maintaining rich data of the human body frame extracted from the original image.

For this research, Alphapose was used as an off-the-shelf pose estimator, and while it was not within scope to develop our own pose estimator, it allows other researchers to build on our results when newer pose estimators are released in the future.

3. Pedestrian Scenario Dataset

3.1. Dataset Curation

The selection of appropriate datasets was based on availability with regards to the quality of video and variety of scenarios. In this paper, we used the Caltech Pedestrian dataset [7], the JAAD dataset [13], and the Daimler Pedestrian Dataset [14] as the base datasets. These were chosen due to the length of the datasets, in terms of videos available as well as video length, but also due to the camera quality of the recorded videos. All videos used in the datasets are captured from the perspective of an ego vehicle, and captured using a dashboard mounted camera, such that all pedestrians were presented from an equivalent perspective.

Using the ground truths from each dataset, each pedestrian was cropped from each frame for each video. A simple naming convention was developed for each image so that they could be easily traced back to the original source video if required. The convention is *DatasetName_VideoNumber_PedestrianNumber_Frame*. This resulted in each pedestrian having a sequence of cropped images containing only themselves.

The final step in this section was to remove any cropped frames that were below 3kB in size. This decision was made as it was judged that these images were far too small to be able to extract any tangible meaning.

3.1.1. Additional Labeling

The subsequent task forms the majority of our contribution towards the novel dataset.

Each sequence of cropped pedestrian frames was viewed and assigned 23 labels, describing three main classes: Primary movement, secondary behavior, and tertiary descriptive classes, as seen in Table 3–5, respectively. Details of the primary movement class were captured using the schematic in Figure 1 and specific labels can be seen in Table 3. All of the labels were assigned manually by the authors after viewing the videos.

These 10 primary classes are those that describe the primary movement of a pedestrian in a crossing scenario. It was also noted whether the pedestrian exhibited irregular behavior and whether or not the primary movement occurred at a crossing.

It was also necessary to collect situational and descriptive labels for the pedestrian themselves. The labels for the secondary and tertiary class can be seen in Table 4 and 5, respectively.

Table 3. Pedestrian movement classes.

Number	Label
1	Crossing from the left
2	Diagonal towards cross left
3	Diagonal adjacent cross left
4	Crossing from the right
5	Diagonal towards cross right
6	Diagonal adjacent cross right
7	Walk towards traffic, no cross
8	Walk adjacent to traffic, no cross
9	Stand left
10	Stand right

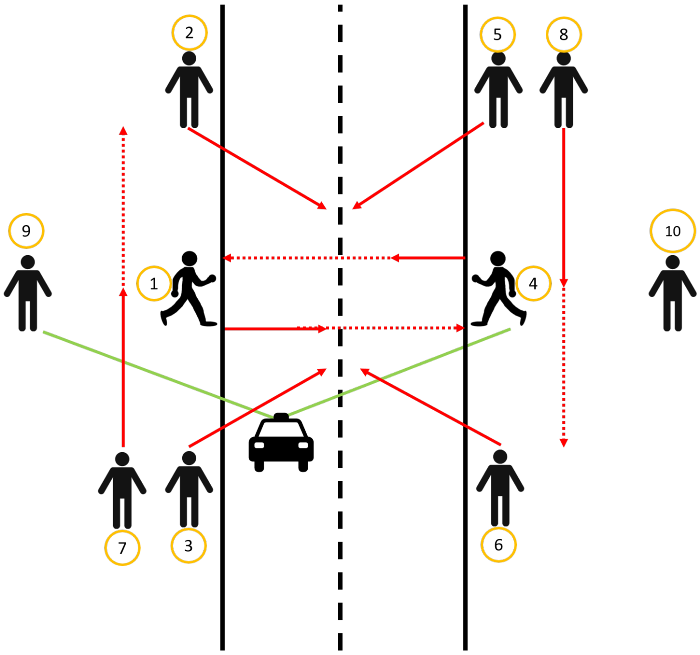

Figure 1. Pedestrian movement classes.

Table 4. Secondary behavior classes.

Action	Label
Speed	No Movement, Slow walk, Walk, Jog, Run
Hesitation	Yes, No
Peeking	Yes, No
Looking	Yes, No
Distraction	Yes, No
Waiting	Yes, No
Waving	Yes, No
Jump back	Yes, No
Intoxicated	Yes, No
Freeze	Yes, No
Trip	Yes, No
Other mobility	Skateboard, Rollerblade, Scooter, Other, N/A

Table 5. Tertiary descriptive classes.

Desciptive	Label
Age Range	0–15, 15–60, 60+
Gender	Male, Female, Unknown
Ethnicity	White, Asian, Black, Mixed race, Unknown
Occluded	Yes, No
Occluded by ped	Yes, No
Full body	Yes, No
Hunched over	Yes, No
With object	Shopping, Dog, Pram, Crutches, Walking frame, Suitcase, Other

The labels for the second classes were for context, which is also useful for training machine vision for use in autonomous vehicles. Innate human behaviors such as *looking* and *hesitating* will be crucial for practitioners to include, so that when on-board systems are trained, they are aware of the likely events that follow when a pedestrian looks at an approaching vehicle.

The tertiary classes collected relate to human descriptives, such as age range, whether or not the pedestrian is occluded at any point in the scene, or whether whether their full body was viewable (for example, if the pedestrian is close to the ego vehicle, only their torso might be visible). These classes can be seen in Table 5.

3.1.2. Pose Dataset

Following the labeling of all the cropped image sequences in the dataset, all of the images were translated to a human pose format, using an off-the-shelf pose estimator. The pose estimator used was Alpha Pose [20]. The same methodologies, as applied in this paper, could also be used with an improved future pose estimator, thus allowing practitioners to create more accurate pose estimations. Creating new or improving on the pose estimation methods themselves are out of the scope of this paper.

The reason for translating the image data into pose data is to simplify the images from an approximate size of 3 kB or greater, to just 17 coordinate points in the image. By reducing the size and dimensionality of the data, we reduce the computing power required for machine learning tasks (Section 5.3). An example of an original cropped image with the pose estimation results overlayed can be seen in Figure 2.

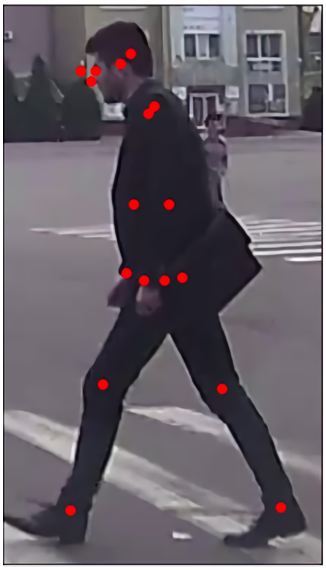

Figure 2. Cropped pedestrian example with pose (from JAAD).

