TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
PULCHOWK CAMPUS

THESIS NO: PUL075MSCSK004

Socially Aware Trajectory Prediction For Multi Pedestrian Environment

by
Bikram Acharya

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING
LALITPUR, NEPAL

August, 2021

Socially Aware Trajectory Prediction For Multi Pedestrian Environment

by

Bikram Acharya

PUL075MSCSK004

Thesis Supervisor

Associate Professor Dr. Dibakar Raj Pant

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer System and Knowledge Engineering

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk Campus

Tribhuvan University

Lalitpur, Nepal

August, 2021

COPYRIGHT©

DECLARATION

I declare that the work hereby submitted for Master of Science in Computer System and Knoweldge Engineering (MSCSKE) at IOE, Pulchowk Campus entitled **"Socially Aware Trajectory Prediction For Multi Pedestrian Environment"** is my own work and has not been previously submitted by me at any university for any academic award.

I authorize IOE, Pulchowk Campus to lend this thesis to other institution or individuals for the purpose of scholarly research.

Bikram Acharya
PUL075MSCSK004
Date: August, 2021

RECOMMENDATION

The undersigned certify that they have read and recommended to the Department of Electronics and Computer Engineering for acceptance, a thesis entitled **"Socially Aware Trajectory Prediction For Multi Pedestrian Environment"**, submitted by **Bikram Acharya** in partial fulfillment of the requirement for the award of the degree of **"Master of Science in Computer System and Knowledge Engineering"**.

.....

Supervisor: Assoc. Prof. Dr. Dibakar Raj Pant,
Department of Electronics and Computer Engineering,
Institute of Engineering, Tribhuvan University

.....

External Examiner: Om Bikram Thapa,
Chief Technology Officer,
Vianet Communications Pvt. Ltd

.....

Committee Chairperson: Assoc. Prof. Dr. Nanda Bikram Adhikari,
Department of Electronics and Computer Engineering,
Institute of Engineering, Tribhuvan

Date: AUGUST, 2021

DEPARTMENTAL ACCEPTANCE

The thesis entitled **"SOCIALLY AWARE TRAJECTORY PREDICTION FOR MULTI PEDESTRIAN ENVIRONMENT"**, submitted by **Bikram Acharya** in partial fulfillment of the requirement for the award of the degree of **"Master of Science in Computer System and Knowledge Engineering"** has been accepted as a bonafide record of work independently carried out by him in the department.

.....

Prof. Dr. Ram Krishna Maharjan

Head of the Department

Department of Electronics and Computer Engineering,

Pulchowk Campus, Institute of Engineering,

Tribhuvan University, Nepal.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to **Assoc. Prof. Dr. Dibakar Raj Pant** for his valuable and constructive guidance, appropriate direction and assistance, the stimulating suggestions and helpful counsel during the thesis period. This thesis work would not have been possible without his support and insightful advice, which has also motivated me to work on difficult problems in pedestrian trajectory prediction. His willingness to give his time so generously is highly appreciated.

I am very grateful and would like to extend my extraordinary appreciation and hearty thanks to **Assoc. Prof. Dr. Nanda Bikram Adhikari**, Program Coordinator, M.Sc. in Computer System and Knowledge Engineering, for his every bit of offered assistance amid the arrangement and preparation of this thesis work.

I would also like to extend my sincere thanks to Department of Electronics and Computer Engineering and faculty members for all the resources, their constant support and encouragement. Also, I would like to take a moment to thank my classmates and colleague for their cooperation throughout the development of this thesis.

Sincerely,
Bikram Acharya
PUL075MSCSK004

ABSTRACT

Human trajectory is the path that a human subject would most likely take to reach a specific destination. This route is estimated or predicted using pedestrian trajectory forecasting techniques. The key is to accurately encode observation sequence, model long-term dependencies from the past trajectories and forecast potential trajectories. Such models helps to learn social impact from other pedestrians, scene limits, and multi-modal possibilities of expected routes and can generalize to challenging scenarios and even output unacceptable solutions. In this thesis , a novel approach of effective hard sampling with contrastive learning to preserve motion representation, which captures desirable generalization properties with little computational overhead is achieved. Further, improving the quality of visual representations in socially aware pedestrian trajectory prediction. ETH-UCY a benchmark dataset, comprising of total 5 different sets ETH, Hotel, Univ, Zara1 and Zara2 and TrajNet++, another benchmark consisting ETH, UCY, WildTrack, L-CAS, and CFF dataset, are used for this thesis. Average Displacement Error (ADE), Final Displacement error (FDE) and Collision Avoidance Metric (CAM) are metrics used for performance evaluation. Experiments were carried out using real-world data and compared to state-of-art to assess the quality of the forecasting algorithm and the effectiveness of process. The result shows that proposed methododology with hard sampling has better collision avoidance in 3 of the 5 sets of ETH-UCY dataset with collision values of Hotel(0.07), Univ(2.62) and Zara1(0.04) compared to that of existing social-nce model Hotel(0.38), Univ(3.08) and Zara1(0.18). Similarly, for TrajNet++'s the result shows better collision avoidance with CAM values Directional-LSTM(4.42) and Social-LSTM(5.20) in comparison to state-of-art. The obtained results indicates a considerable improvement in the accuracy of trajectory predictions with better collison avoidance.

Keywords: Hard Negative Sampling, Pedestrian Trajectory, Contrastive Learning, Motion Representation, Multi Pedestrian Environment, Trajectron++, TrajNet++

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
RVO	Reciprocal Velocity Obstacle
ORCA	Ideal Reciprocal Collision Avoidance
ETH	Eidgenössische Technische Hochschule Zurich University
UCY	University of Cyprus
L-CAS	Lincoln Centre for Autonomous Systems
NMT	Neural Machine Translation
GRU	Gated Recurrent Unit
NCE	Noise Contrastive Estimation
ADE	Average Displacement Error
FDE	Final Displacement Error
D-LSTM	Directional LSTM
S-LSTM	Social LSTM
CVAE	Conditional Variational Autoencoder

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

With the advent in deep learning, routine activities such as pedestrian activity are attracting a lot of popularity as future tasks could be automated or helped by learning models. The multi-agent nature of this challenge adds to the complexities of trajectory prediction-a person must maintain his focus on a multitude of agents in their surrounding. The aim is to forecast potential trajectories of pedestrians (targets) based on their historical movement trajectories in a series(lists of continuous two-dimensional locations).

With time, researchers have achieved success with semantic segmentation [2], object identification [3]and image recognition [4, 5] among other computer vision activities. The estimation of potential trajectories of moving objects has been one of the most popular research subjects in recent years,, it's applications is growing in a variety of work. Human behaviour is incredibly complex and diverse. Pedestrian Route Prediction consists a variety of properties, but only a handful of them can be predicted based on previous trends. The aim of breaking down human behavior into distinct parts, such as mobility patterns, and studying each aspect independently is to decrease the complexity of the problem to a manageable subset. Mobility, as a component of human behavior, is often dynamic, but its variability is lower and may be analyzed with more oriented pattern-recognition techniques. In most situations, human mobility is studied in order to forecast future behavior. Mobility/Trajectory prediction thus is the prediction of people's next position in the area in which they normally travel.

Futher, when navigating in crowded environments, humans have an instinct capacity to predict the future movements of other individuals. To put it in another way, we can graspe the social etiquette of human movement, such as preserving personal space, giving right-of-way, and avoiding passing around individuals in the same

community. Our social experiences cause a variety of dynamic pattern-formation patterns in crowds, such as the introduction of pedestrian lanes with standardized walking directions and pedestrian movement oscillations at bottlenecks. Such capacity to navigate in social settings helps one to not only maintain a safe distance from others, but also to anticipate possible hazards and discomforts.

Trajectory can be presented as time-profile of pedestrian movements thus Trajectory forecasting is concerned with anticipating or determining the most possible direction a human subject will follow to reach their destination given a human subject and their destination. There are many uses for trajectory forecasting however, developing prediction models that are capable of doing so, on the other hand, is a difficult task.

As seen in Figure 1.1, it is possible to use object recognition strategies to address questions about people's positions and counts, but it is difficult to anticipate what will happen in the next frames automatically.

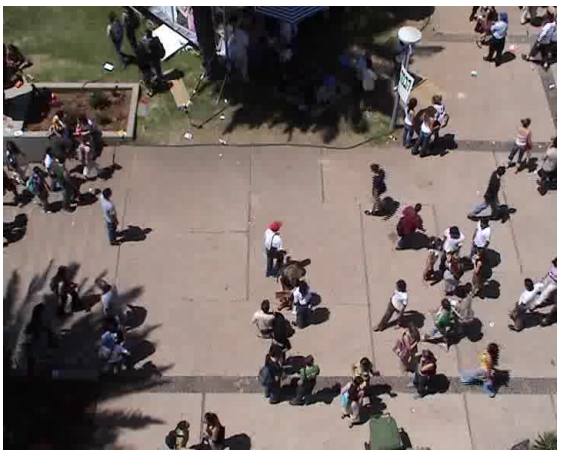


Figure 1.1: A frame of the UCY Pedestrian dataset.

