Intention Prediction of Pedestrians in Challenging Weather Conditions using Deep Learning

by

Ahmed Elgazwy

A thesis submitted to the School of Graduate and Postdoctoral Studies in partial fulfillment of the requirements for the degree of

Master of Applied Science

in

Electrical and Computer Engineering

Faculty of Engineering and Applied Science

University of Ontario Institute of Technology
(Ontario Tech University)

Oshawa, Ontario, Canada

August 2023

THESIS EXAMINATION INFORMATION

Submitted by: **Ahmed Elgazwy**

Master of Applied Science in Electrical and Computer Engineering

Title: Intention Prediction of Pedestrians in Challenging Weather Conditions using Deep Learning

An oral defense of this thesis took place on August 18^{th}, 2023 in front of the following examining committee:

Examining Committee

Chair of Examining Committee	Dr. Zeinab El-Sayegh
Research Supervisor	Dr. Khalid Elgazzar
Research Co-Supervisor	Dr. Alaa Khamis
Examining Committee Member	Dr. Sanaa Alwidian
Thesis Examiner	Dr. Jing Ren

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

Abstract

Assisted and automated driving vehicles have received massive attention over the past few years from the research community to make our roads safer. In this thesis, we introduce a framework for predicting the intention of pedestrians in clear and challenging weather conditions. The framework consists of five deep-learning models, of which two are designed and trained from scratch and three were used pretrained. The framework takes video frames from the dashcam and inputs them to an enhancement pipeline to determine the quality of the images and enhance them if necessary. Then, the framework utilizes pre-trained models (MoveNet, Deep-sort, and Deep-Labv3) for feature extraction. Lastly, all the features are fed into a Transformer-based Intention Prediction Model (TIPM) for pedestrian intention prediction. Results show that TIPM outperforms state-of-the-art models yielding an accuracy of 69% on the JAAD behavior dataset, 82% on the JAAD all dataset.

Keywords: Deep-learning; Image-enhancment; Transformers; Vision transformer; Intention prediction; Assisted and automated driving vehicles.

Acknowledgements

First and foremost, I wish to express my deep appreciation to my supervisors, Dr. Khalid Elgazzar and Dr. Alaa Khamis. This journey's success owes much to their indispensable aid, perceptive insights, unwavering encouragement, and expert guidance. Their mentorship has profoundly influenced both the trajectory and caliber of the work expounded in this thesis.

I also extend my gratitude to my peers in the Internet of Things research lab under Dr. Khalid's guidance. Their collaborative discussions, motivating spirit, and camaraderie have significantly enhanced the gratification and value of the research endeavor.

A special acknowledgment is reserved for my friends: Khalid Atef, Abdalrahman Alsaka, Omar Sameh, Ahmed Shoukry, Karim Alrefaay, and Omar Abdelaziz. I am truly fortunate to have had such steadfast companions who stood by me through every hurdle, celebrated the modest triumphs, and provided solace during challenging times.

Lastly, and of paramount importance, my heartfelt thanks go to my family. My achievements owe immeasurably to the support and direction offered by my mother, Eman, and my younger brother, Mohamed. Their motivational words and devout prayers illuminate my path and their presence bestows upon me fortitude and solace. To my mother, I owe an eternal debt of gratitude for

her unswerving belief in my capabilities and the countless sacrifices made to ensure my triumphs.

"And they ask you about the Spirit. Say, "The Spirit belongs to the domain of my Lord, and you were given only a little knowledge."" [Quran,Surah al-Isra':85].

Dedication

I find it fitting to offer this work in honor of my late father, Prof. Abdelsattar Elgazwy. Throughout my upbringing, he exerted the most profound influence on me, shaping my ardor for science and research. While I yearn for his physical presence, his life's voyage continues to serve as a wellspring of inspiration, and the cherished recollections I hold with him shall forever reside within my heart. May his soul rest in eternal peace.

Author's Declaration

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Ontario Institute of Technology (Ontario Tech University) to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize the University of Ontario Institute of Technology (Ontario Tech University) to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

_____ Ahmed Elgazwy

Statement of Contributions

The contributions that accompany this thesis include two conference papers under process, a submitted journal paper in IEEE on intelligent vehicles, and a journal paper under process. This is described in more detail below.

1. Referred Conference Proceedings:

Ahmed Elgazwy, Hossameldin Ouda, Ammar Elmoghazy, Ghadeer Abdelkader, Austin Page, Khalid Elgazzar, and Alaa Khamis "On Image Enhancement for VRU Detection in Challenging Weather Conditions" Under process. In this paper, the authors introduced a deep-learning image enhancement pipeline, the pipeline included two stages: image classification and image enhancement. For image classification, a multilabel classifier was designed and trained on a custom dataset to detect root problems presented in a frame. For image enhancement, 3 classical image enhancement techniques based on kernels and histogram equalization were used. The multilabel classifier achieved an 80% accuracy on the test dataset and results showed that the pipeline improved both user view and detection precision for pedestrians. This paper is a part of the intention prediction framework presented in Chapter 3, and the results are included in Chapter 4.

2. Referred Conference Proceedings:

Ahmed Elgazwy, Somayya Elmoghazy, Khalid Elgazzar, Alaa Khamis "Pedes-

trian Crossing Intent Prediction using Vision Transformers" Under process. In this paper, the authors introduced a transformer-based intention prediction model. The model utilizes both a transformer encoder and a vision transformer for the task of visual and non-visual feature processing for the intention prediction task. the architecture relies only on a global self-attention mechanism and completely removes recurrent neural networks (RNNs) and convolutional neural networks (CNNs) presented in other models. This modification improves testing time and overall accuracy. Results show that the attention-based approach outperforms baseline models such as single-RNN and multi-RNN and state-of-the-art models such as PCPA and Mask-PCPA. This paper is the basis of the intention prediction framework proposed in Chapter 3. Results are also introduced in Chapter 4.

3. Referred Journal Proceedings:

Ahmed Elgazwy, Khalid Elgazzar, Alaa Khamis "A framework for pedestrian crossing intention prediction in challenging weather conditions using self-attention", under review in IEEE transactions on intelligent vehicles. This paper is an extension of the previous papers "Pedestrian Crossing Intent Prediction using Vision Transformers" and "On Image Enhancement for VRU Detection in Challenging Weather Conditions". In this paper, the authors further expanded their experimentation on the proposed attention model to other attention mechanisms such as local self-attention, and propose a variant of the model that utilizes only vision transformers. combine the image enhancement pipeline with the intention prediction model and test the model on the PIE dataset to test the model robustness. The paper also deploys the proposed pipeline to test the end-to-end inference time. This paper consists of the full framework that's presented in Chapter 3, and its results presented

in Chapter 4.

4. Referred Journal Proceedings:

Ghadeer Abdelkader, Ahmed Elgazwy, Taghreed Alghamdi, Khalid Elgazzar, and Alaa Khamis, "Will they Cross? A Comparative Study on Vision-Based Deep-learning Techniques For Pedestrian Intention Prediction" under review in ACM Journal on Autonomous Transportation Systems, in this paper, the authors present a comprehensive comparison between popular deep learning models used for pedestrian intention prediction. Each model is briefly explained, and tested on both JAAD and PIE datasets. A comparison between each model's accuracy, AUC, F1 score, precision, recall, and inference speed is highlighted. This paper is submitted to. This paper covers numerous aspects of background studies in intention prediction, which are presented in Chapter 2.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Ensuring the safety of vulnerable road users (VRUs), such as pedestrians and cyclists, is a critical aspect of intelligent transportation systems. Accurately predicting VRU's crossing intentions can significantly contribute to the prevention of accidents and the optimization of traffic flow. However, the reliability of such predictions can be compromised under adverse weather conditions, where reduced visibility and altered environmental factors pose significant challenges. In this context, the utilization of image enhancement techniques becomes paramount in maintaining the prediction accuracy of VRU's crossing intentions during inclement weather.

Extensive research was done on VRU intention prediction. Kotseruba et al. [17] provided the most recent benchmark for pedestrian intention prediction in which the Pedestrian Crossing Prediction with Attention (PCPA) model achieved state-of-the-art performance. After that different approaches and architectures were proposed to enhance the performance of PCPA [29, 39, 42]. The proposed work either used non-visual features only for intention prediction [29, 42] or a combination of visual and non-visual features [17, 39]. However, apart from Zhang et al. [42] who used graph neural networks, all of these models used CNN and RNN for visual and non-visual feature extraction respectively. Despite the utilization of various image enhancement techniques to rectify defects in input frames, previous research endeavors have not put forth any approach to uphold prediction accuracy in adverse weather conditions. This challenge poses a substantial concern for two principal reasons: timing limitations and the occurrence of multiple weather conditions concurrently. Presently, not a single model exists that can effectively address the diverse

range of defects that may manifest within a single video frame (i.e., image).

In this thesis, we target the use of vision transformers and self-attention mechanisms to design a pedestrian intention prediction model and utilize the use of CNNs and classical image processing to build an efficient and robust pipeline that can improve or at least maintain prediction accuracy with minimal computational and latency overhead.

1.2 Motivation

Accurate prediction of the crossing intention of VRUs is of utmost importance in the context of assisted and automated driving vehicles. Understanding VRU crossing intentions enables these vehicles to proactively respond to potential hazards, ensuring the safety of both the VRUs and the vehicle occupants. By predicting whether a pedestrian, cyclist, or other VRU intends to cross the road, assisted and automated driving systems can adjust their speed, trajectory, and braking in advance, allowing for timely and appropriate actions. This capability significantly reduces the risk of collisions and enhances the overall safety of road users. However, adverse weather conditions can adversely affect the accuracy of VRU crossing intention predictions, leading to potential safety hazards. In this regard, image enhancement techniques play a crucial role in improving the accuracy of prediction models.

Extensive research was done on the task of predicting the crossing intention of VRU's, the two main approaches for predicting the crossing intention of VRU's are trajectory-based approaches and classification-based approaches. Trajectory-based approach [25,26,37,44] uses RNN's and deep neural networks to predict the future trajectory of the pedestrians and use these trajectories to

estimate the crossing intention. This approach treats the intention prediction task as a regression problem where the future trajectories are represented as a series of future bounding boxes for the pedestrian. In general regression tasks are more difficult than classification tasks which makes the prediction accuracy for regression models lower than their classification counterparts. Classification-based intention prediction [5, 16, 17, 22, 28, 39] uses visual cues extracted from the scene to predict whether the pedestrian is crossing or not. In our framework, we chose to use the classification-based approach for our prediction model as classification models showed better performance than trajectory prediction models.