As AlphaPose pose estimator comes with its own errors [20], every pose prediction made would reflect these errors. To mitigate this, we set the confidence of the pose estimator to be at least 50% confident in a prediction in order to retain the prediction. Alpha Pose reported an average precision on pose predictions of 89.2% at 50% confidence.

It was also necessary to check that the output poses were anatomically viable. Having labeled the whole dataset, it was known that all the pedestrians in the dataset were standing. As such, a set of rules were developed to filter for poses that did not conform, as listed below:

- Keypoints for the shoulders to be in the top 30% of the prediction;
- Keypoints for the hips should be within the 40% to 60% range in the prediction;
- Keypoints for the feet should be in the bottom 40%;
- Keypoints for the feet should not be above the knee; and
- Keypoints for both knees should be within 10% height of each other.

These rules were applied to the entire dataset of poses. Where non-conformant poses were detected (for example due to occlusion in the individual sequences), we interpolated points between the nearest acceptable poses to replace those that were incorrect. The exception to this process was where poses were non-conformant in more than 10 sequential frames, which resulted in the pose sequence being cut to before and after the failed poses. Sequences processed thus were then re-assessed against the original video to check if any relabeling was required. A total of 8 sequences required relabeling, and 2 sequences were removed entirely due to not having enough good frames in succession.

In summary, this activity ensured that all pose sequences were anatomically plausible.

3.2. Dataset Results

In this section, the results are presented for the image and pose datasets in relation to the labels collected for each class. We present the statistics of the dataset, as well as drawing out some comparisons between labels. This shows the diversity of the scenarios collected in the dataset.

3.2.1. Image and Pose datasets

The curated dataset contains a hybrid from the Caltech pedestrian dataset [7], JAAD [13], and the Daimler pedestrian dataset [14]. As the dataset is presented in two forms, it is useful to highlight the statistics of the data with respect to both the raw frames and processed pose estimation.

From the cropped pedestrian image sequences, there are a total of 102,388 individual frames across the 932 pedestrian sequences. These sequences vary in length, with the average number of frames per sequence being 109 frames. The distribution of sequence length, as well as which dataset they occurred from can be seen in Figure 3. From Figure 3, it is clear that the vast majority of sequences are between 0 and 200 frames in length, thus showing the variety and diversity of sequence length available. The videos used to form this dataset were all captured at 30 frames per second.

All frames were passed through Alpha Pose. Due to the rules imposed (see Section 3.1.2), the size of the dataset reduced after pose estimation. This is due to the size and quality of the images available. Where the pose estimator was not able to make a prediction, there would be no pose generated. For particularly poor image sequences, Alpha Pose was unable to make a prediction on any of the frames, while other sequences saw the pose output drastically reduce the number of poses produced when compared to the original image sequence.

Subsequently, the resulting pose sequence dataset contains the same 932 sequences, however the total number of poses was reduced to 88,577, creating an average sequence length of 94 poses. As a result, this means that the size of the pose dataset, in terms of number of samples, was reduced by 13.4%. However, it is clear the distribution remains very similar to the distribution of the image sequences (Figure 4), again with the vast majority of sequences being between 0 and 200 poses in length.

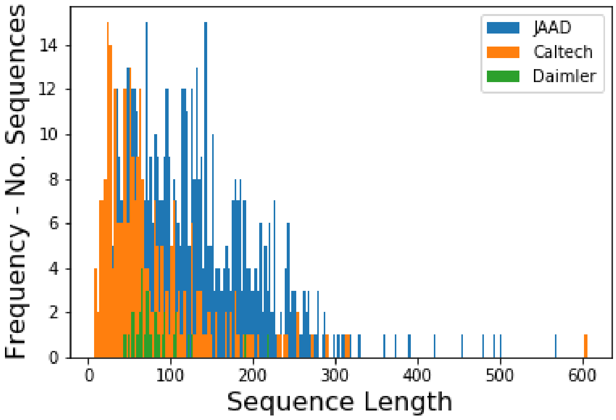

Figure 3. Distribution of sequence length on cropped pedestrian images.

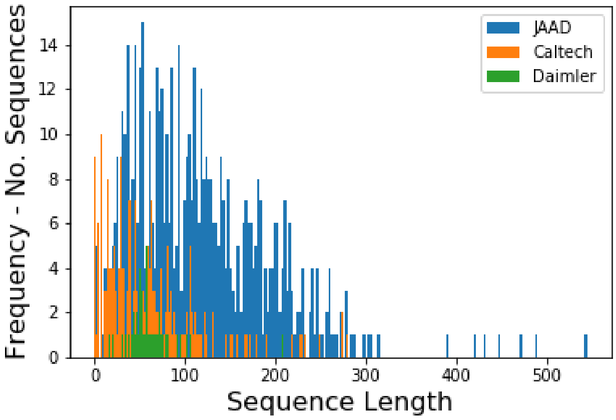

Figure 4. Distribution of sequence length on pedestrian poses.

3.2.2. Labels and Classes

From the extensive labeling and classes collected on this dataset, some useful and interesting statistics emerge. All of the labels were collected by viewing the sequences of images, therefore the quoted numbers below will reflect those of the image dataset. The labels are also valid for the pose sequences, as the the pose dataset is derived directly from the image dataset.

The first comparison that can be drawn is seen in Table 6. In this table, the comparison is between the type of movement observed and the speed at which the pedestrian did the movement. It is clear that crossing from the left (movement class 1) and crossing from the right (movement class 4) are the most common movement types. Normal walking speed (speed class 2) is the most common speed observed.

The labels collected in the dataset also allows us to learn from the types of things pedestrians are carrying or manoeuvring around the roadside. It was most common for a pedestrian to not be carrying anything with 611 samples, however of the cases where the pedestrian is carrying something, shopping is the most common, with 218 samples. In this dataset, we have 22 examples of pedestrians pulling a suitcase at a crossing scenario.

This information is important as pedestrian posture, body shape, and movement style can change extensively depending on the object being maneuvered. For example, someone will move very differently when pulling a suitcase when compared to someone who is not (see Table 5 for objects labeled).

Table 6. Movement class compared to speed.