Assuming that all human trajectories in a scene have been observed for some time, the aim of trajectory prediction is to project potential trajectories that corresponds to social norms. The key difficulties in forecasting pedestrians' paths are encoding observation sequence: *learning how to accurately model long-term dependence in the past trajectory, social impact from other pedestrians, scene limits, and multimodal possibilities of expected routes, as well since they fail to generalize to difficult circumstances and may provide unacceptable solutions.*

Capturing time sequence data and deriving useful information from them is one

of the most fascinating and difficult task in real time scenes and LSTM and RNN have been commonly applied to time sequence data for a variety of issues, including speech recognition, language processing, and machine learning. Similarly, literature shows ample research on extraction of features from human trajectories [6, 7], simulation of human-human/space social relationships [8]. This thesis work focuses on exploring hard negative sampling as data augmentation technique with contrastive learning which can be used to learn motion representation and train various architecture.

1.2 Problem Definition

There is a substantial body of work on pedestrian trajectory prediction; however, there are several areas that can be researched further. A slew of neural network-based models by [9, 10] for learning socially-aware motion representations have been extensively used and their utility for human trajectory forecasting [6, 7] in crowded environments have been demonstrated. Yet, current methods fail to yield desirable outcomes, posing significant safety concerns. The architecture of robust neural models revolves primarily around covariate shift. Also, most models use collected data as described in 3.2. These datasets do not contain enough scenes with complex situations, making it difficult to learn underlying social norms and generalize novel scenes. Similarly, other models uses non-expensive but infeasible interactive data gathering methods such as expert queries and interaction with the environment. Approach suggested by Alahi *et al.*[11] umakes advantage of previous knowledge of socially undesirable occurrences and exploit learning in a robust neural motion model. However, such learning techniques uses both positive and negative samples, which significantly increase batch sizes and computational overhead.

An effective approach may be to use hard negatives sampling with user controlled hardness. Such practical sampling technique captures desirable generalization properties, very little computational overhead and improved the quality of visual representations on image dataset, which is presented by Kalantidis *et al.* [12]. However, such comprehensive confrontation is still lacking in pedestrian trajectory

prediction. The following research topic for this thesis has been created based on the gaps in the literature:

Is hard negative sampling an effective data augmentation technique for pedestrian trajectory prediction? Can unsupervised sampling techniques be used to pick hard negative samples with contrastive learning and preserve motion representation? Will such combination yield promising result for unseen challenging events?

The purpose of this thesis is to offer evidence for the aforementioned research topic through relevant investigations and experiments.

1.3 Objectives

The primary goal of this thesis is to build hard negative sampling for pedestrian trajectory prediction as an approach for learning social motion representation between pedestrians, as well as to further investigate, analyze, and interpret pedestrian interaction with the environment. This objective is can be sub divided into sub-objectives which are:

- To implement hard negative sampling as data augmentation technique.
- To use hard negative sampling based contrastive learning for socially aware motion representation and improve pedestrian trajectory prediction.
- To evaluate the model using Average Displacement Error(ADE), Final Displacement Error(FDE) and Collision Avoidance Metric(CAM).

1.4 Scope of the Work

Many vital technologies, such as cars, social robots, visual monitoring, and so on, include the ability to comprehend busy scenes and anticipate human traffic patterns in a complex real-world context. This work can prove to be valuable for such context.

1.5 Organisation of thesis work

The rest of the thesis is organized as follows. Chapter 2 discusses the brief history of pedestrian trajectory predictions. This section briefly explores and analyzes different works by various authors that led to the establishment of this research work.

In Chapter 3, model's system architecture is defined and background theory is further examined. It then introduces dataset used for this thesis and the details regarding evaluation metrics are explained.

Chapter 4 summarizes the findings and comments gleaned from the research work. Following that, the results from proposed model is compared to the state-of-art.

Chapter 5 concludes this thesis. It highlights the present underlying concerns as well as the suggested work's limits. This chapter also gives information on potential future directions to pursue.

CHAPTER 2

LITERATURE REVIEW

In this chapter, some learned works that influenced the development of this thesis. The parts that follow include a number of strategies for forecasting trajectories in multi-agent settings.

Pedestrian trajectory prediction is a difficult job that has received much interest recently as its implementations have grown in importance. Since these fields have grown in importance and demand over time, approaches for addressing the issue of pedestrian trajectory prediction have advanced, shifting from physics-based models to data-driven models based on deep learning.

Helbing *et al.* [13] pioneered one of the first approaches to pedestrian behavior modeling, known as the Social Forces Model. The suggested method was built on handcrafted features that reflects various powers that work on the pedestrian and treated each pedestrian as a particle and proposed that their motion can be explained by three forces.: acceleration toward the target motion velocity, repulsive effect, and attracting effect.

Bonabeau *et al.* [14] proposed Agent-based modeling which suggested modeling a system as a collection of autonomous decision-making entities referred to as agents, which includes the social force concept and has been used to predict human behavioral behaviors.

Antonini *et al.* [15] modelled human interaction behavior using strong priors in a discrete decision system. Further, Crowd simulations make use of motion models and Funge *et al.* [16] used agent-based approaches for this purpose. These methods modeled each person uniquely, and in order to produce practical simulations, a thorough understanding of various agents was needed.

Physics-based pedestrian behavior simulation has improved over time, with the advent of sophisticated strategies such as BRVO[17], which draws on Reciprocal Velocity Obstacle RVO[18] and the Ideal Reciprocal Collision Avoidance(ORCA)[19].

RVO ensured safe and oscillation-free motion if each agent uses the same collision avoidance rationale. Further, different approaches to social interaction modeling have been explored such as Discrete Choice framework [20], continuum dynamics [21], Gaussian processes [22, 23].

Alahi *et al.* [24] produced Social Affinity Maps to estimate pedestrian destinations by connecting broken or unobserved trajectory. Yi *et al.* [25] used crowd clustering as a cue to better predict trajectory. Robicquet *et al.* [26] defined social sensitivity in order to categorize human motion into several navigation methods. These physics-based models, on the other hand, are constrained because they rely on hand-crafted functions and hence can only represent a fraction of all conceivable behaviors

Until 2016, the most popular tool for forecasting pedestrian trajectories was physics-based, but it is now possible to forecast potential trajectories using evidence. Using a data-driven approach entails understanding how people walk by training a machine learning algorithm for real-world pedestrian trajectories. Data-driven methods can explicitly extrapolate the laws and nuances of human walking behavior that would be difficult to formalize from data. Learning how people walk solely from observable trajectories necessitates three key components: a machine learning algorithm with sufficient representation capacity, an efficient optimization strategy, and a sufficient amount of real-world evidence. Deep learning models for pedestrian trajectory prediction in the literature depend primarily on the use of Recurrent Neural Networks (RNN), specifically Long Short-Term Memory (LSTM)[27] cells.

Alahi *et al.*[28, 29], motivated by the use of recurrent neural networks (RNNs) in a variety of sequence prediction tasks [28, 29], proposed Social LSTM[6] model was one of the first to use such an approach, pioneering the use of deep learning in pedestrian trajectory prediction. The pedestrian trajectory, along with social details, is fed into an LSTM in this model. To model social interaction, social knowledge is depicted as a grid of nearby pedestrians. Extraction of features from human trajectories [6, 7], simulation of human-human/space social relationships [8], and understanding the mutual activities of heterogeneous social actors [30] have been the subject of most of the recent existing trajectory prediction study.

More recently, with desire to instigate socially aware motion representations, a number of neural networks [31, 1] have been investigated. Several design choices like feature pooling [32, 7], attention mechanism [33, 34] and spatio-temporal graphs [35, 36] have shown promising results in crowded environment. However, the robustness of these approaches, on the other hand, remains a major problem. Similarly, other type of work introduces additional loss functions or extra value function for self correction.

Luo *et al.* [37] introduced extrapolated value function with conservative extrapolation which leads to lead to policies with self-correction. It introduced Value Iteration with Negative Sampling to initiate a reinforcement learning algorithm, which outperformed previous studies.

Zeng *et al.* [38] proposed the use of structured planning costs, which was then used to plan a safe maneuver based on the projected future distributions of actors. Thus introduced loss directly effected the output of models and Liu *et al.* [11] proposed an enhanced method that derive motion representation with no changes to the primary task; they draw negative samples at random, whereas we propose a more informed sampling method based on prior information.

Unlike [11] which used both positive and negative samples, which significantly increases batch sizes and computational overhead, our goal is to select hard negative samples similar to ground truth for motion representation. Selecting hard negatives samples with user controlled hardness is a practical sampling technique which captures desirable generalization properties with very little computational overhead.

CHAPTER 3

METHODOLOGY

The methodology chapter formulates solution to the problem that this research work seeks to solve.