1.3 Problem Statement

According to the World Health Organization's 2018 Global Status Report on Road Safety, it was highlighted that there is a continuous increase in the number of road traffic fatalities every year. It was reported that approximately 1.35 million deaths occur annually worldwide as a result of traffic accidents [43]. Deep learning techniques have shown promising results in predicting the crossing intention of pedestrians by extracting various cues from input frames captured by a dashcam, including pedestrian bounding boxes, pose key points, ego vehicle speed, local context, and semantic segmentation of the scene. However, several challenges must be addressed to achieve accurate predictions, including the hard real-time constraint and maintaining prediction accuracy during bad weather conditions using image enhancement techniques.

One of the key challenges in pedestrian crossing intention prediction is obtaining accurate and reliable inputs from the input frames. Extracting

pedestrian bounding boxes and pose key points requires overcoming occlusions, variations in lighting conditions, and background clutter. Additionally, effectively incorporating ego vehicle (i.e. the vehicle that contains the sensors that perceive the environment around the vehicle) speed, local context, and semantic segmentation into the prediction model requires a robust fusion of diverse information sources and contextual cues. Another significant challenge is the hard real-time constraint imposed on the prediction model to allow for action-taking. The model needs to achieve accurate predictions while satisfying the stringent time constraints to ensure the practical viability of the system.

Furthermore, adverse weather conditions pose a considerable challenge to maintaining prediction accuracy. Factors such as poor visibility, rain, snow, or fog can significantly degrade the quality of input frames, making it difficult for the prediction model to accurately extract relevant features and make reliable predictions. Image noise, weather-related artifacts, and low contrast further complicate the task, impacting the accuracy and robustness of the model.

State-of-the-art approaches, as seen in [17, 39], employ RNNs and CNNs to process visual cues extracted from input frames, enabling the prediction of a target pedestrian's crossing intention. However, the use of RNNs for handling temporal features presents several challenges. Firstly, RNNs struggle with long-term dependencies in input sequences, as discussed in [4]. Moreover, training RNNs requires substantial time, and their inference is comparatively slower than self-attention-based models. Additionally, RNNs and CNNs encounter domain adaptation issues due to their sequential and local nature, which introduces bias. In contrast, global self-attention models process the entire sequence in one step. Furthermore, current state-of-the-art models do

not address challenges arising from adverse weather conditions. Overcoming these challenges involves employing image enhancement techniques, which play a crucial role in maintaining accurate predictions during bad weather. These techniques aim to improve input frame quality and clarity, reducing noise, enhancing contrast, and preserving important features related to pedestrian detection and localization. Lastly, our study stands out by considering real-time testing of proposed models, an aspect overlooked in previous literature.

Therefore, the primary problem addressed in this study is fast and accurate pedestrian crossing intention prediction using deep learning techniques. The study aims to tackle the challenges of providing an accurate prediction model that can generalize to different testing scenarios, maintain its prediction accuracy during severe weather conditions and meet the hard real-time constraint by having a low inference time and low end-to-end prediction time taking into consideration the processing time needed to extract different inputs needed for the model to make a prediction.

1.4 Thesis Contribution

This thesis makes the following main contributions:

1. We propose a novel pedestrian prediction framework based on the transformer architecture. The framework utilizes both vision transformers and self-attention mechanisms to produce an accurate prediction with low latency achieving state-of-the-art performance on JAAD dataset [27] against several models while maintaining a low inference time.

2. Provide a fully annotated dataset to train a multilabel classifier to detect root problems in input frames. The dataset contains 75 dashcam videos

collected from over 4 hours of footage. The dataset contains footage with both severe weather and clear conditions in a balanced manner to remove bias during training.

3. Provide a solution for maintaining prediction accuracy during unfavorable weather conditions by employing an image enhancement pipeline to detect and rectify underlying issues in input frames.

4. Provide a novel feature fusion method (total fusion) in the prediction model that provides better performance in terms of accuracy and inference time.

1.5 Thesis Organization

The organization of this thesis is as follows: In Chapter One, we describe the problem of prediction of VRU intentions and the utilization of image enhancement techniques to maintain prediction accuracy. We discuss the existing gaps and challenges in computer vision-based approaches, outline our approach to addressing these challenges, and highlight our contributions to the field.

Chapter Two is dedicated to the examination of various image enhancement methods, wherein we assess the advantages and disadvantages of employing deep learning or classical image processing for this particular task. We take into consideration the stringent timing constraint that arises from the inherent nature of the problem we aim to address. Furthermore, we provide an extensive background study on different intention prediction models and the techniques employed in each model, emphasizing the pros and cons of each technique.

In Chapter Three, we introduce our framework and its components. We

delve into a detailed discussion of the techniques and tools employed in constructing our framework, the main components of our framework which are the image enhancement pipeline and the intention prediction model are described thoroughly, and the reason behind choosing their inner components including the transformer encoder, the vision transformer, the image enhancement modules and the image multilabel classifier is discussed in detail.

Chapter Four presents the comprehensive results obtained from testing the entire framework on the JAAD dataset, as well as the individual components, and thoroughly examines and interprets the results. We also provide the results of testing the framework on a deployment environment against state-of-the-art models. Our model shows improvement over state-of-the-art models in terms of accuracy achieving 82% on JAAD all datasets and 20 ms for the inference time.

Lastly, in Chapter Five, we conclude by providing valuable insights into our methodology, discussing its effectiveness, potential limitations, and suggesting potential directions for future research and expansion of our work.

Chapter 2

Background and Related Work

2.1 Introduction

In this chapter, we review the literature surrounding the prediction of pedestrian crossing intention and different image enhancement techniques used to overcome various defects in video sequences. We start by investigating different enhancement techniques that use classical image processing and then compare these methods with their deep-learning counterparts. After that, an extensive review of state-of-the-art deep learning methods utilized for pedestrian crossing intention prediction is provided and research gaps are outlined. Lastly, we give an introduction to transformers, self-attention mechanisms, and vision transformers which are the main building blocks for our prediction model.

2.2 Image Enhancement For unfavorable Weather Conditions

In recent years, computer vision techniques have received enormous attention in traffic domain applications. The prime beneficiary of this technology is autonomous vehicles, although improvements are being made every day to achieve a fully self-driving vehicle, there are still many challenges related to object detection in unfavorable conditions and real-time performance. Low-visibility weather conditions, such as rain, fog, and snow, are some of these challenges. The stringent requirements for real-time performance in autonomous vehicles also pose a significant challenge for any perception model that could be implemented in these vehicles.

In 2018, an unprecedented accident happened in the United States during

an Uber self-driving vehicle test. Although the driver was distracted by his mobile phone, the car detected a pedestrian holding a bicycle and crossing the road at a very late stage resulting in the pedestrian's death. Figure 2.1 shows the first frame in which a pedestrian is detected from the Uber accident recording using the YOLOv4 detector. According to the reports, it is clear that it was a system design error, due to multiple reasons: (1) The system was not trained on detecting a human holding a bicycle, (2) It was nighttime and visibility was poor, (3) The system did not take any actions even though the object was detected six seconds before the collision [32].

Image enhancement techniques can be divided into two main categories: (1) classical methods such as traditional filters, classical image processing techniques, and classical machine learning techniques, (2) deep learning-based models. Histogram equalization is one of the most popular classical methods used for low-resolution images [11]. It enhances the brightness and color contrast in an image. This method was tested on a dataset with foggy images and showed a significant improvement in object detection. However, as mentioned earlier, in real-life situations, an image could contain more than one defect. Using only one classical image enhancement method would not fix all the issues in an image. In fact, it is even worse to pass an image through the wrong filter which is commonly known as a destructive or catastrophic enhancement. CNN-based methods can achieve better outputs and have the potential to fix multiple defects in a given frame with the proper training dataset. However, most CNN-based architectures that are used in image enhancement require a long inference time to achieve good results.

The recent advancements of deep learning and neural networks enabled researchers to develop many new methods for image enhancement. These

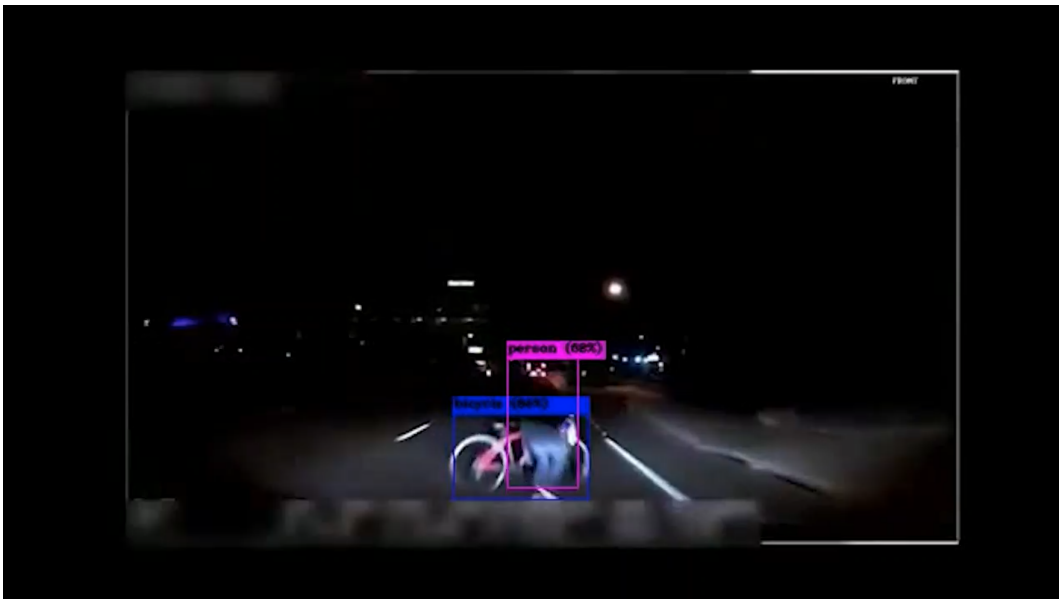

Figure 2.1: Uber accident recording passed to YOLOv4

methods improve the performance of image processing in many applications including the self-driving vehicles. However, time constraints and image quality (especially if an image may suffer from multiple quality issues simultaneously) are the two major challenges that need to be addressed to meet the real-time requirement of autonomous vehicles.