		Speed					
		0	**1**	**2**	**3**	**4**	**Total**
Movement Class	1	0	14	170	22	0	206
	2	0	3	42	3	0	48
	3	0	5	28	4	2	39
	4	0	18	226	21	6	271
	5	0	6	28	5	0	39
	6	0	5	54	4	0	63
	7	0	8	64	0	1	73
	8	0	20	56	1	0	77
	9	20	11	1	0	0	32
	10	50	31	3	0	0	84
Total		70	121	672	60	9	932

4. Generative Adversarial Networks

Generative Adversarial Networks have been gaining popularity in recent years, since their inception in 2014 [23]. The GAN, in its most basic form, is two neural networks that are trained simultaneously, a Discriminator and a Generator. The Generator, G, captures the data distribution by testing the generated samples on the Discriminator, and the Discriminator, D, estimates whether the sample came from the training data, or from G.

When training the GAN, G has the objective of maximizing the probability that D will make a mistake, while D has the objective of maximizing the probability that it can identify a generated, fake sample from G. Thus, this creates a minimax, two-player game. GANs are designed so that they reach a Nash equilibrium [24] where each player cannot reduce their cost without changing the other players' parameters [25]. In practice, a GAN is successfully trained once G has adequately recovered the training data distribution, while D outputs a confidence result of 50% when presented with either a training data sample, or a sample from G.

In this work, the majority of training methods are based on the Wasserstein GAN [26]. This is a method that changes the loss function from that of a Jansen–Shannon function to a Wasserstein function. By doing this, the risk of experiencing 'exploding' gradients is negated. 'Exploding' gradients occur in GANs during training and causes the gradients in both the discriminator and generator to diverge and tend towards infinite values, thus rendering training a fruitless task.

5. Ped-Cross GAN

This section introduces the second contribution of this paper. Building upon the foundations of GANs outlined in Section 4, a novel GAN architecture is defined. This novel GAN, named Ped-Cross GAN, is used to generate human pose crossing sequences in sequence lengths of 5 at a time. These sequences can either be from entirely newly generated human pose sequences, or the pose sequence can be generated from a given starting and end pose extracted from the original pedestrian scenario dataset.

The GAN itself is formed of a Discriminator and Generator. The novelty of the Ped-Cross GAN comes from the architecture from within the Generator. This is defined in Section 5.1. The training and decisions around training are outlined in Section 5.2, and the hyperparameters are then defined in Section 5.3 to allow reproducibility of the results.

5.1. Network Architecture

Traditionally, the Generator, G, in a GAN is provided with random noise, Z, which is then passed through the generator to put the data points in the order or orientation that the Discriminator, D, can understand. In this task, G has two requirements: It needs to generate human like pose structures and to make sure sequential human pose structures are plausible for generating the movement of a human being. These are two difficult tasks for the GAN to succeed at, therefore, a method was devised that approached this challenge.

Instead of completing both tasks at the same time, that of generating a human pose and putting those poses in a plausible sequence, it was decided to segment the task of generating the human pose, and to do this in a previously trained GAN. Using the GAN created by Spooner et al. (2019) [27], a human pose generator was trained as per the training criteria outlined in that paper.

A single generated pose then formed the input data for G in the Ped-Cross GAN. The same pose was duplicated the number of times for the desired sequence required from the GAN. For instance, the results in this paper are based upon the generation of a pose sequence of five poses, therefore, the starting generated pose would be duplicated so that the starting input in G was five of the same poses.

Before the poses were passed from the pretrained pose generator to G in the Ped-Cross GAN, they were checked for anatomical accuracy, such that the generated poses were human like. If they were not, then samples were generated until a suitable set of poses were generated. This ensured that every human pose that G and D saw was human-like, and thus, significantly simplifying the training required in G.

For G in Ped-Cross GAN, the network architecture is a fully connected, feed-forward four layer network, which takes input as a sequence of five human poses and outputs at the same dimension. The total number of neurons in the network was 6996. Due to feeding in already generated human poses, the task of the generator is to learn the distribution of the training data, which the discriminator learns. The task for G is more of a task of adjusting and reordering the generated poses into an order that will fool D.

For D in Ped-Cross GAN, the network architecture was based around a Long-Short Term Memory (LSTM) network. The LSTM has been well used in similar sequential, time-based machine learning problems. It is well regarded for its ability to retain relevant information over a period of time steps. Introduced for the first time in 1997 [28], they have been used extensively in machine vision tasks and more recently in GANs [29]. The LSTM used as D in Ped-Cross GAN was a single layer LSTM with 400 hidden units, with a fully connected output layer to extract the classification score. A diagram of Ped-Cross GAN can be seen in Figure 5.

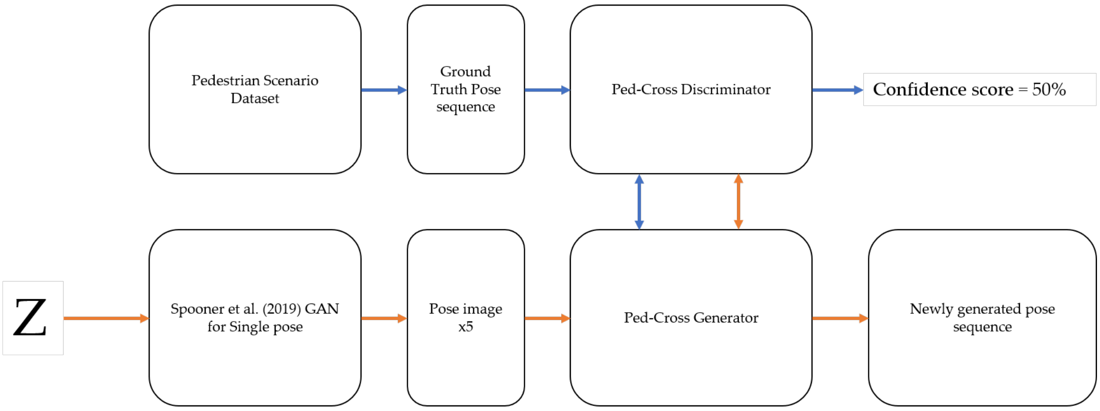
Pedestrian Scenario Dataset
Ground Truth Pose sequence
Ped-Cross Discriminator
Confidence score = 50%
Z
Spooner et al. (2019) GAN for Single pose
Pose image x5
Ped-Cross Generator
Newly generated pose sequence

Figure 5. Schematic of Ped-Cross Generative Adversarial Networks (GAN).