3.1 System Block Diagram

Our proposal in solving the challenges presented in the previous section can be found in Figure 3.1. It describes model architecture and training approaches for predicting

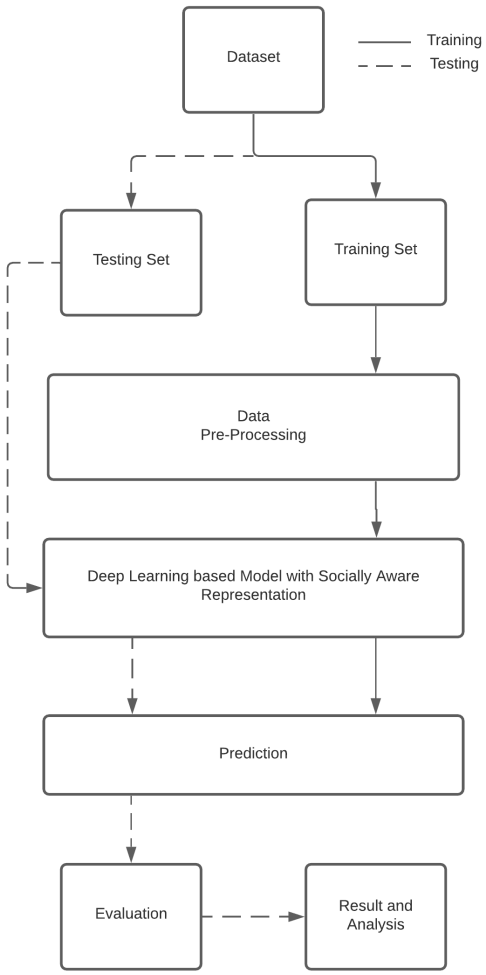


Figure 3.1: Methodology

pedestrian trajectories. The proposed architecture predicts future trajectories of pedestrians in a scene, given their previous motion states, by converting agent's past and present trajectories into encoded motion representations based on shared social information.

The methodology uses different encoding architecture to embed motion representation with socially aware interaction on augmented data generated using hard negative sampling. As human trajectories are complex, the proposed method could be well suited for trajectory predictions, and might help avoid possible hazards or discomforts.

3.2 Dataset

Datasets are essential for assessing a method, so significant effort must be made to ensure that the data gathered accurately represents the problem to be solved. The use of data-driven methodologies necessitates the availability of sufficient quantities of high-quality data. Most of the datasets for pedestrian trajectory prediction are generally available in two formats: Image coordinate and Real world coordinates. Image coordinate means pedestrians are represented by pixels location each pedestrian occupies in the camera image. However, the real world coordinates are the annotation of pedestrians by their position in meters with origin in an arbitrary point of world as shown in Fig. 3.2, where $P_i(X_t, Y_t)$ gives the annotated position for P_i pedestrian present in frame t and X_t, Y_t presents its real world coordinate.

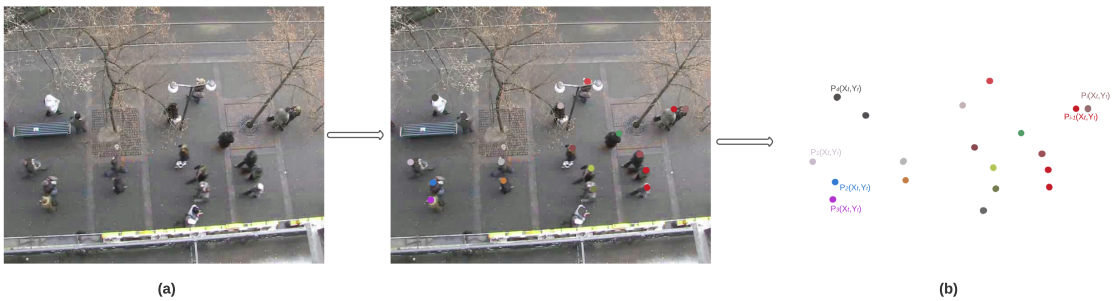


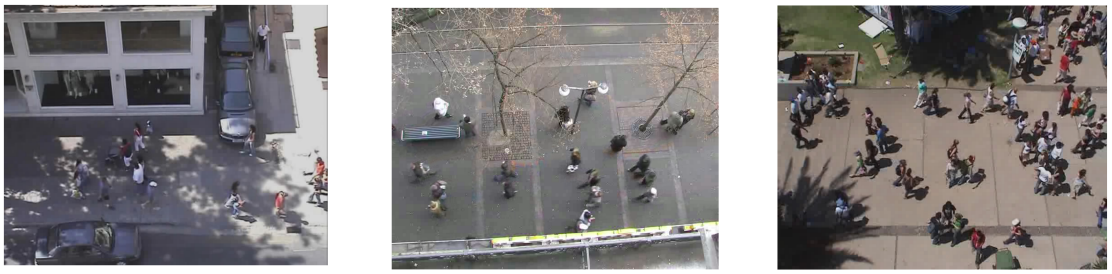
Figure 3.2: Dataset Generation (ETH-Hotel): (a)Frame t of Video (b) annotated by position in meters $P_i(X_t, Y_t)$

Depending on the use case of application, suitable coordinate format is selected.

Image coordinate contains high spatial information so they are primarily used in video surveillance application, however for application like autonomous driving and robotics, the annotated position is adequate thus uses real world coordinate. This thesis work focuses specifically on pedestrian trajectory and makes use of real-world coordinates. Thus, for this thesis work, our approach is evaluated on publicly available benchmark datasets with real-world coordinates i.e ETH[39]-UCY[40] dataset and TrajNet++ [41] dataset.

3.2.1 ETH and UCY Dataset

ETH-UCY is the most commonly used datasets in literature.Current state-of-art algorithms are evaluated on these datasets. ETH consists two scenes namely ETH; shows front of ETH Zurich's main building and HOTEL; the entry of a hotel in the city of Zurich, taken from bird's eyes view, where every frame contains annotations of pedestrian's position for every 0.4 sec and a total of 750 different pedestrian over 25 minutes. Similarly, UCY contains three scenes(Zara1,Zara2 and Univ) also captured using bird's eye view, which consists of 900 different pedestrians and their trajectories. Similar to ETH, every frame of this dataset is annotated with pedestrians positions. These datasets are used together which contains 5 subsets of real-world pedestrian trajectories widely used. These dataset preparation and preprocessing has been done using Trajectron++ [42], which is multi-modal trajectory forecasting models. Some frames of different dataset in ETH-UCY are shown in Fig. 3.3 below.



(a) UCY-Zara Dataset **(b)** ETH-Hotel Dataset **(c)** UCY-Univ Dataset

Figure 3.3: Images from different Datasets

The snapshot of real world coordinates of ETH-UCY dataset is shown in Fig. 3.4 below. Figure shows a snapshot of raw ETH data, where columns represents frame

number, pedestrian, x and y coordinates respectively

0	1	1.41	-5.68
0	2	0.51	-6.94
0	3	2.3	-4.59
0	4	2.74	-2.5
0	5	-1.59	0.93
0	6	-1.72	1.32
0	7	-2.45	2.26
0	8	-1.45	-0.76
0	9	1.16	-8.54
0	10	-0.11	-9.9
10	1	1.28	-6.35
10	2	0.55	-7.59
10	3	1.94	-4.12
10	4	2.76	-1.77
10	5	-1.59	0.93
10	6	-1.72	1.32
10	7	-2.43	2.46
10	8	-1.45	-0.76
10	9	1.18	-9.17
10	10	-0.1	-9.88
20	3	1.53	-3.49
20	4	2.73	-1.07

Figure 3.4: Snapshot of Real World Coordinate for ETH Dataset

3.2.2 TrajNet++ Dataset

Similarly, this work was also evaluated on TrajNet++ benchmark [1], which integrates several common pedestrian trajectory datasets and solely evaluates trajectories with pedestrian interactions.. TrajNet++ is a large-scale interaction-centered trajectory forecasting benchmark that includes concrete agent agent scenarios. TrajNet++ data is made up of trajectories taken from real-life movies and stored in the form of geographical coordinates and framerate of 2.5 frames per second is used. There are several freely accessible subsets of the dataset, including ETH, UCY, WildTrack[43], L-CAS[44], and CFF[24]. Details on TrajNet++ dataset is presented in Table 3.1 below.

Table 3.1: Dataset Details of TrajNet++

Dataset	Sub-Division	Tracks	Video (min)
ETH	ETH-HOTEL	∼650	25
	ETH-UNI		
UCY	UCY-ZARA	∼100	16
	UCY-STUDENT		
WildTrack	-	∼650	60
L-CAS	-	∼1100	49
CFF	-	∼42 million Trajectories	-

TrajNet++ also have a comprehensive assessment framework to put the gathered approaches to the test in order to make a fair comparison. Further more, by defining

a hierarchy of trajectory categorization, it is possible to properly index trajectories. Not only does detailed categorization enable to further sample trajectories, but it also allows us to gain insight into the model's success in various scenarios. In the TrajNet++, we have used standard train and test split for the communicating subcategory (Type-III). Type III sub category is a non-linear trajectory interaction scene in which primary trajectory undergoes social interactions. Interacting trajectories are further categorized into the following sub-categories (shown in Fig 3.5 for a more detailed categorization compatible with frequently encountered social experiences.