In this Chapter, we highlight the work done for image enhancement using both classical and deep-learning approaches while discussing the limitations of each technique. Then we provide a simple study that shows the effect of adverse weather conditions on the accuracy of the detection modules.

2.2.1 Classical image enhancement algorithms

Classical methods rely mainly on performing mathematical operations locally on the image pixels to fix a particular issue in the image, which could be low light, poor image contrast, or any other issues that could affect image quality.

Mathematical operations carried out in the frequency domain could involve transforming the image using Fourier transform, process the image by applying filters for example, then re-generate the image again. Although this process is fast and does not require any training datasets, it is not scalable and can not be used to fix more than one issue at a time. In our framework, we tackle three concurrent types of image defects that could happen simultaneously in challenging weather [24].

- Low light, which is captured mainly during the night in streets without proper lighting or on extremely foggy days
- Low-resolution, which mainly occurs when the images are captured with an unclear lens (e.g., fog, snow falling, or heavy rain falling)
- Blurriness, which could be caused by rain, fog, or a combination of these weather conditions

There are many classical image enhancement techniques that were developed to overcome each one of the above defects, but not all of them at once. Therefore the first step in the proposed pipeline is to detect what issues are present in the video frames passed to use the proper sequence of classical methods to enhance the images.

Low light was the main cause for the Uber accident that killed a pedestrian and it remains a significant challenge for AVs until today. It gets even more challenging when an image is taken during a foggy or rainy night, which further escalates the effects of low light on the image quality. Although deep learning algorithms have achieved great results in this area like MIRnet model [41], LLNet [19], the main downside for these models is the high computational time needed which makes it difficult to run in real-time applications. Thus, simple

(a) Original image before enhancement

(b) Enhanced image using Sepia filter

Figure 2.2: Enhancing low light conditions using Sepia kernel

filters like sepia kernels can be used to enhance the brightness and improve the image quality with a very fast response time as shown in Figure 2.2 (b). However, the downside of using these kernels is the limited performance they

can offer which may not be acceptable in severe dark conditions.

In regards to the low-resolution problem, the histogram equalization method has been used by many researchers, where the image is first transformed to gray-scale [13], as the technique relies on stretching the gray levels in an image and enhancing the contrast. Therefore, this technique is mainly used in applications that do not require colorful bright images [12]. A workaround for this limitation is to first change the RGB image representation into a YCbCr representation where the Y matrix carries the luminance of the image and the Cb and Cr matrices carry the chrominance of the image. The histogram equalization is then carried on the luminance part of the image (Y) and then the image is transformed back to RGB representation. Multiple variations of this method have been proposed to overcome some challenges in the base model [13]: local histogram equalization, local histogram stretching, and non-linear mapping methods. Generally, all histogram equalization variations aim to compute a function for each pixel in relation to the neighboring pixels. Figure 2.3 shows the result of applying the AHE method on a low-resolution frame and then passing the frame to YOLOv4 for object detection.

The Adaptive Histogram Equalization (AHE) [24] was proposed for images that have darker parts in certain pixels than others. AHE's approach is to locally identify the gray-scale mapping of a pixel on an image according to its neighboring pixels to be able to effectively enhance the image's important details [11]. In this paper, the AHE method is used mainly on low-resolution images as it showed an improvement in regards to the detection frame of a pedestrian using an object detection module in foggy situations.

Lastly, the third image defect considered is a blur which is caused by various weather conditions like rain, snow, fog, or any combination of these. There

Figure 2.3: Using AHE method and YOLOv4

are multiple types of blur: Average blur, Gaussian blur, and Motion blur The blurry or defective image (d) with a blur could be described using the following equation [1].

$$d = b * I + x \tag{2.1}$$

where I represents the original image, b is the blur factor, x is the noise and $*$ is the convolution process [1].

Image de-blurring techniques are essential for many fields, such as enhancing traffic images captured in rainy weather conditions. The research done by Palumbo et.al [23] shows how various techniques like Bias Field Correction, Intensity Standardization, and Noise filtering are used to de-blurring images. It is also important in military applications, traffic surveillance, and photography; therefore intensive research has been already made in this area and various techniques have been developed. Popular methods for image de-blurring include Iterative Van Cittert Algorithm, Iterative Landweber Algorithm, It-

erative Richardson-Lucy Algorithm, Iterative Poisson Map Algorithm, and Laplacian Sharpening Filters [1].

2.2.2 Deep-learning image enhancement algorithms

In the last decade, deep learning models have made great progress with their ability to learn generalized features from big datasets. The main two categories that exist are: (1) Convolution Neural Networks and encoder-decoder models, (2) high-resolution feature processing, both of which have some advantages and disadvantages. For example, the first approach faces a challenge in the reconstruction process of the original image resolution. Although a generalized number of spatial features is learned by this resolution reduction. The second approach tackles this issue by avoiding any down-sampling on the input image. However, the trade-off comes at the cost of encoding the image details [41]. MIRnet is one of the popular models used for image enhancement mainly to solve poor lighting conditions. MIRnet is introduced in [41] as a new multi-scale approach, which maintains the original image resolution and features by using parallel convolution streams. Similar to MIRnet, there are several modes that follow the same approach such as Scale-recurrent network (SRN-DeblurNet), Single guided network (SGN), and Deep multi-scale convolutional neural network [21]. However, MIRnet differs in the ability to process data across all the network levels, unlike the above models.

In regard to the low-resolution problem, many deep-learning image resolution enhancement algorithms have been developed by researchers. These algorithms are categorized as follows: Prediction models, Edge-Based methods, Image statistical methods, and Patch-based models. Chih-Yuan Yang et.al [38] developed a benchmark for all these methods and showed that the

Patch-based models achieve state-of-art performance. Generally, the super-resolution convolutional neural network (SRCNN) performs image enhancement in three steps: (1) Patch extraction and representation, (2) Non-linear mapping, and (3) Image reconstruction. SRCNN first extracts features and patches from the defect image to generate a representation in PCA, Haar, etc. Then, a feature map is created by mapping high-dimensional vectors during the second step of non-linear mapping. Lastly, the reconstruction of the image process starts from the high-resolution feature maps created in the previous step [14].

For the blurring defect in images, Syed Waqas Zamir et.al [40] propose a restoration transformer (Restormer) model for image deblurring. Restormer is a deep learning model for image restoration based on transformers in its architecture. This reliance on transformers instead of normal neural networks aims to enhance the overall performance of the model in terms of inference time and accuracy.

However, transformer architecture normally has a significant computational complexity especially with high-resolution images, making it hard to use with high-resolution image restoration tasks [40]. But the developers of the Restormer model have made changes in the normal transformer architecture, making it the state-of-the-art model in:

- Single image motion deblurring.
- Defocus deblurring.
- Restoring images that are blurred because of raindrops.
- Image denoising.

The main drawback of the Restormer model is the long inference time. It takes about 40 seconds to perform image de-blurring on a single image.

From the previous sections, we conclude that classical image enhancement algorithms have a very fast response time but lack in terms of high performance and generalization to different image conditions. On the other hand, deep-learning approaches have high performance in terms of image enhancement but usually require a long time to produce results which makes them more suitable for offline applications. In our framework, we chose to use classical image enhancement modules in our pipeline as they are easier to implement and can be replaced later by more sophisticated techniques depending on the application. For instance, for applications that may apply the pipeline offline (i.e., the image quality is more important than the inference time), then replacing the classical image enhancement modules with deep learning image enhancement modules will be the ideal solution.

2.3 Deep Learning In Pedestrian Crossing Intention

The recent progress achieved in the field of computer vision has had a significant impact on various applications, particularly autonomous vehicles (AVs). The effectiveness of AVs' perception models is heavily reliant on the processing of a large volume of input frames at high speeds to acquire the necessary information for crucial decision-making. While the performance of perception models has seen a significant improvement in numerous tasks, there remain obstacles that impede the widespread adoption of AVs in the market. One of the key challenges is predicting the behavior of vulnerable road users (VRUs)

such as pedestrians and cyclists. The accurate and expeditious anticipation of VRUs' movements would enhance the safety and reliability of driving experiences.

The recent advancements in visual sensor resolution, combined with their affordable price, have established them as the primary choice for autonomous vehicles (AVs). Among the various sensors commonly used in AVs, front-facing cameras are not only reliable but also cost-effective, taking this into consideration developing a behavior prediction model that relies on visual inputs from front-facing cameras will be the most practical option for deployment. Kotseruba et al. [17] provided the most recent benchmark for pedestrian intention prediction in which the PCPA model achieved state-of-the-art performance. After that, different approaches and architectures were proposed to enhance the performance of PCPA [29, 39, 42]. The provided work either used non-visual features only for intention prediction [29, 42] or a combination of visual and non-visual features [17, 39]. However, apart from Zhang et al. [42] who used graph neural networks all of these models used CNN and RNN for visual and non-visual feature extraction respectively which may not be optimal for this task. Also, the inference time of the model is not considered in the previous work. In this section, we provide an extensive study of various deep-learning approaches for pedestrian crossing intention prediction. Table 2.1 provides a summary of the selected DL models with respect to criteria and input features that are used for predicting the crossing/non-crossing intentions of pedestrians.