5.2. Training

For training and testing of the Ped-Cross GAN, a simplified version of the dataset was selected. The reasoning behind this was to first prove the concept of generating new pose sequences based on a main movement, before extra classes we added. Extra classes, such as those outlined in Table 4 and 5 fall out of scope for the results presented in this paper however they will form the basis of future research.

For this research, it was decided that a sequence length of five frames would be adequate for training the GAN and generating samples. The reasoning for this was that after every five frames, if a practitioner was to want a longer sequence, they would be able to use the final frame of the five as a starting point for the next five frames, and so on. Therefore, subsequent generated poses could, in theory, be generated indefinitely until the desired sequence length was achieved.

As a result, the only class labels that were used were from the primary movement class. From inside this primary movement class, we used Class 1 (Crossing from the left) and Class 4 (Crossing from the right), as seen in Table 3 and Figure 1. This resulted in a sample size of 206 sequences for crossing from the left and 271 sequences for crossing from the right. The sequences are all of varying lengths, as can be seen in Figure 4. Therefore, by only taking five frames from each sequence in each class would mean losing an enormous amount of useful data from the respective samples.

The decision was made to slice each sequence in each class into subsequent sequences, all with a length of five frames. This drastically extended the usable dataset for the two classes at hand. Further to this, during the curation of the dataset, it was noted that in the majority of examples, it would take several frames for the pedestrian to begin their characteristic movement for the specified class and several frames at the end of each sequence where the characteristic movement was not recognizable. Therefore, to negate this immeasurable drawback, 10 frames from the beginning and 10 frames from the end of each sequence were removed and not considered in the training of the GAN. The resulting samples for each class were 4268 for Crossing from the Left, and 4749 for Crossing from the Right.

To create a meaningful training and testing set to evaluate the success of Ped-Cross GAN, the dataset was divided in to two subsets, 80% for training and 20% for testing, as is the convention with deep learning algorithms. An equal distribution was randomly removed from both classes, culminating in 7215 samples for training, and 1802 samples for testing. This testing set will be used to validate the GANs results in Section 6.1.

5.3. Hyperparameters

The hyperparameters for the Ped-Cross GAN have been finely tuned to provide optimal results, which can be seen in Section 6. Throughout training, Stochastic Gradient Descent (SGD), ADAM optimizer [30], and the RMSProp optimization algorithm [31] were all used and compared. For the training of Ped-Cross GAN, it was found that RMSProp provided the most optimal results. RMSProp provided far more stable training with the loss from the Generator tending towards zero, whereas ADAM and SGD experienced some volatility. Both the discriminator and generator used the RMSProp optimization algorithm.

The sequences were trained at a length of five frames, such that D would see sequences of five frames from the Pedestrian Scenario dataset, and G would feed in five poses with the ambition of creating a pose sequences that matches the distribution of the Pedestrian Scenario dataset.

Ped-Cross GAN was trained for 5000 epochs, during each epoch, D would see samples from the Pedestrian Scenario dataset five times for every time it would see a generated sample. This is due to the way the GAN trains. If the two are trained to the same level, it means that G would become too good at fooling D too quickly, sometimes even quicker than D has time to learn a meaningful distribution from the dataset. By creating this imbalanced game, it forces G to work harder to fool D, as D will have a much better idea of what a true sample looks like. For this research, a 5:1 training ratio was found to be the right balance. It was found that if D was trained more than this, D became too good at identifying the generated samples from G, hampering the training of G.

Other notable parameters were the learning rate, which was set at 0.0001, and the batch size, which was set to be 32.

6. Ped-Cross GAN Results

In this section, the results for the Ped-Cross GAN will be introduced and analyzed. To deliver insight in to the generated results, the results are validated in a number of ways, which will be outlined in Section 6.1. The validated results themselves will be shown in Section 6.2, while selected generated results will be shown in visual form in Section 6.3.

6.1. Validation Method

When training and testing GANs, it is important to avoid a self fulfilling prophecy. That being, when the generated results are tested on the very same data in which the discriminator was trained using in the GAN itself. To avoid this, the subset of the Pedestrian Scenario dataset was divided, to keep 20% of the samples to one side, so that they could be used for testing and validation. In this case, this meant that 1802 testing samples were available, 1802 samples which Ped-Cross GAN would have never seen prior.

The fully trained Ped-Cross GAN is capable of generating as many, or as few, new samples as is required. For that reason, several different number of samples were generated for testing the success of the GAN. While 5000 samples for each class was the number of samples that provided the most favorable and balanced results, validation efforts were also carried out on fewer and greater samples for each class.

The validation methods used were in the form of a simple LSTM classifier network. This network would see each pose sample in a sequence length of five, and output a classification score. In this case, the classifier would classify whether the pose sequence was one of crossing from the left, or a pose sequence of crossing from the right.

The architecture of the validation LSTM was very similar to the LSTM in the discriminator in Ped-Cross GAN. The reason for this is that both networks are essentially doing a very similar classification task. The discrimintator in Ped-Cross GAN is trying to classify between two classes, whether a sample is real or fake. Whereas, the validation LSTM is also classifying between two classes, the two movement classes previously defined.

Therefore, the LSTM was constructed as a single LSTM block, which contained 400 hidden units. It accepted an input dimension of 34, a sequence dimension of 5, and an output dimension of 1. The only slight difference between this LSTM and the discriminator was that of the final layer. In the discriminator, a non integer value on the classification was acceptable as the Generator could use this to learn. In this LSTM, a firmer decision on the classification of a sample was desired, so therefore the final layer was a fully connected layer, with a softmax activation function, so that the classification for any particular sample would be mutually exclusive of any other class. The classifier was trained with a batch size of 16 and for 20 epochs.

This validation using the classifier was carried out in two ways. The two methods sound very similar in practice, however they harbor different results and importantly, different insights that can be taken from the training in Ped-Cross GAN.

The first validation method was to train the classifier on the 1802 samples, which were kept from the Pedestrian Scenario dataset. This trained classifier was then used to test the 10,000 newly generated samples from Ped-Cross GAN. This will be called normal validation throughout the discussion.

The second validation was to train the classifier on the 10,000 newly generated samples from Ped-Cross GAN. Then this trained classifier was used to test how well it could classify the real 1802 samples kept aside from the Pedestrian Scenario dataset. This will be called reverse validation throughout the discussion.

6.2. Validation Results

6.2.1. Normal Validation

The method for normal validation was that of training the classifier on 1802 samples in the testing set from the Pedestrian Scenario dataset. The classifier was then tested on the generated samples from Ped-Cross GAN. The number of correctly classified samples would give an indication into the capability of Ped-Cross GAN.