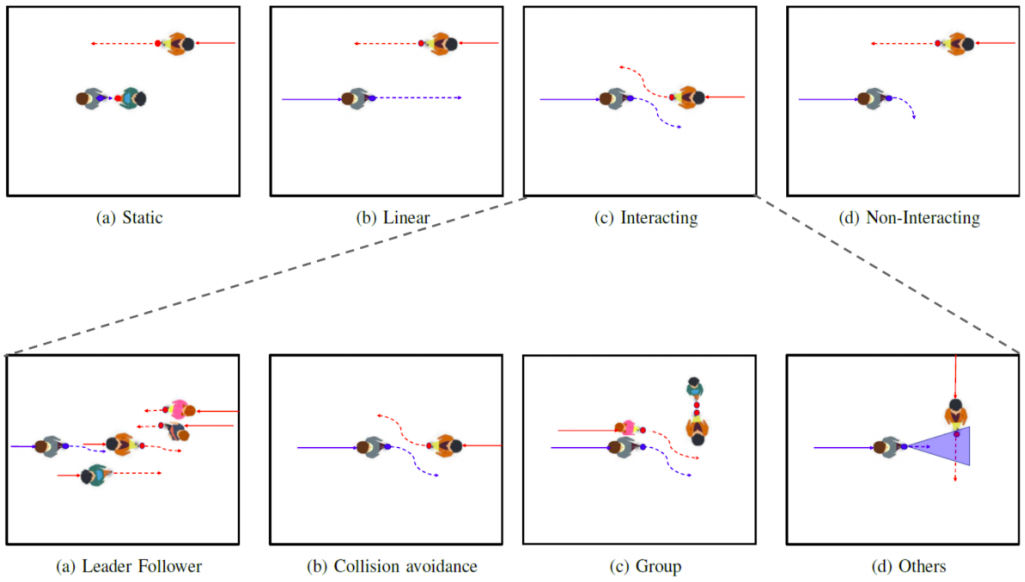


Figure 3.5: TrajNet++ Type-III Crowd Interactions in Real World

3.3 Preprocessing

The raw benchmark dataset explained in Section 3.2 is prepocessed and converted to structure that can be easily used with Trajectron++ and TrajNet++ tools. Brief discussion of preprocessing process for both dataset is described below.

3.3.1 ETH and UCY Preprocessing

The benchmark dataset ETH-UCY is publicly available in text format(.txt) and contains frame number, pedestrian number, X and Y real world coordinate of

the pedestrian for every frame as shown in Figure 3.4. Raw dataset is then converted using Trajectron to a pickle file containing the scene information for every frame. Scene information means the real world coordinate (position), velocity and acceleration of every pedestrian for every frame. The overall process of preprocessing is shown in Figure. 3.7 below.

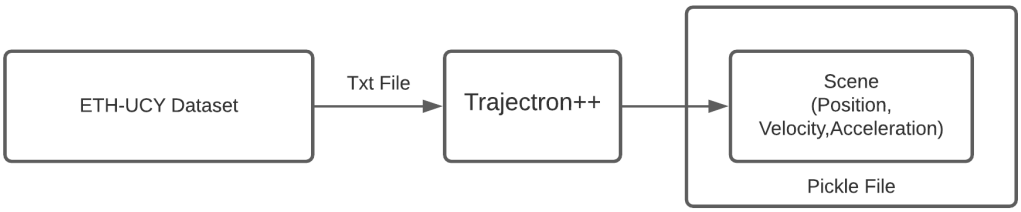


Figure 3.6: Preprocesing of ETH-UCY Dataset Using Trajectron++

3.3.2 TrajNet++ Preprocessing

The raw dataset consisting of ETH-UCY, WildTrack,L-CAS and CFF is publicly available in text format which is processed to ndjson dataset file by TrajNet++. NDJSON file contains Scene and Track information for every pedestrain in all frame and type of categorization it belongs to as shown Figure 3.5 . The Figure 3.7 below show the block diagram of preprocessing.

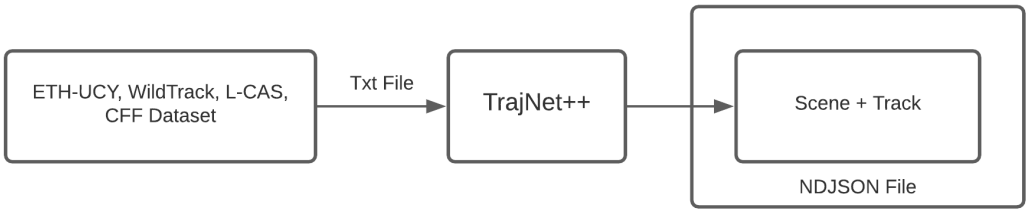


Figure 3.7: Preprocesing of ETH-UCY, WildTrack, L-CAS, CFF Dataset Using TrajNet++

The two different data representation i.e scene and track contains data as follows:

1. Scene

   The json structure of scene is as follows:

{“scene”: {“id”: 266, “p”: 254, “s”: 10238, “e”: 10358, “fps”: 2.5, “tag”: 2}}

where,

id	scene id
p	pedestrian ID
s,e	starting and ending frames id of pedestrian “p”
fps	frame rate
tag	trajectory type

Each scene has a primary pedestrian, which is identified by the scene's pedestrian ID. The scene is labeled (tagged) in relation to this primary pedestrian.

2. Track

   The json structure of track is as follows:

   {“track”: {“f”: 10238, “p”: 248, “x”: 13.2, “y”: 5.85, “pred_number”: 0, “scene_id”: 123}}

   where,

f	frame id
p	pedestrian ID
x,y	x and y coordinates in meters
pred_number	prediction number
scene_id	scene id

   X,Y coordinates are in meters of pedestrian ”p” in frame ”f”. Prediction Number is handy when offering several guesses rather than a single prediction. A maximum of 3 predictions are permitted. Scene ID is handy when offering predictions about various agents in the scenario, rather than just the principal pedestrian prediction

3.4 Deep Learning Models With Social Representation

3.4.1 TrajNet++ Benchmark [1]

Trajnet++ is a LSTM baseline with variety of interaction modules: *directional*, *occupancy* and *social pooling*. Among the various models two top-ranking uni-modal

forecasting models on different interacting categories of the Trajnet++ benchmark was selected i.e Social-LSTM [6] and Directional-LSTM [1].

Fig. 3.8 represents the model pipeline used by various interaction modules in TrajNet++, where sequence encoder is used to encode the past and present trajectory position of primary and secondary agents. Interaction encoder stores the encoded feature to capture interaction between pedestrian. Thus encoded social representation is passed to decoder. Depending on the architecture used by decoder, future trajectories are predicted.

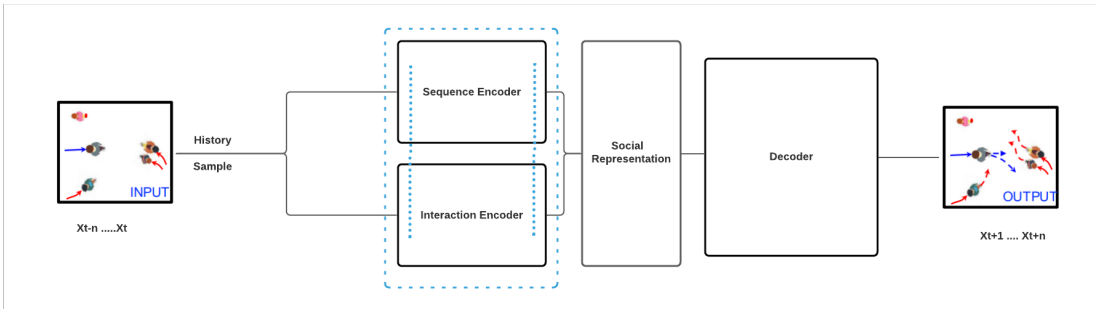


Figure 3.8: TrajNet++ Data Driven Pipeline

LSTM ENCODER DECODER

The LSTM Encoder-Decoder architecture has become a reliable and commonly used tool for neural machine translation (NMT) and sequence-to-sequence (seq2seq) prediction in general. The Fig 3.9 show the encoder decoder architecture which consists of 3 parts: encoder, intermediate (encoder states) vector and decoder.

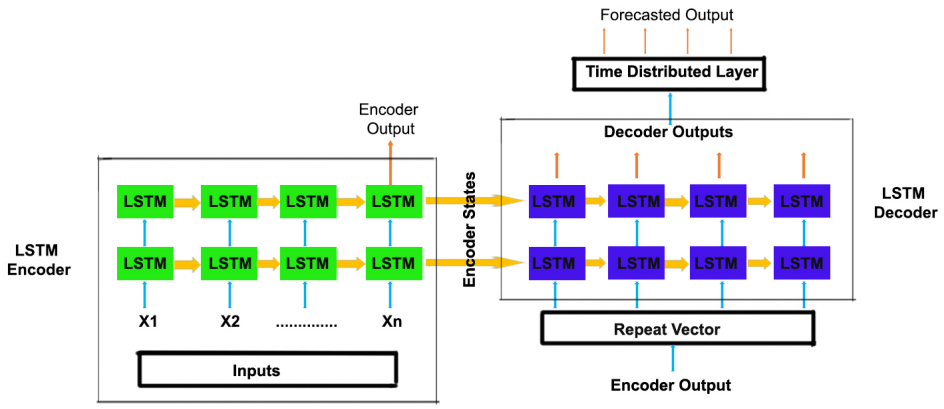


Figure 3.9: LSTM Encoder-Decoder Architecture

– **Encoder**

For increased performance, a stack of several recurrent units LSTM cells), each of which receives a single element from the input list, accumulates information for that element, and propagates it forward. The formula is used to compute the hidden states h_i:

$$\mathbf{h_t} = f(W^{(hh)}h_{t-1} + W^{(hx)}x_t) \tag{3.1}$$

– **Encoded States**

This is the encoder's final hidden state for the model. The aforementioned formula is used to calculate it. This vector aims to contain all information from the input elements in order to aid the decoder in generating proper predictions. For the decoder, it serves as the model's original hidden state.