Table 2.1: Summary of the selected DL models

DL Models	Selection Criteria	Input Type
Convolution-3D	Spatio-Temporal-based	Local context
Static-VGG16	Basic VGG16 Architecture	Local context
SFRNN	Scene and Dynamics Features Fusion	Local context, global surround, pose, bounding box,vehicle speed
SingleRNN	Trajectory-based	Local context, pose, bounding box, vehicle speed
Convolutional LSTM	LSTM-based	Local context
PCPA	Attention-based	local box, bounding box, pose
Two-streams RNNs	Trajectory-Based	Local context, global surround, pose, bounding box, vehicle speed
Position velocity LSTM	Trajectory-based	Bounding box, pedestrian velocity vector
Mask PCPA 4 2D	Attention-based with Features Fusion	Local context, global surround, pose, bounding box, vehicle speed

2.3.1 Static-VGG16

Static-VGG16 [31] is one of the simplest deep learning models. The Static model architecture is formed of a VGG16 backend and fully connected layers, where action estimation and prediction are based on the final image in the observation sequence. Even though the static-VGG16 model is considered the simplest among other baseline models, it still outperformed other complicated architectures such as conv-LSTM.

2.3.2 SingleRNN

Kotseruba et al. [16] proposed a single RNN network. The network receives a vector as input, which is a combination of various features including bounding box coordinates, pedestrian pose key points, ego vehicle speed, and the local context of the scene surrounding the target pedestrian. The input vector is then passed through a RNN to capture temporal features, followed by the transfer of these features to a FCL for the final prediction. Nonetheless, the concatenation of all the extracted features into a single vector may potentially cause confusion for the model. Additionally, the use of only one RNN to capture all the temporal features may not be the most optimal approach.

2.3.3 SFRNN

Joe.NG et al. proposed an alternative approach in [22], which involved using a stack of RNNs instead of a single RNN. In this approach, each RNN layer takes the hidden states from the previous RNN layer as input. Furthermore, the authors suggested the use of features extracted from optical flow instead of raw frames.

In [28], Amir Rasouli et al. presented a modified version of the stacked RNN network. In their approach, the features extracted from one input are fed to a GRU cell. The output of the GRU cell is then concatenated with the next features, resulting in a more refined representation of the temporal characteristics. The authors also proposed the incorporation of the surrounding context as one of the inputs. The surrounding context is defined as the features extracted from the input frame after masking the bounding box of the target pedestrian. These extracted features play a vital role in teaching the

model to distinguish the target pedestrian from its surrounding environment.

2.3.4 Two-streams RNNs

A. Bhattacharyya et al. [5] utilized a MultiRNN network that comprises of two streams. The first stream is utilized to forecast the odometry of the ego vehicle, which involves predicting its speed and steering angle. On the other hand, the second stream utilizes the predicted odometry alongside the pedestrian bounding box coordinates to predict the trajectory of the pedestrian. This trajectory prediction involves determining the coordinates of the next bounding box. However, a limitation of this method is that the trajectory prediction is reliant on the accuracy of the odometry prediction, which may result in error propagation in the event of inaccurate odometry prediction.

2.3.5 Position velocity LSTM

Position velocity LSTM is a multi-task learning model first proposed by Smail Ait Bouhsain et al. [7]. The model is based on concatenating data from previous pedestrian bounding boxes and velocity vectors to predict the pedestrian intention as well as the pedestrian's future trajectory. Bounding boxes are extracted from the true labels provided by the dataset for a specific number of frames. The velocity vectors are calculated by subtracting consecutive bounding boxes that belong to the same pedestrian. The model architecture is described as an encoder-decoder model where there are two encoders one for the position and the other for the velocity and then two decoders for future trajectory and intentions respectively. Every encoder or decoder consists of a LSTM block with 256 hidden units. The model takes **N** number of previous

bounding boxes and velocity vectors and predicts the intentions and future trajectories for the next **M** frames where **N** and **M** are arbitrary values that can be set manually before training.

2.3.6 Convolutional LSTM

Convolutional LSTM (ConvLSTM) [30] is an extended version of a fully connected LSTM deep learning model to address the spatiotemporal sequence prediction problem. The introduction of convolution structures into the encoding-prediction design (input-to-state and state-to-state transitions) provides an end-to-end trainable solution. The output prediction depends on feeding the extracted attributes to the convolutional LSTM which is considered an input into a fully connected layer by the final hidden state.

2.3.7 Convolution-3D

Another approach for detecting pedestrian intentions involves using 3D convolutional networks (3D-CNN) [9, 33]. Du Tran et al. [33] proposed the usage of 3D-CNN for action classification tasks. The 3D-CNN is capable of extracting both spatial and temporal features from the input frame sequence using 3D convolutions and 3D pooling layers. However, this approach is limited to extracting only visual features. To use 3D-CNN for pedestrian intention prediction, the input frame sequence must be cropped around the target pedestrian. Otherwise, the same frame sequence will be fed to the network multiple times if multiple pedestrians exist in the same scene. Additionally, neglecting pose key points and global context as features can significantly affect the accuracy of the model predictions.

Joao Carreira and Andrew Zisserman [9] proposed a two-stream 3D-CNN architecture that utilizes both the raw frame sequence and the cropping optical flow frames. Each input is separately fed into a 3D-CNN for feature extraction, and the resulting features are concatenated and used for the final prediction. While the inclusion of optical flow frames improves the accuracy, it also increases the processing time considerably.

2.3.8 PCPA

Kotseruba et al. [17] presented a recent benchmark for evaluating various intention prediction models and introduced an attention-based approach that employs 3D-CNNs and RNNs to extract spatial and temporal features from a sequence of input frames. The proposed model takes four inputs, namely the bounding box and pose key points of the target pedestrian, the speed of the ego vehicle, and the local context of the scene. The model achieved state-of-the-art performance on both JAAD and PIE datasets. However, this model didn't take the global context of the scene into account, moreover, the model solely relied on a single approach for fusing the extracted features, known as later fusion. Later fusion involves independently extracting features from the inputs using convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Subsequently, these features are concatenated to form the final prediction.

2.3.9 Mask PCPA 4 2D

Yang et al. [39] filled the discussed gaps in the PCPA model [17] when they proposed to use the semantic segmentation of the scene as one of the inputs

and investigated several fusion methods where the hierarchical method proved to be the most efficient.

2.3.10 Graph based models

Other methods have relied on using graph neural networks (GNNs) [18, 42]. These approaches use either pose key points [42] or bounding box [18] to represent the pedestrian node in the graph. However, these methods may not effectively capture visual inputs such as local and global context, which can limit the performance of the models.

2.3.11 A comparative study on the discussed models

Lastly, the mentioned models were compared using the JAAD dataset, and the results are shown in Table 2.2, results show that the attention-based models [17, 39] and 3D convolutional models [9, 33] perform better than the other approaches. Despite the progress made in the aforementioned areas, there remain several gaps that need to be addressed. Firstly, the inference time of the model requires further attention to enhance its efficiency. Secondly, the model's performance in adverse weather conditions needs to be improved, as these conditions can pose significant challenges to accurate prediction. Lastly, the exploration and integration of newly proposed self-attention mechanisms and transformer architectures offer potential avenues for advancing the prediction capabilities and overall performance of the model.

Table 2.2: Evaluation and Performance Metrics on JAAD Dataset

DL Models	Accuracy	AUC	F1 Score	Precision	Recall
Mask-PCPA-4-2D	**84.49**	70.65	53.11	**58.54**	48.59
C3D	84.03	77.30	**59.45**	53.46	66.95
Static_VGG16	82.81	73.95	55.10	50.71	60.32
SFRNN	74.95	76.68	52.56	39.29	79.35
SingleRNN	76.64	76.30	53.15	40.93	75.78
Conv-LSTM	19.79	50.96	30.12	17.77	**98.90**
PCPA	78.43	**80.16**	57.31	43.82	82.83
Two-stream RNNs	75.07	77.70	53.41	39.67	81.73

2.4 Transformer And Attention Based Algorithms

Self-attention-based mechanisms and transformers have emerged as pivotal components in various natural language processing (NLP) tasks, revolutionizing the field's approach to capturing and modeling contextual dependencies. Self-attention, also known as intra-attention, enables an input sequence to attend to its own elements, facilitating the identification of important relationships and capturing long-range dependencies. Transformers, a type of neural network architecture built upon self-attention, have become particularly influential in NLP due to their ability to effectively process sequential data. By employing self-attention mechanisms, transformers can simultaneously model dependencies between all positions in a sequence, making them highly suitable for tasks such as machine translation, sentiment analysis, and question-answering. The self-attention mechanism in transformers enables them to efficiently encode and aggregate information from the entire input, resulting in state-of-the-art performance in various NLP benchmarks. As a result, the

integration of self-attention-based mechanisms and transformers has significantly advanced the field of NLP, providing a robust foundation for capturing complex contextual relationships in textual data. On the other hand, Vision transformers have recently emerged as a groundbreaking approach for image recognition and understanding tasks, challenging the traditional dominance of convolutional neural networks (CNNs) in computer vision. Inspired by the success of transformers in natural language processing, vision transformers extend the application of self-attention mechanisms to visual data. By representing images as sequences of patches, vision transformers enable the modeling of global contextual relationships among image elements. This approach eliminates the need for handcrafted hierarchical features and enables end-to-end learning. Vision transformers have demonstrated impressive performance across a range of computer vision benchmarks, including image classification, object detection, and semantic segmentation. The self-attention mechanism in vision transformers allows them to capture long-range dependencies and effectively model the interactions between image patches, enabling the extraction of fine-grained features and contextual information. Through their ability to learn from both local and global image information, vision transformers have established themselves as a powerful paradigm for visual recognition tasks, pushing the boundaries of computer vision research. In this section, we give an introduction to attention mechanism, self-attention, transformers, and vision transformers which are the main building blocks of our prediction model.