Over the 20 epochs of training, it can be seen in Figure 6, that the classifier trains well, and begins to converge at around 10 epochs.

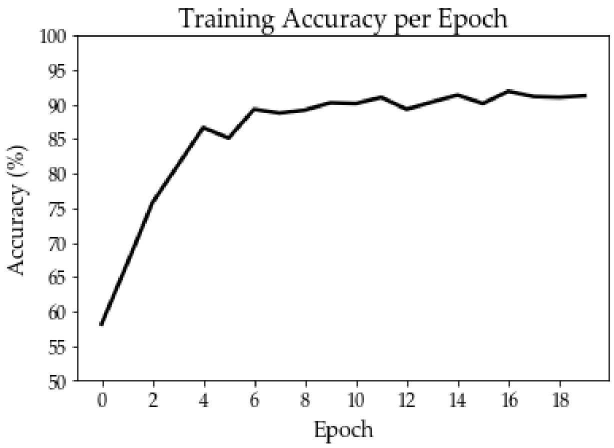

Figure 6. Classifier training for normal validation.

Following training on the 1802 real data samples, 10,000 generated samples from Ped-Cross GAN were classified. Figure 7 shows the results for the generated samples. It can clearly be seen that the generated results have performed well, scoring an overall correct classification rate of 99.2%. Out of the 10,000 samples tested, only 40 samples were classified incorrectly.

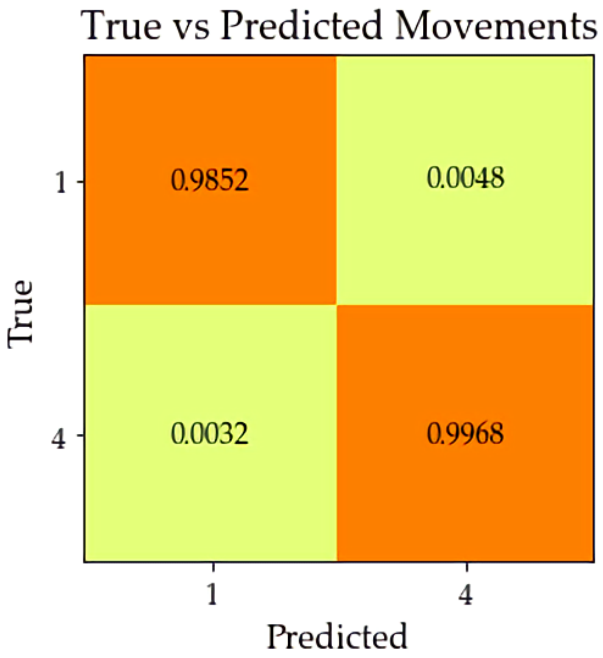

Figure 7. Results of generated/real samples on trained classifier.

Therefore, these results highlight that the Generator in Ped-Cross GAN was able to replicate the training distribution of the Pedestrian Scenario dataset well enough to score very highly when classified against never before seen data.

6.2.2. Reverse Validation

The method for reverse validation was to test the 1802 real samples from the Pedestrian Scenario dataset on a classifier trained entirely on data generated as a result of Ped-Cross GAN. The same 10,000 samples that were generated for the normal validation were used to train the classifier.

Figure 8 shows the accuracy of the model throughout training of the classifier. It can be seen that the classifier registers a classification accuracy of 100% after just 3 epochs, on a dataset of over 10,000 samples. This can be regarded as a somewhat surprising result, considering the number of generated samples that the classifier was trained on.

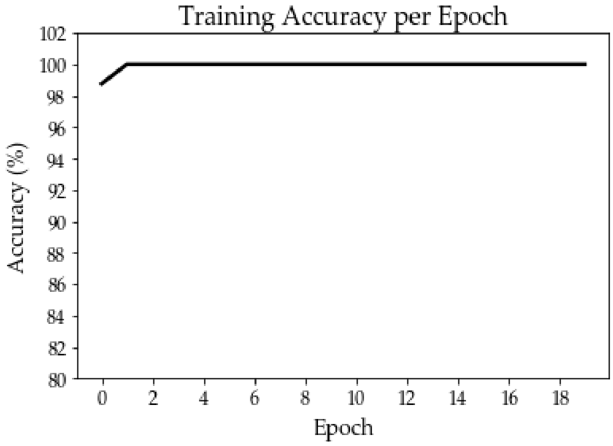

Figure 8. Classifier training for reverse validation.

In the opposite way as the normal validation, the reverse validation used the classifier trained on generated data to test the real data in the testing subset of the Pedestrian Scenario dataset. Figure 9 shows the confusion matrix of the results. It can clearly be seen that the results do not offer the same reflection on Ped-Cross GAN, as the results from the normal validation method. In this instance, the classifier correctly classified 1176 real poses correctly out of a possible 1802 and therefore misclassified 626 poses. This resulted in an overall accuracy of 65.26%.

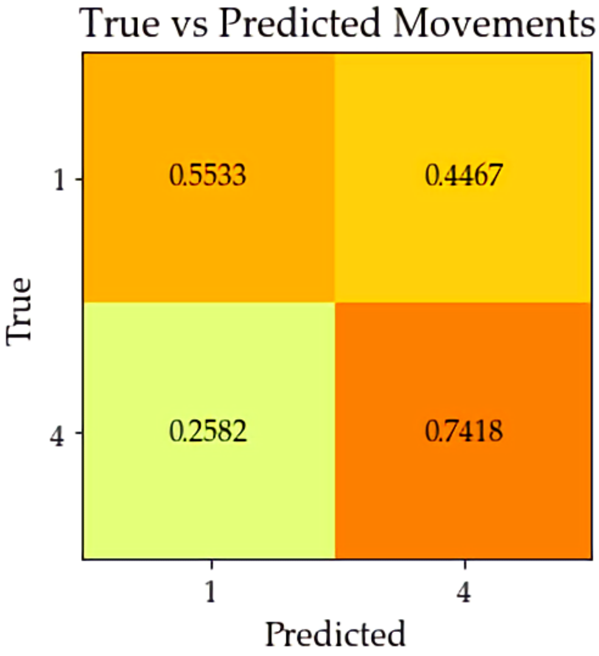

Figure 9. Results of real/generated samples on a trained classifier.

6.3. Visual Results

The results presented in Sections 6.2.1 and 6.2.2 show a disparity in results when attempting to validate Ped-Cross GAN in two very similar ways. One way to display some of the reasons behind these two results is to visually inspect some of the generated samples, and to try understand why they led to the results seen in Section 6.2.2.