– **Decoder**

A stack of periodic units, each predicting an output y_t at a time phase t. Each recurrent unit takes a hidden state from the preceding unit and creates an output as well as its own hidden state. Every hidden state $\mathrm{h}_i is computed using the formula.$:

$$\mathbf{h_t} = f(W^{(hh)}h_{t-1}) \tag{3.2}$$

3.5 Sequence Encoding

Since the model is trained and tested on using trajectron and trajnet++, the input trajectory sequence will be as shown in Fig 3.1, dataset block. Sequence Encoder encodes temporal information.

3.6 Interaction Encoder

This thesis work makes use of non grid based interaction model, which capture the social interactions in a grid-free manner, thus the spatial information is preserved.

3.7 Socially Aware Representation

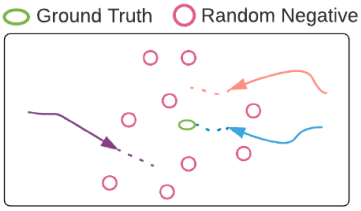
Socially aware representation tries to capture strong interaction between agents and attempts to learn social behaviours from data. This thesis work makes use of contrastive learning, a technique to learn an embedding space using similarity measures and selection of hard negative samples to approximate viable neighbourhood relationship.

- **Hard Negative Sampling**

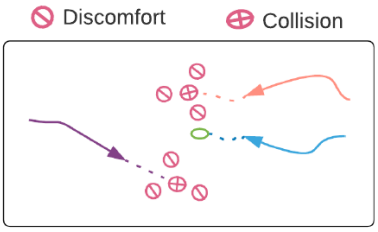
  It is likely that negative samples with a different label than the anchor but embedded close are the most helpful and give considerable gradient information during training. If the embedding believes the negative samples to be equivalent to the anchor at that moment, those are the most helpful negative samples.In addition to this, hardness is a crucial idea to keep in mind. Soft negatives have a higher learning signal than soft negatives, but they also inflict more harm than good by mimicking the repair of erroneous negatives.Gradually modifying the amount of hardness allows the user to find a compromise between enhanced learning signal and the harm produced. Due to their near proximity to the anchor, problematic locations have a high probability for sharing the same embedding. The embedded trajectories of principle actor(anchor) and other pedestrian in every scene is used for generating hard negative sample. Let every trajectory present in a scene represented by *traj(s)*, features of the scene is preserve using embedding function *f(x)* where function represents embedding in respect to anchor pedestrian x. Thus, $f(x^+)$ and $f(x^-)$ represents positive and negative embedding. The A suitable hardness control variable $\beta is selected using equation as shown below$: $\mathbf{L_{(f,q)}} = \mathbb{E}^{x\sim p}_{x^+\sim p^+_x}\Bigg[-log\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + \frac{Q}{-\tau}(\mathbb{E}_{x^-\sim q^\beta}[e^{f(x)^T f(x^+)}]-\tau^+\mathbb{E}_{v\sim q^+_\beta}[e^{f(x)^T f(x^+)}])}\Bigg]$(3.3)Fig. 3.10 shows the hard negative sampling and compares it to random sampling and social sampling used by social-nce [11].

Contrastive Representation Learning

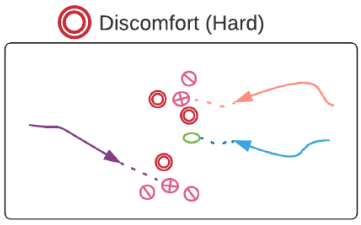
Learning a parametric function that maps raw data into a feature space to extract abstract and usable knowledge for downstream tasks is characteristic of representation learning[45]. To train an encoder, recent contrastive learning methods often



(a) Random Sample



(b) Social Sample



(c) Hard Negative Sample

Figure 3.10: Different Sampling Techniques

use the concept of noise contrastive estimation in an embedding space, namely the InfoNCE loss[46] given by equation below:

$$\mathbf{L_{NCE}} = -log\frac{exp(sim(q,k^+)/\tau)}{\sum_{n=0}^{N} exp(sim(q,k_n)/\tau)} \tag{3.4}$$

where the encoded query q is brought close to one positive key $k_0 = k^+$ and pushed apart from N negative keys $k_1, \ldots, k_N$, τ is a temperature hyperparameter, and $sim(u,v) = u^T v/(||u||||v||)$ is the cosine similarity between two feature vectors.

Equation 3.3 is further modified by SocialNCE[11], a variant of InfoNCE tailored for socially-aware motion representation learning. Thus, the loss function becomes:

$$\mathbf{L_{SocialNCE}} = -log\frac{exp(\psi(h_t^i)\cdot\phi(s_{t+\delta t}^{i,+},\delta t/\tau)}{\sum_{n=0}^{N}exp(\psi(h_t^i)\cdot\phi(s_{t+\delta t}^{i,n},\delta t/\tau)} \tag{3.5}$$

where for every embedding of history observations $q = \psi(h_t^i)$,$\psi(\cdot)$ is an MLP projection head and embedding of a future event $k = \phi(s_{t+\delta t}^i, \delta t)$, $\phi(\cdot)$ is an event encoder modeled by an MLP. Fig. 3.11 gives the hard social contrastive learning in multi-pedestrian context.

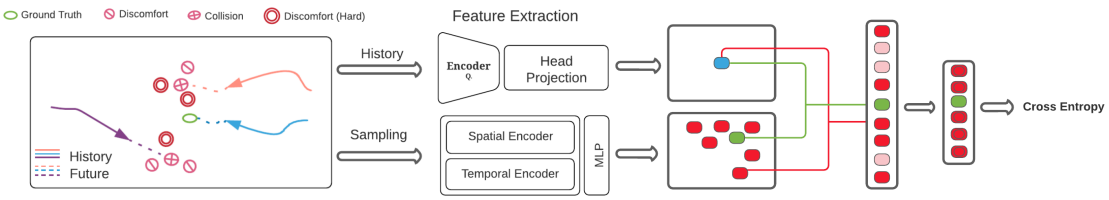


Figure 3.11: Hard Social Contrastive Learning in Multi Pedestrian Context

Thus, the overall loss function for this method becomes L_{Loss} which is given by equation below:

$$\mathbf{L_{Loss}} = L_{Task} + L_{SocialNCE(HardSample)} \tag{3.6}$$

where,

L_{Task} = Loss of model used i.e S-LSTM, D-LSTM and MultiModalCVAE[42] etc

$L_{SocialNCE(HardSample)}$ = Contrastive Loss using Hard Negative Samples

3.8 Evaluation Metrics

The most widely used metrics for human trajectory forecasting are as follows:

1. **Average Displacement Error (ADE)**

   This metric, like the one used in [1], calculates overall predicted time steps average L_2 distance between ground truth and model prediction. Fig. 3.12 shows the ADE calculation between predicted trajectory and ground truth.

   The formula for ADE calculation is as follows:

$$\mathbf{ADE} = \frac{\sum_{i=1}^{n}\sum_{t=T_{obs}+1}^{T_{pred}}[(\bar{x}_i^t - x_i^t)^2 + (\bar{y}_i^t - y_i^t)^2]}{n(T_{pred} - (T_{obs}+1)} \tag{3.7}$$

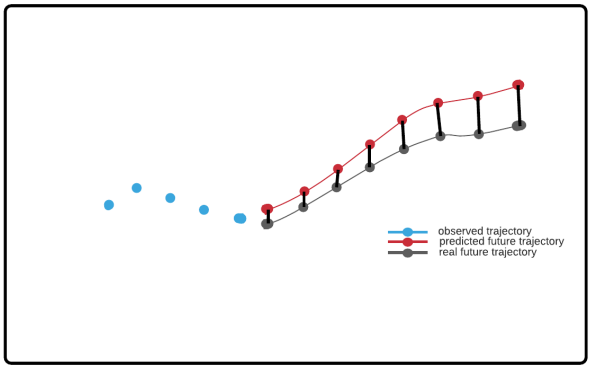


Figure 3.12: ADE Calculation

where,

n = number of pedestrian

T_{obs} = Observed Trajectories

T_{pred} = Predicted Final Destination

$\bar{x}_i, \bar{y}_i$ = predicted spatial coordinate x,y for pedestrian i

x_i, y_i = ground truth spatial coordinate x,y for pedestrian i

2. **Final Displacement Error (FDE)**

   At the completion of the forecast cycle, the distance between the predicted final destination T_{pred} and the ground truth destination. Fig. 3.13 shows the FDE calculation between predicted trajectory and ground truth.