2.4.1 Transformer and self attention mechanism

Attention mechanism

The introduction of the attention mechanism aimed to enhance the efficacy of the encoder-decoder model used in machine translation. Its fundamental concept was to enable the decoder to flexibly utilize the most pertinent segments of the input sequence. This was achieved through a weighted combination of all the encoded input vectors, wherein the highest weights were assigned to the most relevant vectors. The attention mechanism was first introduced to guide recurrent neural networks in order to produce hidden states that describe the relationship between different parts of the input sequence in a better way. The first attention mechanism was introduced by Bahdanau et al [2]. This attention mechanism relied on the step-by-step computation of alignment scores, attention weights, and context vectors in the following manner:

- **Alignment scores:** The alignment scores $e_{t,i}$ are computed by using a linear transformation function $a(.)$ (i.e a feed-forward layer) on the encoded hidden state h_j and the decoder hidden state of the previous time step S_{t-1}. Equations 2.2,2.3 show the calculation of the alignment scores where V_a, W_a, U_a are the weight matrices of the feed-forward layer.

$$e_{t,i} = a(S_{t-1}, h_j) \tag{2.2}$$

$$a(S_{t-1}, h_j) = v_a^T tanh(W_a.S_{t-1} + U_a.h_j) \tag{2.3}$$

- **Attention weights:** The scores computed in the previous step signify the relevance of the encoder's hidden state to the decoder's hidden state, this score is normalized to ensure that all the attention weights are be-

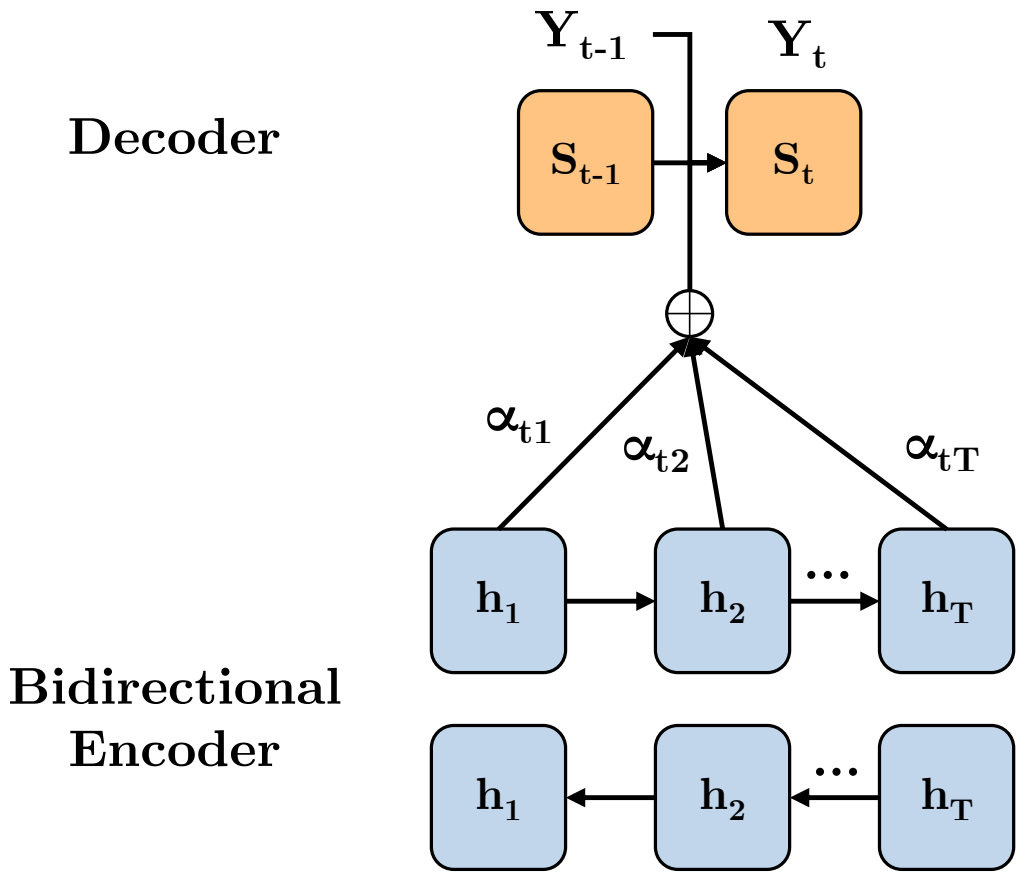

Figure 2.4: Illustration of Bahdanau et al attention mechanism in encoder-decoder network

tween 0 and 1 which is done using a softmax layer. The output from the softmax is the attention weights $\alpha_{t,i}$

$$\alpha_{t,i} = softmax(e_{t,i}) \tag{2.4}$$

- **Context vector:** The encoder hidden states h_i are weighted using the attention weights $\alpha_{t,i}$ and then summed as seen in equation 2.6 to form the context vector which is used in guiding the decoder to produce the

hidden state of the next time step S_t as shown in figure 2.4.

$$c_t = \sum_{i=1}^{T} \alpha_{t,i}.h_i \tag{2.5}$$

In general, the attention mechanism uses three components: the query **Q**, key **K**, and value **V**. In Bahdanau et al [2] the query is the decoder hidden state S_{t-1}, key, and value are the encoder hidden state h_j.

Transformers and self attention

Self-attention as a concept relies on relating different positions of a single sequence in order to compute a representation of the sequence [34], so the key difference between general attention and self-attention is that self-attention is used on a single sequence to find the relation between different parts of this sequence, where general attention estimates the relevance of one sequence to another one. Transformers [34] introduced by Ashish Vaswani et al were the first architecture that relied solely on self-attention and completely discarded RNNs from the encoder-decoder architecture. The transformer model uses self-attention in both encoder and decoder networks and uses general attention when connecting the encoder and decoder parts as shown in figure 2.5.

The authors in [34] utilized the scaled dot product attention over the addition attention or feed-forward attention used in [2] the output from the attention block is then given from equation 2.6

$$Attention(Q, K, V) = Softmax(\frac{Q.K^T}{\sqrt{d_k}}).V \tag{2.6}$$

where d_k is the dimension of the key or query vector. Instead of using one at-

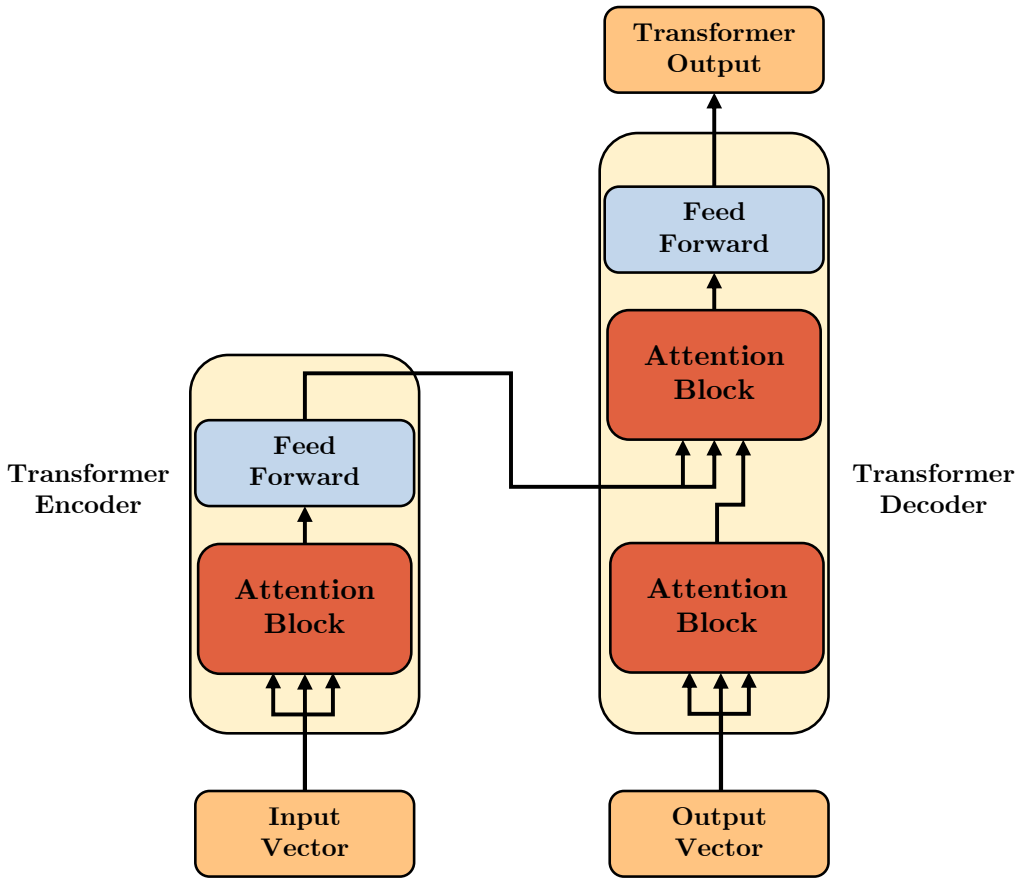

Figure 2.5: illustration for the transformer architecture, self-attention is used in the encoder and decoder whereas general attention is used in connecting the two components

tention head in the encoder or decoder the concept of multi-head attention is also used in [34] where each head learns information about the input sequence from different representation subspaces, this is done by applying different linear transformations on the Q, K, V matrices then apply self-attention to the output matrices and then apply another transformation on the concatenated outputs form the attention block. Equations 2.7,2.8 show the mechanism of multi-head attention for (h) heads.

$$Multihead(Q, K, V) = Concat(head_1, head_2, ..., head_h).W^O \tag{2.7}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.8}$$

The use of multi-head attention enables the model to attend to multiple parts of the input concurrently, further improving its capability to manage long-term dependencies [39]. In other words, the prior attention functions boost the memorization of sequential patterns by focusing on specific parts of the input characteristics. Recently, transformer models based on self-attention cells have become increasingly popular and have replaced RNNs in various applications. The study in [20] demonstrated that transformers outperformed RNN-based models in extracting temporal features in pedestrian trajectory prediction.

2.4.2 Vision transformers

Transformers have been applied in computer vision applications, where they have replaced state-of-the-art architectures that utilize CNNs [15]. In these applications, the image is regarded as a word sequence in NLP applications by dividing it into linear embeddings of multiple patches and providing them to the transformer. It is worth mentioning that vision transformers (ViT) have a large number of parameters, giving the model significant flexibility to fit the training data. Consequently, ViT can surpass CNNs only when trained on large and diverse datasets, ranging from 4 to 300 million images, to avoid over-fitting [15]. These large datasets permit transformers that lack the essential inductive bias of CNNs to exhibit excellent performance and generalization abilities when applied to smaller tasks. The study in [15] demonstrates that ViT achieves the best accuracies on datasets such as ImageNet and ImageNet-Real when pre-trained on larger datasets such as ImageNet-21k.