The generated pose sequences in Figure 10 and 11 have been chosen by the authors of this paper. After viewing hundreds of samples, the pose sequences were cherry picked, with a view to give a good idea of what was observed in the generated pose sequences.

It is clear to see that the Generator in the Ped-Cross GAN has provided some errors, especially when looking at Class 1 (crossing from the left). Figure 10 shows some selected generated pose sequences. In the first four rows, the results are promising, where it clearly looks like a human crossing from the left. However, when consulting the final four rows, there are several issues that are clearly apparent. Specifically, row 6 appears to start the generated sequence well, before it encounters an error, which causes the human-like form to disappear.

The results in Figure 11 are far better. Unlike Class 1, Class 4 (crossing from the right) did not show any erratic visual errors. In all the visual trials, not one observed sample from Class 4 appeared to show any great error. On the one hand this is a good result, it means that some confidence can be had in Ped-Cross GAN and how it had applied its learning to its Generator. However, on the other hand, it is noted that there is a strong similarity between generated sequences for crossing from the right.

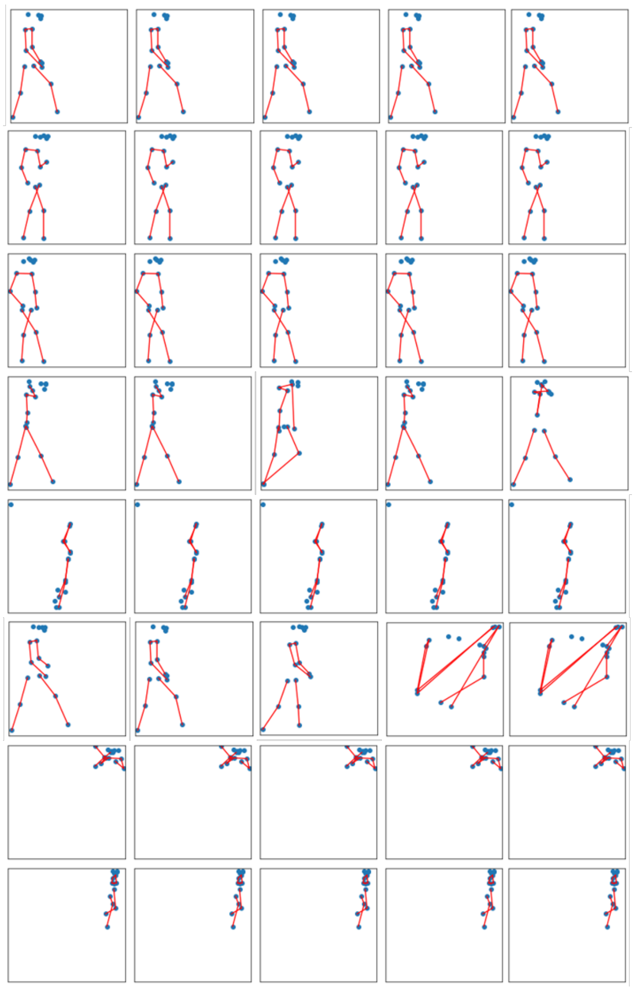

Figure 10. Selected generated poses for Class 1.

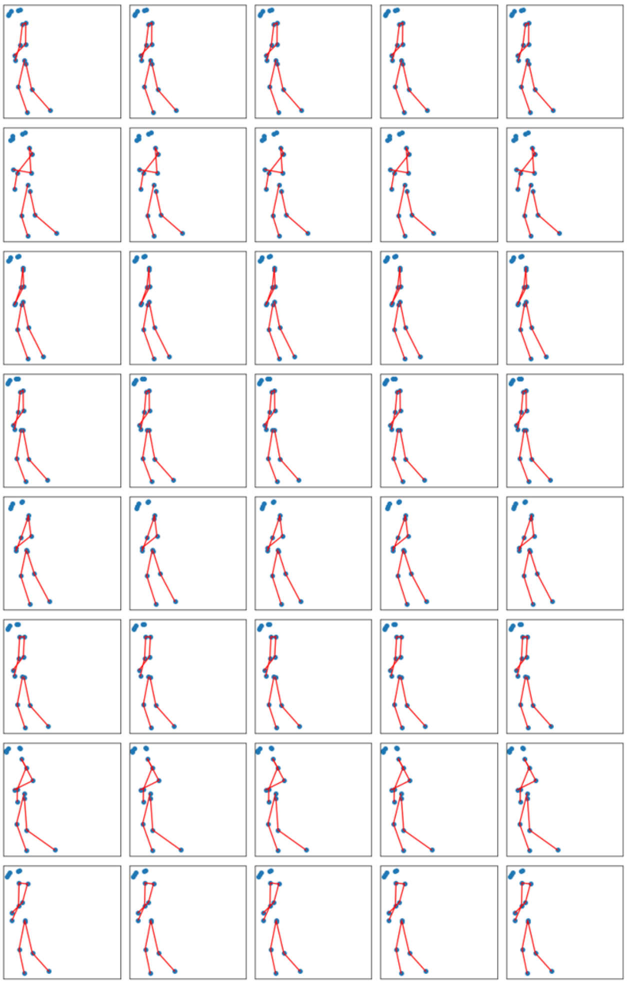

Figure 11. Selected generated poses for Class 4.

7. Discussion

The results presented in Sections 6.2.1 and 6.2.2 both show very different reflections for how the Ped-Cross GAN has been trained. This section will look into some of the reasons as to why such results were attained.

7.1. Normal Validation

The results for normal validation are good. Reporting a classification accuracy of 99.2% shows that the Ped-Cross GAN was able to excellently replicate the distribution of the Pedestrian Scenario Dataset. Therefore, if there is a confidence in the human nature of the poses in the dataset, then there is confidence in the human nature of the generated poses.

For the classifier to output a classification accuracy of 99.2%, the correct classification of so many of the erratic pose sequences, seen in Figure 10, cannot be attributed to chance. It is highly unlikely that the classifier would be able to guess the correct classification of these pose sequences to the degree of only making 40 errors out of 10,000 samples. It is, therefore, possible to conclude that these erratic pose sequences must already exist in the pedestrian scenario dataset and that the errors must be so prominent that examples of the errors were present in both the training and testing subsets. It is inefficient to visually check over 100,000 poses, a cursory check of the pedestrian dataset highlighted that there were these errors present in the dataset. It is therefore possible to conclude that the anatomical rules outlined in Section 3.1.2 are not stringent nor strenuous enough.