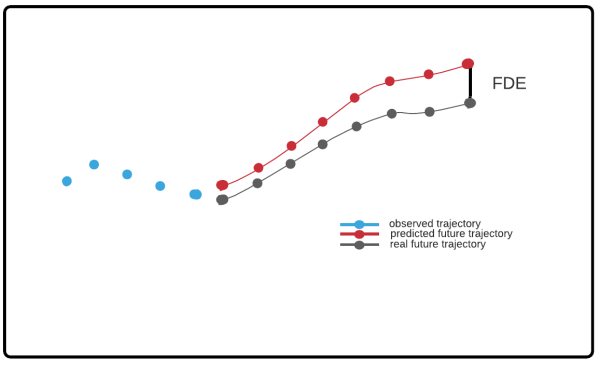


Figure 3.13: FDE Calculation

The formula for FDE calculation is as follows:

$$\textbf{FDE} = \frac{\sum_{i=1}^{n} \sqrt{(\bar{x}_i^{T_{pred}} - x_i^{T_{pred}})^2 + (\bar{y}_i^{T_{pred}} - y_i^{T_{pred}})^2}}{n} \tag{3.8}$$

where,

n = number of pedestrian

T_{pred} = Predicted Final Destination

$\bar{x}_i, \bar{y}_i$ = predicted spatial coordinate x,y for pedestrian i

x_i, y_i = ground truth spatial coordinate x,y for pedestrian i

3. **Collision Avoidance Metric**

   Collision I - Prediction collision (Col-I)

   This metric shows whether or not the expected model trajectories intersect, indicating whether or not the model knows the concept of collision avoidance.

   Collision II - Groundtruth collision (Col-II)

   For a ground truth potential scene, this statistic measures the percentage of collisions between the primary pedestrian's projection and the neighbors.

3.9 Tools Used

The tools and software's that will be used in this project work are listed below:

- Python Libraries: PyTorch, Numpy, Tensorboard
- Text Editor: Pycharm

CHAPTER 4

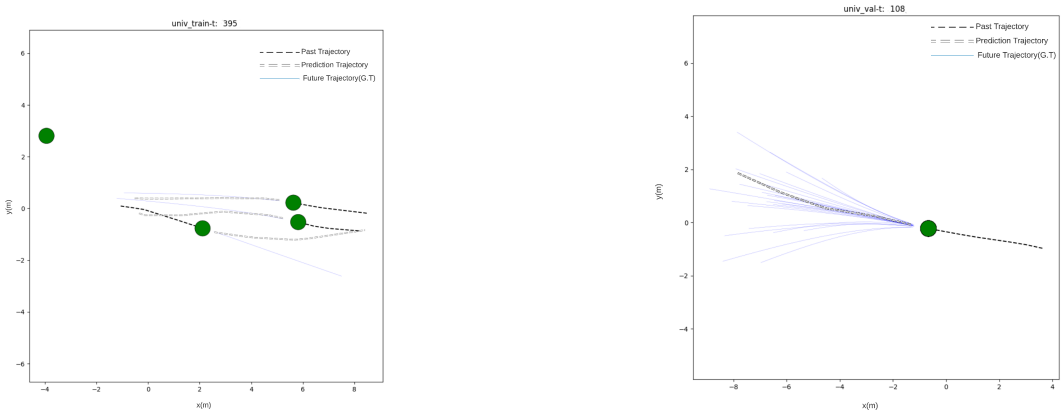
RESULTS AND DISCUSSION

4.1 Results and Discussions

The figure below shows trajectory results for different pedestrian, scene on different dataset.

Univ Dataset

Figure 4.1a a shows pedestrian present in frame 395 of dataset. Different color and line codes are done to represent past, prediction and ground truth trajectories. This frame was used for training the model. Similarly, figure 4.1b is the figurative representation of frame 108 and contains multiple ground truth because those trajectories are sampled for multi modal trajectories. This frame consists of single pedestrian.

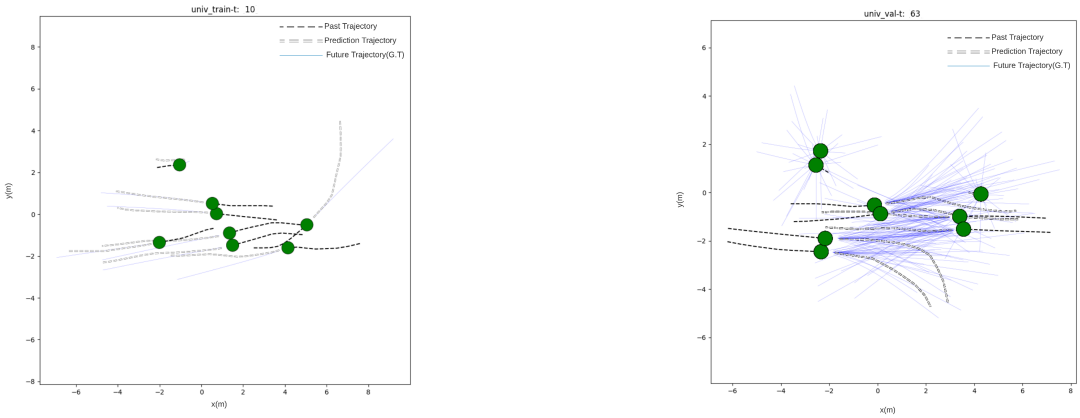


(a) Univ Dataset Training **(b)** Univ Dataset Evaluation

Figure 4.1: Univ Dataset Training and Evaluation for Few Pedestrian Environment

Figure 4.2a shows pedestrian present in frame 10 of dataset. Different color and line codes are done to represent past, prediction and ground truth trajectories. This frame was used for training the model. Similarly, figure 4.2b is the figurative representation of frame 63 and contains multiple ground truth because those trajectories are sampled for multi modal trajectories.

This frame consists of multiple pedestrian.

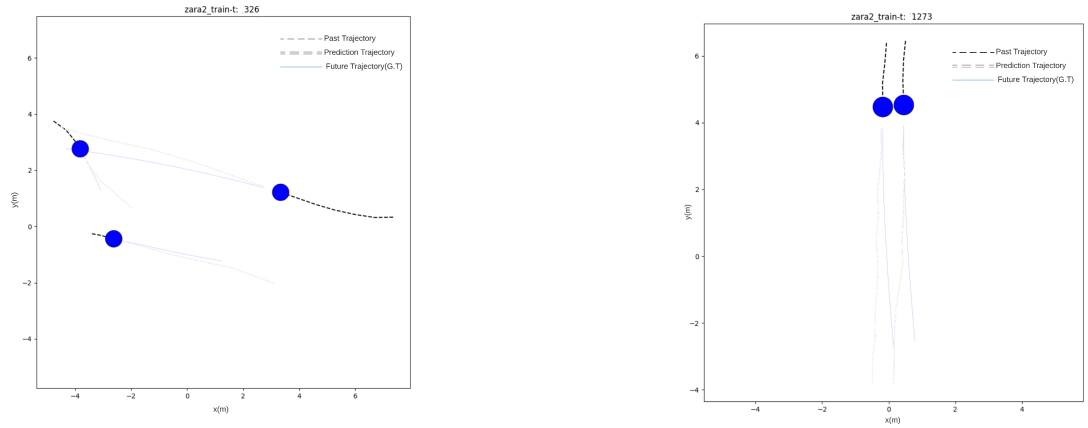


(a) Univ Dataset Training **(b)** Univ Dataset Evaluation

Figure 4.2: Univ Dataset Training and Evaluation for Multi Pedestrian Environment

Zara Dataset

Figure 4.3a a shows pedestrian present in frame 326 of dataset. Different color and line codes are done to represent past, prediction and ground truth trajectories. This frame was used for training the model. Similarly, figure 4.3b is the figurative representation of frame 1273 and contains multiple ground truth because those trajectories are sampled for multi modal trajectories.

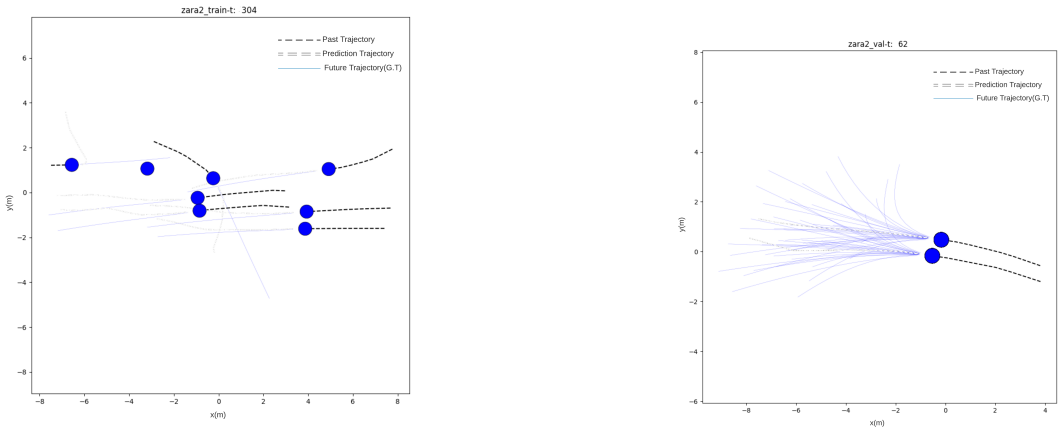


(a) Zara Dataset Training **(b)** Zara Dataset Evaluation

Figure 4.3: Zara Dataset Training and Evaluation for Few Pedestrian Environment

Figure 4.4a shows pedestrian present in frame 304 of dataset. Different color and line codes are done to represent past, prediction and ground truth

trajectories. This frame was used for training the model. Similarly, figure 4.4b is the figurative representation of frame 62 and contains multiple ground truth because those trajectories are sampled for multi modal trajectories.