2.5 Summary

In this chapter, the focus is on the literature surrounding pedestrian crossing intention prediction and image enhancement techniques in video sequences. The chapter begins by exploring various enhancement techniques that rely on classical image processing methods. These techniques aim to address different defects present in video sequences, improving the overall quality and clarity of the image. The chapter then delves into a comparison between these classical image-processing methods and their deep-learning counterparts. Deep learning has gained significant attention in recent years due to its ability to automatically learn complex patterns and features from data. The comparison helps to understand the advantages and limitations of each approach in the context of pedestrian crossing intention prediction. Moving forward, the chapter provides an extensive review of state-of-the-art deep learning methods utilized specifically for pedestrian crossing intention prediction. This section highlights the latest advancements in the field and discusses the effectiveness of various deep learning architectures, algorithms, and models in accurately predicting pedestrian crossing intentions.

Furthermore, the chapter identifies research gaps in the existing literature, pointing out areas where further investigation and improvements are needed. This analysis helps to identify opportunities for future research and development in pedestrian crossing intention prediction. Lastly, the chapter introduces transformers, self-attention mechanisms, and vision transformers as the primary building blocks for the proposed prediction model. Transformers have emerged as powerful tools in various domains, including computer vision, due to their ability to capture long-range dependencies and contextual informa-

tion. The chapter provides an overview of these concepts and establishes their relevance to the prediction model.

Chapter 3

Methodology

3.1 Introduction

In this chapter, a thorough examination of the methodology pertaining to the proposed framework is presented. Commencing with an intricate explication of the problem definition, a comprehensive understanding of its nuances and fundamental elements is established. Subsequently, meticulous attention is paid to the intricate design of the image enhancement pipeline, where the overarching architecture is illustrated in great detail. Furthermore, a comprehensive overview of each individual component within the pipeline is provided, accompanied by a comprehensive discussion of the dataset used for both training and testing purposes.

Subsequent to the elucidation of the image enhancement pipeline, the focus is shifted toward an exhaustive exploration of the pedestrian intention prediction model. Its intricate workings and underlying principles are thoroughly examined, ensuring a comprehensive understanding of its capabilities and limitations.

3.2 Problem Formulation

Due to the visual nature of the problem we are trying to solve, we mostly rely on visual sensors to extract our input frames. Taking this into consideration we find that front dashboard cameras are the most practical option due to their informative output about the scenes combined with their affordable price. For these reasons, front dashboard cameras were chosen to be the main source of input in our framework. Other visual sensors such as lidar can be used to provide further information about the scene such as the depth of an object but they are extremely expensive which hinders them from being

adopted for mass production. Our problem is defined as follows: Given a sequential series of frames captured from the front dashboard camera of the ego vehicle, our objective is to employ an image enhancement pipeline that classifies these frames as either clean or defective. In the case of a frame being classified as defective, the pipeline should further identify the specific underlying issue responsible for the defect and implement appropriate corrective measures accordingly. Consequently, the pipeline produces clean or enhanced frames, which are subsequently utilized for the purpose of pedestrian intention detection while they traverse the roadway. The prediction model employed estimates the probability that a particular pedestrian, denoted as i within the observed scene will either cross the road or not cross it, within n frames prior to the occurrence of the crossing or non-crossing event (C/NC). It is worth noting that n represents the number of frames elapsed from the last observed frame up until the C/NC event [39]. Where the inputs are divided into non-visual inputs and visual inputs.

In addition to the visual inputs, non-visual cues are incorporated into our model to enhance its predictive capabilities. These non-visual inputs encompass the vehicle speed (V), the bounding box surrounding the target pedestrian (Bi), and the pose key points for the target pedestrian (Pi). On the visual aspect, we take into consideration both the local context (LC) and the global context (GC). The local context refers to a cropped region surrounding the target pedestrian within the input frame, effectively capturing the immediate surroundings. In contrast, the global context involves image segmentation that encompasses the target pedestrian within the input frame, providing a broader contextual understanding. However, to better comprehend the significance of the global context in our model's input, we conduct a detailed analysis

in Chapter 4. In this chapter, we examine the implications of excluding the global context, and we present our findings and insights regarding its impact on the overall model performance. During both the training and testing phases of the pedestrian intention prediction model, the pedestrian bounding box and vehicle speed are extracted from the ground truth annotations available within the dataset. The local context (LC) is acquired by cropping the frame around the corresponding bounding box, thereby capturing the immediate surroundings of the pedestrian. To obtain the pose key points, a pre-trained openpose model [8] is utilized, enabling the extraction of accurate pose information. In addition, the global context (GC) is obtained through the utilization of a pre-trained DeepLabv3 model [10]. This model effectively performs image segmentation, encompassing the target pedestrian within the input frame. By doing so, a broader contextual understanding is achieved, which aids in the prediction process. However, when evaluating the end-to-end architecture, a different extraction method is employed for the bounding box and pose key points during testing. The objective of this approach is to reduce the processing time, optimizing the overall efficiency of the system. Subsection 4.4 presents a comprehensive and detailed illustration of the specific setup utilized for the end-to-end testing procedure. Following the input extraction phase, the model inputs for a single frame within a sequence of m frames can be represented by the following equations:

- The bounding box is given by the coordinates of the top left and the bottom right corners.

$$B_i = \{X_{ibr}, X_{itl}, Y_{ibr}, Y_{itl}\} \tag{3.1}$$

- The pose key points are given by a 36D vector of 2D coordinates that contain 18 pose joints.

$$P_i = \{(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), ..., (x_{i18}, y_{i18})\} \tag{3.2}$$

- The vehicle speed is given by a single categorical value that corresponds to one of five states (stop, moving slow, moving fast, decelerating, and accelerating).

$$V = v_1 \qquad \text{Such that} \quad v_1 \in [0 - 4] \tag{3.3}$$

- The local context is an RGB image taken around the pedestrian bounding box. in the conducted experiments the local context size is fixed to (224,224,3).

$$LC = lc_1 \qquad \text{Such that} \quad lc_1 \in \mathbb{R}^{224,224,3} \tag{3.4}$$

- The Global context is an RGB image produced by superimposing the segmentation mask of the frame on the raw frame. It has the same size of the local context image.

$$GC = gc_1 \qquad \text{Such that} \quad gc_1 \in \mathbb{R}^{224,224,3} \tag{3.5}$$

3.3 Image Enhancement

Before diving into the pipeline and its details, we made some pre-analysis experiments to justify our pipeline's objectiveness.

3.3.1 The effect of noise on the object detection

The purpose of this study is to confirm that image quality plays a vital role in detecting pedestrians using YOLOV4 [6]. We have chosen YOLOV4 as an object detector for several reasons

1. It can be trained through different GPU capacities which are super efficient for engineers to save time as their training is done much faster (it allows the model to be trained on 1080Ti or higher which means more CUDA resources to be used).

2. It allows the utilization of accuracy enhancement methods and hardware optimization methods by setting flags in its training commands.

3. Enhancing the training mechanism to be able to train the data on a single GPU and by making small changes to some of the batches' normalization techniques that can speed up the training process such as CBN (Cross Iteration Batch Normalization).

We have also picked seven of the JAAD video dataset [27] that represent different weather conditions (sunny, foggy, rainy) and at night time in order to simulate all possible combinations for the real-time scenario. We extract the frames from these videos at a rate of 30fps and pass them to YOLO to find the first frame containing a pedestrian. Then, we pass the generated frames

into a blurring filter with a kernel of size 20. This process is repeated until we complete all frames.

Table 3.1 shows the results of our pre-analysis which proves that image quality plays a vital role in detecting pedestrians in earlier frames. That means camera quality and camera lens' clarity contribute greatly to YOLO's detections and its performance and that motivated us to dive even further into our approach for creating a multi-staged image enhancement pipeline.

Table 3.1: Results from the analysis done on the JAAD dataset

Video Number	**Weather condtions**	**Description**	**Detection frame before blurring**	**Detection frame after blurring**
23	Rainy Night	Driving at low speed on a rainy night through a dark street with little light on it and a pedestrian crossing road quickly at the end of the footage	134	N/A
18	Sunny Morning	Driving on a sunny morning but get blocked by a crossing car for a while before going around the blocking car and continuing on the road	3	N/A
22	Snowy Morning	Driving in a plaza's parking lot on a snowy morning with some people walking on the parking lot to get to their cars	14	246
62	Sunny Morning	Turning left on a small cross-road with some people crossing the street	1	372
279	Sunny Morning	Driving through a neighborhood with some people crossing the road from the right of the car and the front of the car.	1	6
281	Snowy Morning	Driving on a snowy road slowly with a jay-walker slowly crossing the street	10	98
5	Sunny Morning	Driving in a plaza's parking lot at noon with people walking in front of the car with their groceries	1	11

3.3.2 The proposed pipeline

To maintain prediction accuracy during bad weather conditions we propose an image enhancement pipeline that detects and corrects the root problems in a

given frame. The pipeline consists of two phases: a detection phase and an enhancement phase.

During the detection phase, a multi-label classifier is employed to identify specific conditions that may impact the quality of the frame, including low light, low resolution, and blurriness. Furthermore, the classifier is capable of detecting frames that are already clean and require no further enhancement. In such cases, the pipeline can skip the subsequent enhancement phase. The architecture of the multi-label classifier is depicted in Figure 3.1, and it takes two inputs: the raw image, which is resized to dimensions of 448 $\times$ 448 $\times$ 3 pixels, and the image's histogram. The histogram represents the frequency distribution of pixel values across the three channels of the image and serves as a valuable indicator of unique image characteristics. For our classifier, we have utilized 100 bins for the histogram analysis. For example, Dark images typically exhibit a high frequency of low pixel values in the histogram, while low-resolution images often display clusters of pixels. On the other hand, blurred images tend to demonstrate a Gaussian distribution pattern in the histogram. Hence, there exists a direct correlation between the histogram characteristics and the underlying image problems.

To process the inputs and generate the final prediction, our model incorporates CNNs and FCLs. These components analyze and extract features from the inputs, and the resulting processed features are concatenated for the final prediction. the output of the model is 3×1 array of binary values. The structure of the output flags vector $\mathbf{V}$ is as follows, $\mathbf{V} = \boldsymbol{\{b,\ r,\ l\}}$, where b indicates blurriness (0 is not blurry, 1 is blurry), r represents low-resolution, l represents low light.