On the one hand, this conclusion can be regarded as negative, however it is also positive from the perspective of the training that occurred in the Ped-Cross GAN. For the

Ped-Cross GAN to be able to generate an array of different pose sequences, which were able to score 99.2% on the trained classifier, means that the training in the GAN can be regarded as a success. It could be construed that the erratic results, seen in Figure 10, are a reflection of how well the Ped-Cross GAN trained. Not only was it able to generate plausible, human-like pose sequences in rows 1–4, but the knowledge gained was such that it was also able to replicate the pose errors apparent in the dataset. This, therefore, means that when trained with the correct anatomical data, the Ped-Cross GAN would be capable of generating new pedestrian crossing scenarios which can be used for simulated testing of CAVs.

7.2. Reverse Validation

In normal validation, the classifier was trained on the full variety of sequences captured in the two classes. When tested on the generated samples, because a vast number of samples were very similar, this meant that they would have been captured by the same neurons in the vector space of the classifier. In other words, when the normal classifier correctly classified one sequence, it would also have no trouble in correctly classifying all of the other generated samples that were similar to the first. The same would apply for all slight variations in generated samples.

As we know, the reverse validation does the exact opposite of this. Now that it is known that many of the samples are similar to one another, it means that the variation in data available to the classifier is limited. Therefore, when it comes to testing the 1802 real samples, the variation in the test data is far richer than that of the generated pose sequences used to train the classifier. Thus, it struggled to effectively classify the real samples, resulting in an accuracy score of 65.26%.

For reverse validation, the results did not prove to be as good as those presented by the normal validation. This is due to a number of factors. The main factor identified relates to the variation, or lack thereof, in the generated results.

By consulting the visual results shown in Figure 11, it is clear that many of the generated sequences are very similar. Naturally, in a crossing scenario, so tightly labelled as is in the case of those used from the Pedestrian Scenario Dataset, it would be expected that the generated samples would look fairly similar. However, upon a visual inspection of the results, the generated pose sequences were far more similar than expected. It appears to show that the generated pose sequences are all very slight variants of the same precise movement.

It shows that the Ped-Cross GAN was able to learn the inherent nature of the class that it was asked to generate. However, it is apparent that it has simplified the knowledge learned to satisfy a few examples and is good enough to fool the discriminator during training, but is lacking the variation expected.

7.3. Training of Ped-Cross GAN

The failure to generate enough variation in the samples created the issues observed in the results. If the GAN is not able to vary the samples which are output, then the worth of the GAN is diminished. The advantage of using GANs is that you can generate new samples and data points based on the learning from real data. The generated data points should fall within the scope and distribution of the real data. The generator should be able to generate samples across the full range of distribution of a dataset, and not localize about a single area within the data distribution, as it appears to have done so in this training.

This is a common issue in the training of GANs and one that is not always simple to rectify. The explanation for the issue is a simple one. The generator, G, has a sole purpose to fool the discriminator, D. As is the same when playing or conducting any game, when the player devises a strategy to win the game, the player will continue to use this strategy, as it is proven to lead to success. The minimax game completed between G and D in a GAN is no different. When G has figured out a way to fool D, it will continue to focus on this route so that it can prolong its success. In the case of Ped-Cross GAN, the successful arrangement of data points to fool D appears in the form of the similar pose sequences, as seen in Figure 11.

The results presented in this paper have encountered this issue mainly due to over-simplifying, what is complex human pose sequence data, into just two classes. When this research progresses, and more classes are added from Table 4 and 5, this will enable almost every pose sequence to be different from each other. By being able to differentiate between the samples, the generator will have far more information to learn from, creating a multi-dimensional vector space, rather than the simple two-dimensional space created in this paper.

Further to this, amendments to the training parameters of the GAN can be made. Namely to introduce weight clipping and penalties to the generator when it over performs. This effectively means that each time G significantly fools D, G will be punished such that it is discouraged from continuing to force itself down the same route. This is akin to closing the door on that route, but not locking it, and given the correct input noise, G would still be able to use this route, however with other input noise, it will have to find another way.

The results from the Ped-Cross GAN have highlighted promising research avenues, as well as shone a light on aspects that need improvement. It is clear that the Ped-Cross GAN is very capable of learning the movement of a pedestrian in human pose form, and to then generate new pedestrian crossing scenarios based on a given label. Not only was the Ped-Cross GAN able to learn how to generate new pedestrian crossing scenarios, but it was able to do so in a manner which was indistinguishable from the Pedestrian Scenario dataset, as seen in Section 6.2.1. However, visual results have highlighted the need for more rigorous anatomical tests to be completed on the Pedestrian Scenario dataset. The anatomical tests used in this research are seen in Section 3.1.2.

8. Conclusions and Future Work

In the research area of autonomous vehicles and automated driving, the challenge of addressing pedestrian safety will always be a difficult task when implementing new technology. To implement such new technology on the roads used by millions of people everyday requires extensive testing, validation, and verification, all of which can only be completed with adequate data for the task at hand.

When testing such technology, there are a multitude of difficulties to consider, such as testing for rare events. It is not feasible, nor safe, to conduct real world testing of scenarios such as those including children or drunken pedestrians, therefore, these tests must be conducted in simulation. However, testing in simulation is not without its own issues. Hard coding a test by an engineer introduces a bias into that test, which limits the fidelity of a test to that very specific scenario, or the very issue of where the relevant scenarios come from. There is a need for CAVs to be able to identify the finer details of a pedestrian, such as intent and other factors that may affect their movement.

The research presented in this paper aims at tackling that issue. The Ped-Cross GAN has demonstrated the ability of a generative adversarial network to generate new pedestrian crossing scenarios in the form of human poses. The results highlighted in Section 6 show the encouraging performance in generating new pedestrian crossing scenarios, as well as some short comings that will become the basis of future research.

With the introduction of the Pedestrian Scenario dataset, the specific movements, behaviors, and descriptives of pedestrians is captured and extensively labeled. These additional labels take the datasets from simple video data to identify a pedestrian in an image, to an extensive dataset which will show insight into how pedestrians act and behave in a road environment.

After disseminating the results, it is clear that there are some issues that need addressing, namely the confidence in every human pose in the dataset being anatomically true. Given the poses generated and displayed in Figure 10, it is clear that the anatomical rules outlined in Section 3.1.2 were not extensive, nor strenuous enough.

Following the correction of the dataset, this research will progress to improve on the Ped-Cross GAN, and implement additional labels, as seen in Table 4 and 5. This could drastically improve the use case for the Ped-Cross GAN.