(a) Zara Dataset Training **(b)** Zara Dataset Evaluation

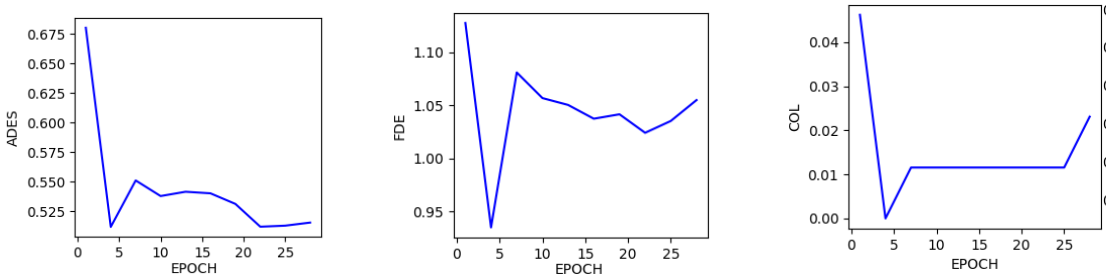
Figure 4.4: Zara Dataset Training and Evaluation for Multi Pedestrian Environment

4.2 Evaluation Metric

The evaluation of proposed architecture was done using ADE, FDE and Collison Avoidance Metric described above. The changes in ADE, FDE and Collision Avoidance is shown using curve below. The decreasing nature of all metrics in curves as shown in Figure [4.5,4.6,4.7,4.8,4.8] suggests that the model is learning.

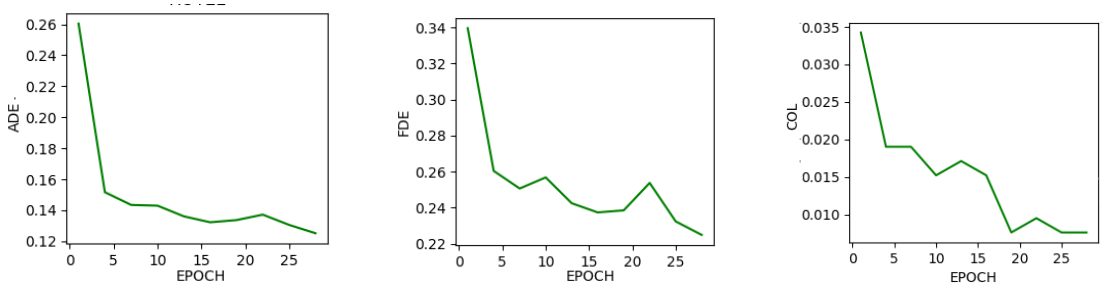
4.3 Validation

The metrics presented in Table 4.1 shows that proposed hard negative sampling performs better collision avoidance than social-nce model in all datasets except ETH , and ADE and FDE for all dataset are in considerable range with social-nce.



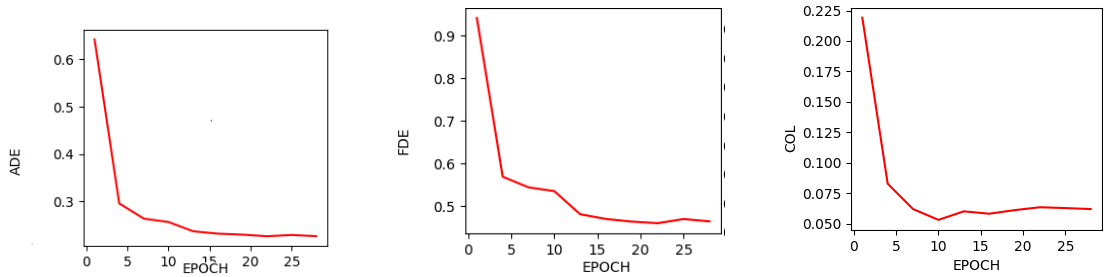
(a) ADE for ETH **(b)** FDE for ETH **(c)** CAM for ETH

Figure 4.5: Evaluation Metrics Curve For ETH Dataset



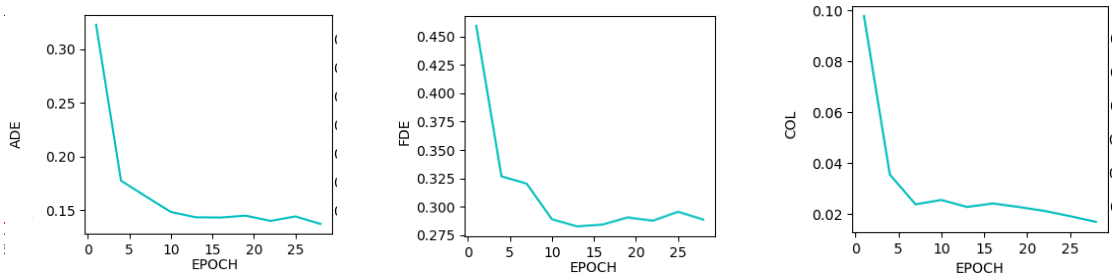
(a) ADE for HOTEL **(b)** FDE for HOTEL **(c)** CAM for HOTEL

Figure 4.6: Evaluation Metrics Curve For HOTEL Dataset



(a) ADE for UNIV **(b)** FDE for UNIV **(c)** CAM for UNIV

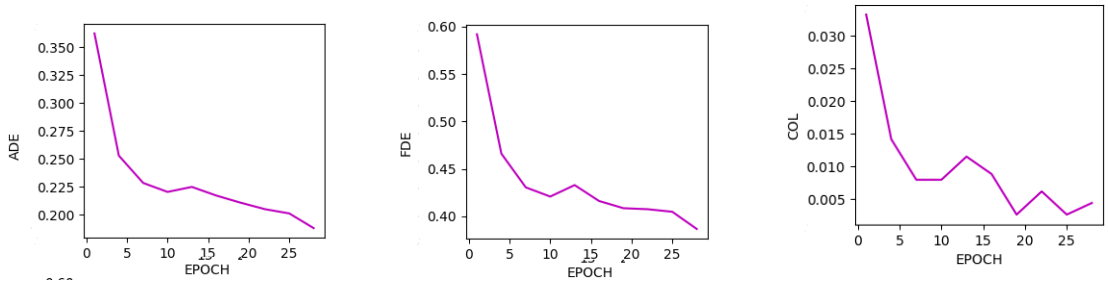
Figure 4.7: Evaluation Metrics Curve For UNIV Dataset



(a) ADE for ZARA1 **(b)** FDE for ZARA1 **(c)** CAM for ZARA1

Figure 4.8: Evaluation Metrics Curve For ZARA1 Dataset

Further, the method is validated on TrajNet++, the corresponding metrices are shown in Table 4.2



(a) ADE for ZARA2 **(b)** FDE for ZARA2 **(c)** CAM for ZARA2

Figure 4.9: Evaluation Metrics Curve For ZARA2 Dataset

Table 4.1: Comparison With Social-NCE using Trajectron++

Dataset	Social_NCE			Thesis Work		
	ADE↓	FDE↓	COL↓	ADE↓	FDE↓	COL↓
ETH	-	**0.71**	0.00	0.515	1.055	0.23
HOTEL	-	**0.177**	0.38	0.125	0.225	**0.3**
UNIV	-	**0.435**	3.08	0.227	0.465	**0.56**
ZARA1	-	0.330	0.18	0.188	**0.32**	**0.08**
ZARA2	-	**0.255**	0.99	0.137	0.289	1.70

Table 4.2: Comparison With Social-NCE and LSTM based encoder-decoder using TrajNet++

Model	Interaction	ADE↓ (m)	FDE↓ (m)	COL↓ (%)
LSTM-LSTM	Direction Pooling [1]	0.58	1.25	6.4
LSTM-LSTM	Social Pooling [6]	0.55	1.18	6.9
Social-NCE	Direction Pooling [1] + Contr. Learning [11]	-	**1.22**	4.59
Social-NCE	Social Pooling [6]+ Contr. Learning [11]	**0.53**	**1.14**	5.31
Thesis Work	Direction Pooling [1]	**0.56**	1.22	**4.42**
Thesis Work	Social Pooling [6]	**0.53**	1.15	**5.20**

4.3.1 Analysis

From the results as presented in Section 4.1, we can observe better collision avoidance performance by the proposed methodology. Further analyzing evaluation metrics and curves as shown in Section 4.3 and validating it with social-nce as presented in Table [4.2,4.1], we can conclude that taking hard negative samples that are similar to ground truth helps model to better learn motion representation, which in turn improves collision avoidance. Additionally, the proposed method outperforms social-nce on collision avoidance and ADE when both are implemented in TrajNet++ than Trajectron++ as shown by Table [4.2,4.1], this might be due

to characteristics of dataset. For this work we only used Type III categorization for TrajNet++, however Trajectron++ uses ETH-UCY dataset which contains all types of trajectories and padestrian.