During the enhancement phase, we employ three distinct image enhance-

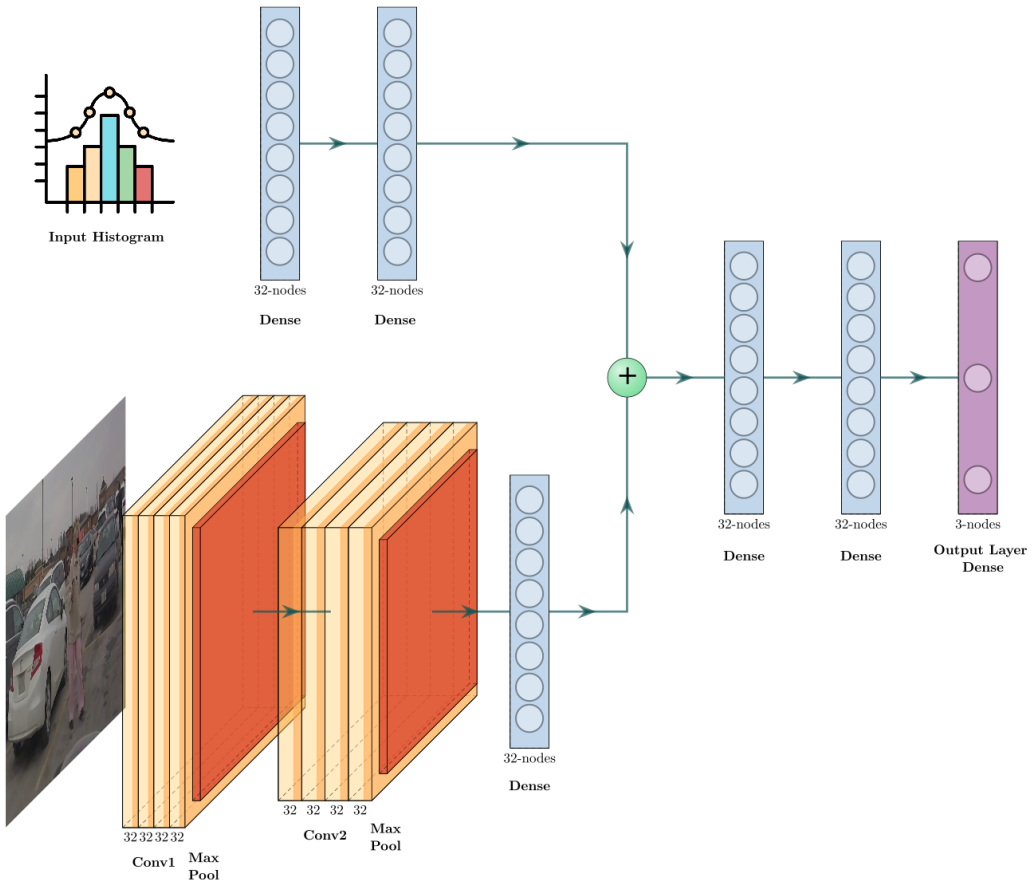

Figure 3.1: Multi-label classifier architecture: The architecture is based on CNNs and fully connected layers.

ment modules to address the three specific problems of interest. Each module focuses on tackling one problem independently. The following modules are utilized:

- De-blurring: We utilize a sharpening kernel with a kernel size of 3 to effectively address the issue of blurriness in the input frame. This module enhances the clarity and sharpness of the image, thereby reducing the blurriness.

- Low-resolution enhancement: To overcome the problem of low resolution, we apply the histogram equalization method. This technique enhances

the contrast and distribution of pixel values in the image, resulting in an improved perception of details and overall image quality.

- Low-light enhancement: The sepia kernel with a kernel size of 3 is employed to enhance images with low-light conditions. This module optimizes the color tone and brightness, leading to improved visibility and enhanced details in low-light scenarios.

Algorithm 3.1 Pipeline process

Input: Input frame F
Output: Enhanced frame EF

1: **procedure** PIPELINE(F)
2: $Image\_problem \leftarrow Multi - label(F)$
3: $X \leftarrow x$
4: $N \leftarrow n$
5: **if** $Image\_problem$ is Blurry **then**
6: $F \leftarrow Deblur(F)$
7: **end if**
8: **if** $Image\_problem$ is LowResolution **then**
9: $F \leftarrow EnhanceResolution(F)$
10: **end if**
11: **if** $Image\_problem$ is LowLight **then**
12: $F \leftarrow EnhanceLight(F)$
13: **end if**
14: $EF \leftarrow F$ (no enhancement)
15: **return** EF
16: **end procedure**

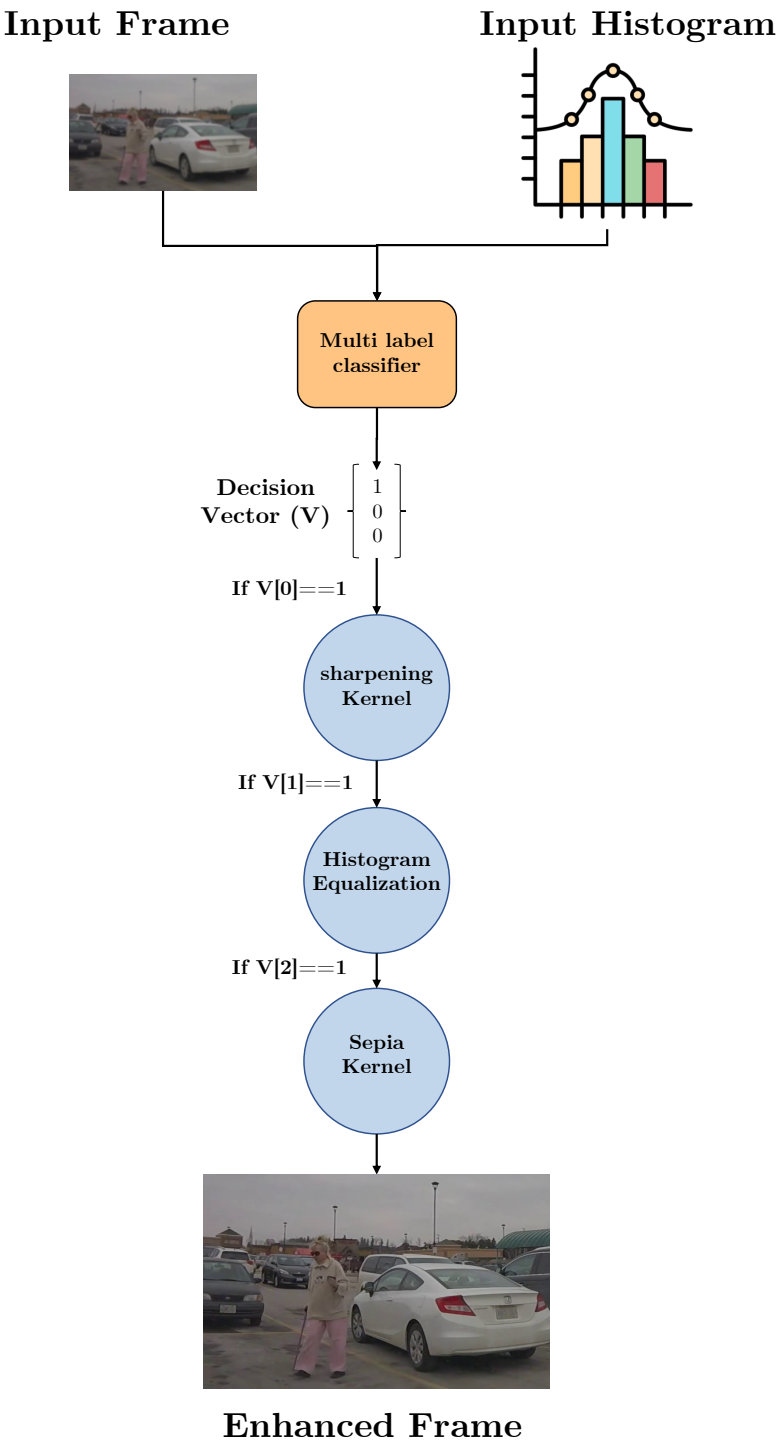

Figure 3.2: Pipeline procedure: The pipeline takes the raw image and the image histogram as inputs and leverages CNNs to detect root problems in input frames then uses classical image enhancement techniques to rectify these defects.

These enhancement modules have been selected to provide effective improvements to the input frames without significantly increasing the processing time of the overall framework. Figure 3.2 and algorithm 3.1 illustrate the overall pipeline procedure, showcasing the sequential execution of the enhancement phase. Moreover, the independence of these modules allows for easy replacement of any component in the pipeline, enabling flexibility and adaptability. The order in which the enhancement modules are applied plays a crucial role in the quality of the output frame. Experimentation with different permutations revealed that the order shown in Figure 3.2 provides the best results. The experiments were done on multiple images where a deformed image is introduced to different permutations of the enhancement components and the UIQI [35] is computed between the output image and the ground truth image. Table 3.2 shows the average UIQI score for each permutation over the test images.

Table 3.2: Results of testing the image enhancement pipeline.

Permutation	**Universal image quality index score**
Sepia-Sharpening-Histogram eq.	69%
Sepia-Histogram eq.-Sharpening	65%
Sharpening-Sepia-Histogram eq.	70%
Sharpening-Histogram eq.-Sepia	78%
Histogram eq.-Sharpening-Sepia	74%
Histogram eq.-Sepia-Sharpening	66%

3.3.3 Dataset

For the training of the image enhancement multi-label classifier, We collected and processed our own dataset due to the lack of a good dataset that includes pedestrians in unfavorable conditions. We named it Difficult Detection

Dataset (DDD)[1]. DDD contains 75 dashcam videos collected from over 4 hours of footage in different challenging conditions, such as low visibility, foggy, and rainy weather. Figure 3.3 displays different scenes extracted from our dataset, showcasing the diverse and challenging scenarios encountered during data collection. The primary objective of curating this dataset was to provide a difficult environment for the object detection process, which serves as the central focus in demonstrating the effectiveness of our image enhancement pipeline. Furthermore, the dataset incorporates clear footage to mitigate bias during training and ensure a balanced distribution of weather conditions commonly encountered during the winter season. The ratio between clear frames and

Figure 3.3: Sample frames from the DDD dataset showing different weather conditions included in the dataset.

unclear frames within the DDD dataset is approximately 23.3% to 76.7% respectively. All videos were recorded using a dashcam, capturing footage in 2K

[1]The dataset is available for download at `https://drive.google.com/drive/folders/1JRViqE5BpzIG2J4WzblsXAgToIWDrGCL?usp=sharing`

resolution at a frame rate of 30 frames per second, with a wide field of view of 170 degrees. The dataset encompasses a total of 32,250 frames, equivalent to nearly 18 minutes of video footage.