The ultimate objective for this research is to be able to train the Ped-Cross GAN to a point where a practitioner will be able to generate new pedestrian crossing scenarios which may have not been captured in video data. They will also be able to generate rare and important crossing events and scenarios, such as those involving elderly people, where the available data is limited.

It is envisaged that these generated pose sequences will be used in simulated environments, to test and validate the connected and autonomous vehicles, such that they are able to perform in a safe manner on our roads. This will enable the simulation of rare events, firstly, in a safe controlled environment and secondly, without ever needing to capture the rare event 'in the wild'.

Author Contributions: All authors have read and agreed to the published version of the manuscript. Conceptualization, S.K.; Data curation, J.S.; Formal analysis, J.S., V.P. and M.C.; Methodology, J.S. and M.C.; Project administration, V.P., S.K. and A.D.; Resources, V.P., M.C. and S.K.; Software, M.C.; Supervision, V.P., M.C., S.K. and A.D.; Validation, J.S. and A.D.; Visualization, J.S.; Writing—original draft, J.S.; Writing—review & editing, V.P., S.K. and A.D.

Funding: This research was funded by a joint initiative called the Centre for Connected and Autonomous Automotive Research (CCAAR). The joint initiative is between Coventry University, the Faculty Research Centre for Future Transport and Cities, and HORIBA MIRA Limited.

Institutional Review Board Statement: This study was conducted under the ethical approval of Coventry University, project code: P94776.

Informed Consent Statement: Not applicable.

Data Availability Statement: 20% of the Pedestrian Scenario Dataset will be freely available at https://ccaar.co.uk/projects/pedestrianscenariodataset/. The full Pedestrian Scenario dataset will be available upon request, and dependant on the use case, may incur a fee. Please contact the authors for more information.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. *Road Traffic Injuries*. Available online: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (accessed on 5 January 2021)
2. Singh, S. *Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey*; National Highway Traffic Safety Administration:Washington, DC, USA, 2015; pp. 1–2. Available online: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115 (accessed on 5 January 2021)
3. Combs, T.S.; Sandt, L.S.; Clamann, M.P.; McDonald, N.C. Automated Vehicles and Pedestrian Safety: Exploring the Promise and Limits of Pedestrian Detection. *Am. J. Prev. Med.* **2019**, *56*, 1–7. [CrossRef] [PubMed]
4. Griggs, T.; Wakabayashi, D. How a Self-Driving Uber Killed a Pedestrian in Arizona. *New York Times*, 2018.
5. Levin, S.; Wong, J.C. Self-Driving Uber Kills Arizona Woman in First Fatal Crash Involving Pedestrian. *The Gaurdian*, 2018.
6. Bradshaw, T. Self-Driving Cars under Scrutiny after Uber Pedestrian Death. *Financial Times*, 2018.
7. Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: A benchmark. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, Miami, FL, USA, 20–25 June 2009; pp. 304–311. [CrossRef]
8. World Health Organization. *Global Status Report on Road Safety—2015*; Technical Report; World Health Organisation: Geneva, Switzerland, 2005; ISBN 9789241564854.
9. European Commission. Road safety in the European Union—Trends, statistics and main challenges. *Mobil. Transp.* **2015**, 1–24. [CrossRef]
10. National Highway Traffic Safety Administration (NHTSA). *Traffic Safety Facts—Pedestrians*; 2018. Available online: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812493 (accessed on 5 January 2021)
11. Reynolds, S.; Tranter, M.; Baden, P.; Mais, D.; Dhani, A.; Wolch, E.; Bhagat, A. *Reported Road Casulatites Great Britain: 2016*; Technical Report September; Department for Transport: London, UK, 2007.

12. Dalal, N.; Triggs, W. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR05, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
13. Kotseruba, I.; Rasouli, A.; Tsotsos, J.K. Joint Attention in Autonomous Driving (JAAD) *arXiv* **2016**, arXiv:1609.04741.
14. Schneider, N.; Gavrila, D.M. Pedestrian path prediction with recursive Bayesian filters: A comparative study. *Lect. Notes Comput. Sci.* **2013**, *8142 LNCS*, 174–183. [CrossRef]
15. Zhang, L.; Lin, L.; Liang, X.; He, K. Is faster R-CNN doing well for pedestrian detection? *Lect. Notes Comput. Sci.* **2016**, *9906 LNCS*, 443–457. [CrossRef]
16. Li, J.; Liang, X.; Shen, S.; Xu, T.; Feng, J.; Yan, S. Scale-Aware Fast R-CNN for Pedestrian Detection. *IEEE Trans. Multimed.* **2018**, *20*, 985–996. [CrossRef]
17. Du, X.; El-Khamy, M.; Morariu, V.I.; Lee, J.; Davis, L. Fused Deep Neural Networks for Efficient Pedestrian Detection. *arXiv* **2018**, arXiv:1805.08688.
18. Rasouli, A.; Kotseruba, I.; Tsotsos, J.K. Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017, Venice, Italy, 22–29 October 2017; pp. 206–213. [CrossRef]
19. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.e.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 172–186. [CrossRef] [PubMed]
20. Fang, H.s.; Xie, S.; Tai, Y.w.; Lu, C. RMPE: Regional Multi-person Pose Estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2353–2362. [CrossRef]
21. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. *Lect. Notes Comput. Sci.* **2014**, *8693 LNCS*, 740–755. [CrossRef]
22. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693. [CrossRef]
23. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
24. Nash, J.F. Equilibrium points in n-person games. *Proc. Natl. Acad. Sci. USA* **1950**, *36*, 48–49. [CrossRef] [PubMed]
25. Fedus, W.; Rosca, M.; Lakshminarayanan, B.; Dai, A.M.; Mohamed, S.; Goodfellow, I. Many Paths to Equilibrium: GANs Do Not Need to Decrease a Divergence At Every Step. *arXiv* **2017**, arXiv:1710.08446.
26. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* **2017**, arXiv:1701.07875.
27. Spooner, J.; Cheah, M.; Palade, V.; Kanarachos, S.; Daneshkhah, A. Generation of pedestrian pose structures using generative adversarial networks. In Proceedings of the 8th IEEE International Conference on Machine Learning and Applications, ICMLA 2019, Boca Raton, FL, USA, 16–19 December 2019. [CrossRef]
28. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
29. Lin, X.; Amer, M.R. Human Motion Modeling using DVGANs. *arXiv* **2018**, arXiv:1804.10652.
30. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
31. Hinton, G.; Srivastava, N.; Swersky, K. Lecture 6e: Neural Networks for Machine Learning. *Coursera* **2014**. [CrossRef]