CHAPTER 5

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

This thesis work proposed a methodology to effectively sample data using hard negative sampling as data augmentation technique. This work proposed an enhanced method that derives motion representation with no changes to the primary task, thus this method can be easily implemented along with any deeplearning models. From the result, it can be observed proposed method with hard sampling has better collision avoidance in 3 of the 5 sets of ETH-UCY dataset with collision values of Hotel(0.07), Univ(2.62) and Zara1(0.04) compared to that of existing social-nce model Hotel(0.38), Univ(3.08) and Zara1(0.18). Similarly, for TrajNet++'s the result shows better collision avoidance with CAM values Directional-LSTM(4.42) and Social-LSTM(5.20) in comparison to state-of-art. The obtained results indicates a considerable improvement in the accuracy of trajectory predictions with better collison avoidance. Thus, the experiment shows that it is a practical sampling technique which captures desirable generalization properties.

5.2 Limitation

The proposed method was experimeted on ETH-UCY and TrajNet++ dataset, and since TrajNet++ consists type III categorization of pedestrian and ETH-UCY consists all types of pedestrian experiment, the collision avoidance metric is better for TrajNet++, so further analysis can be done for pedestrian types other than III. Similarly, this work was limited to real pedestrian dataset, so the work can be further extended for autonomous driving prediction tasks. This thesis work only used encoder-decoder models, so the proposed method can be further implemented to other deeplearning models like Graph Neural Networks(Spatio Temporal Graph Neural Network) that are commonly used for trajectory prediction problems.

REFERENCES

[1] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective, 2021.

[2] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey, 2020.

[3] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A survey of deep learning-based object detection. *IEEE Access*, 7:128837–128868, 2019.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[5] Nilanjan Dey, Girish Mishra, Jajnyaseni Kar, Sayan Chakraborty, and Siddhartha Nath. A survey of image classification methods and techniques. 07 2014.

[6] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016.

[7] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks, 2018.

[8] Chiho Choi and Behzad Dariush. Looking to relations for future trajectory forecast, 2019.

[9] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6271–6280, 2019.

[10] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, S. Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks, 2019.

[11] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social nce: Contrastive learning of socially-aware motion representations, 2020.

[12] Yannis Kalantidis, Mert Bülent Sariyildiz, Noé Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *CoRR*, abs/2010.01028, 2020.

[13] Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, May 1995.

[14] Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280–7287, 2002.

[15] Gianluca Antonini, Santiago Venegas, Michel Bierlaire, and Jean-Philippe Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision*, 69, 08 2006.

[16] John Funge, Xiaoyuan Tu, and Demetri Terzopoulos. Cognitive modeling: Knowledge, reasoning and planning for intelligent characters. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 29–38, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[17] Sujeong Kim, Stephen Guy, Wenxi Liu, Rynson Lau, Ming Lin, and Dinesh Manocha. Predicting pedestrian trajectories using velocity-space reasoning. *The International Journal of Robotics Research*, 34, 01 2014.

[18] Jur van den Berg, Stephen J. Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In Cdric Pradalier, Roland Siegwart, and Gerhard Hirzinger, editors, *Robotics Research*, pages 3–19, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[19] Javier Alonso-Mora, Andreas Breitenmoser, Martin Rufli, Paul Beardsley, and Roland Siegwart. *Optimal Reciprocal Collision Avoidance for Multiple Non-Holonomic Robots*, pages 203–216. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[20] Gianluca Antonini, Michel Bierlaire, and Mats Weber. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8):667–687, 2006.

[21] Adrien Treuille, Seth Cooper, and Zoran Popovic. Continuum crowds. *ACM Trans. Graph.*, 25:1160–1168, 07 2006.

[22] Meng Keat Christopher Tay and Christian Laugier. *Modelling Smooth Paths Using Gaussian Processes*, pages 381–390. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[23] Raúl Quintero Mínguez, Ignacio Parra Alonso, David Fernández-Llorca, and Miguel Ángel Sotelo. Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition. *IEEE Transactions on Intelligent Transportation Systems*, 20(5):1803–1814, 2019.

[24] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2211–2218, 2014.

[25] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Understanding pedestrian behaviors from stationary crowd groups. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3488–3496, 2015.

[26] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 2016.

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[28] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator, 2015.

[29] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, Deva Ramanan, and Thomas S. Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2956–2964, 2015.

[30] T. Leung and G. Medioni. Visual navigation aid for the blind in dynamic environments. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 579–586, 2014.

[31] Alexandre Alahi, Vignesh Ramanathan, Kratarth Goel, Alexandre Robicquet, Amir A. Sadeghian, Li Fei-Fei, and Silvio Savarese. Chapter 9 - learning to predict human behavior in crowded scenes. In Vittorio Murino, Marco Cristani, Shishir Shah, and Silvio Savarese, editors, *Group and Crowd Behavior for Computer Vision*, pages 183–207. Academic Press, 2017.

[32] Nachiket Deo and Mohan Trivedi. Convolutional social pooling for vehicle trajectory prediction. pages 1549–15498, 06 2018.

[33] Changan Chen, Yuejiang Liu, Sven Kreiss, and Alexandre Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning, 2019.

[34] Riashat Islam, Komal K. Teru, Deepak Sharma, and Joelle Pineau. Off-policy policy gradient algorithms by constraining the state distribution shift, 2019.

[35] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. pages 2375–2384, 10 2019.

[36] V. Kosaraju, Amir Sadeghian, Roberto Martín-Martín, I. Reid, S. H. Rezatofighi, and S. Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *NeurIPS*, 2019.

[37] Yuping Luo, Huazhe Xu, and Tengyu Ma. Learning self-correctable policies and value functions from demonstrations with negative sampling, 2019.

[38] Wenyuan Zeng, Shenlong Wang, Renjie Liao, Yun Chen, Bin Yang, and Raquel Urtasun. Dsdnet: Deep structured self-driving network, 2020.

[39] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268, 2009.

[40] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. *Computer Graphics Forum*, 26(3):655–664, 2007.

[41] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective, 2021.

[42] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data, 2021.

[43] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

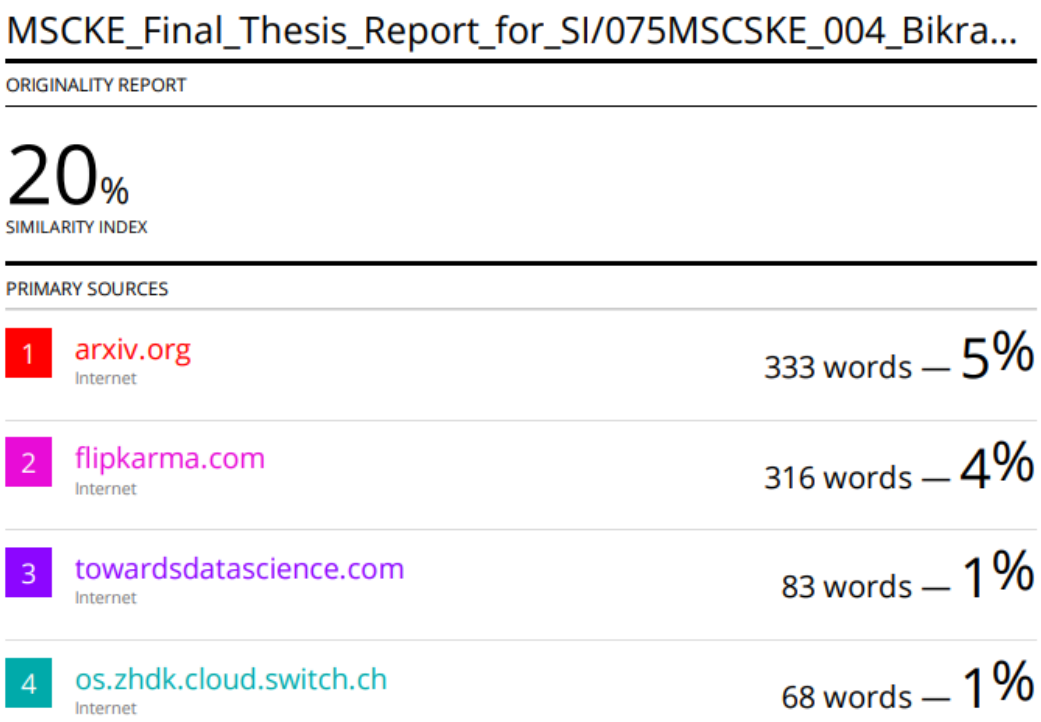
[44] Li Sun, Zhi Yan, Sergi Molina Mellado, Marc Hanheide, and Tom Duckett. 3dof pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data, 2017.

[45] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014.

[46] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

APPENDIX A

A-1 Similarity Index


MSCKE_Final_Thesis_Report_for_SI/075MSCSKE_004_Bikra...

ORIGINALITY REPORT

20%
SIMILARITY INDEX

PRIMARY SOURCES

1	arxiv.org Internet	333 words — 5%
2	flipkarma.com Internet	316 words — 4%
3	towardsdatascience.com Internet	83 words — 1%
4	os.zhdk.cloud.switch.ch Internet	68 words — 1%