3.4 Intention Prediction

Our pedestrian intention prediction model leverages the transformer architecture [34] and specifically employs vision transformers [15] to process visual features. For the processing of non-visual features, the model utilizes the encoder component of the transformer, as illustrated in Figure 3.6. The encoder incorporates multi-head self-attention blocks and FCLs to handle sequential inputs. In our task, the non-visual inputs are already in vector form, thereby eliminating the need for an embedding layer in the encoder. We experimented with different configurations for the number of attention heads, number of fully connected networks, and number of hidden neurons in these networks, and the results are presented in chapter 4. As for the visual inputs, our model employs a pre-trained Vision Transformer (ViT) model for feature extraction. The ViT model is fine-tuned using the ImageNet-1K dataset, and the classification layer is removed to utilize it solely as a feature extractor. During the feature processing step, the extracted features are fused together. Various fusion mechanisms were tested, including:

- Early fusion: all the non-visual features are concatenated before processing them using the transformer encoder.
- Later fusion: all the non-visual features are concatenated after processing them using the transformer encoder.

- Hierarchical fusion: where non-visual features are concatenated in a hierarchical fashion as seen in Figure 3.4.

- Total fusion: we propose to fuse the non-visual features with visual features before processing them using a vision transformer, this occurs by applying the pose key points and the bounding box coordinates to the extracted local context of the pedestrian, which reduces the inputs of the system to three inputs only: the ego vehicle speed, the augmented local context containing information about the bounding box and pose key points and the global context of the scene. The resulting architecture is shown in Figure 3.5. This fusion method is proposed to improve the inference time of the model without affecting the prediction accuracy.

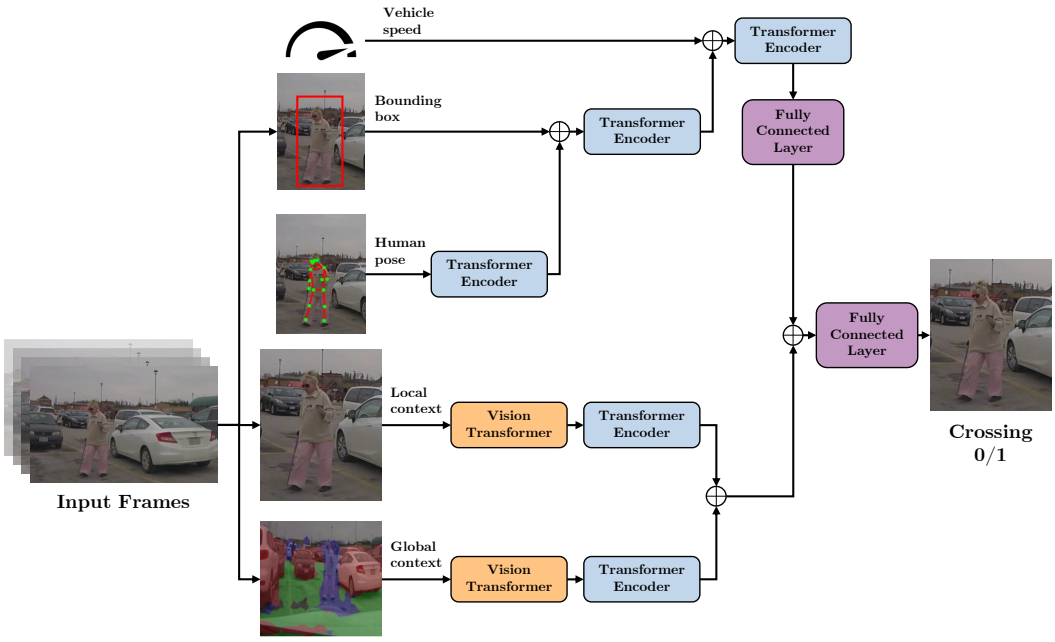

Figure 3.4: Proposed Framework: The framework utilizes transformers and vision transformers to predict the pedestrian crossing intention and uses image enhancement to maintain accuracy during adverse weather conditions.

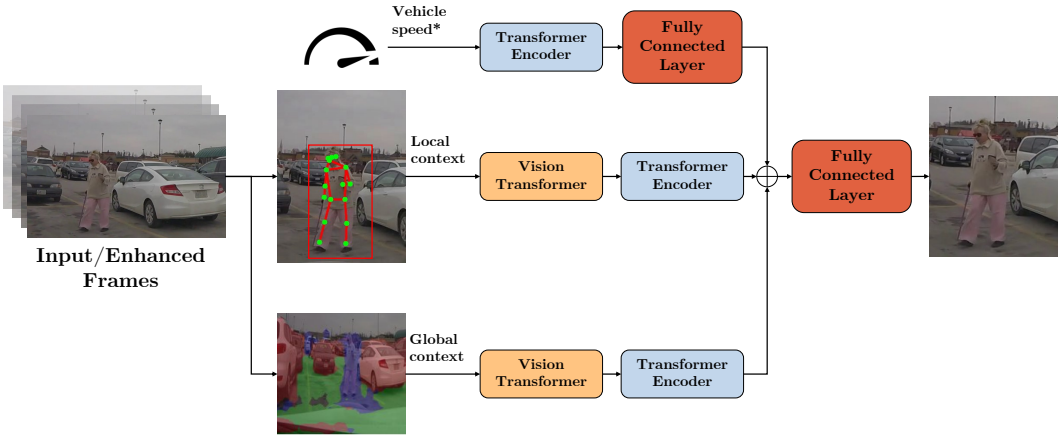

Figure 3.5: Applying Total fusion on the pedestrian intention prediction model, this variation is aimed to save inference time while maintaining high prediction accuracy.

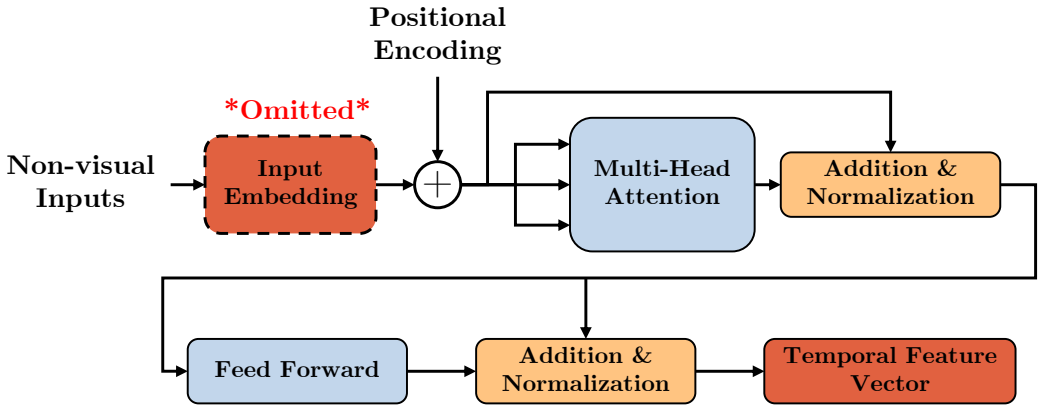

Figure 3.6: Transformer encoder model (T.E), the proposed architecture uses this model for non-visual feature extraction without the input embedding layer, the figure is edited form [34].

Following the processing of various inputs, the processed visual and non-visual features are concatenated to form the final input for the prediction step. A fully connected layer (FCL) with a sigmoid activation function is employed for making predictions. During the training phase, binary cross-entropy loss (L) (Eq.3.6) is utilized to calculate the error, and the Adam optimizer is employed to update the model weights. The adoption of the Adam optimizer is preferred in this context, as it has been shown to yield

superior results compared to stochastic gradient descent for VRU intention detection.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{3.6}$$

where N represents the total number of samples, y_i is the ground truth label of the i-th sample, and $\hat{y}_i$ is the predicted probability by the model for the i-th sample.

3.4.1 Summary

This chapter presents a detailed examination of the methodology underlying the proposed framework. It begins by providing an intricate explication of the problem definition, aiming to establish a comprehensive understanding of its nuances and fundamental elements. This sets the foundation for the subsequent discussions and analyses.

The chapter then focuses on the intricate design of the image enhancement pipeline. The overarching architecture of the pipeline is illustrated in great detail, highlighting its structure and flow. Meticulous attention is given to each individual component within the pipeline, providing a comprehensive overview of their functionalities and roles. Additionally, the chapter discusses the dataset used for both training and testing purposes, shedding light on its composition, size, and relevance to the framework.

After elucidating the image enhancement pipeline, the chapter shifts its focus to the pedestrian intention prediction model. A thorough exploration of the model's workings and underlying principles is conducted, ensuring a comprehensive comprehension of its capabilities and limitations. This examina-

tion encompasses the algorithms, architectures, and methodologies employed within the model, providing insights into how it predicts pedestrian intentions.

Chapter 4

Performance Evaluation and Discussions

4.1 Introduction

This chapter provides an overview of the findings obtained from comprehensive testing conducted on each constituent element of the framework. Initially, the performance of the image enhancement pipeline is examined, and its effect on reducing detection time with the implementation of the YOLOv4 detector is demonstrated. Subsequently, the outcomes of evaluating the intention prediction model using both the JAAD behavior and JAAD all datasets are presented. Lastly, an assessment of the overall framework's effectiveness is conducted by deploying it on a local server, where its accuracy and real-time performance are thoroughly assessed.

4.2 Image Enhancement

The evaluation of the image enhancement pipeline consisted of two distinct stages: the assessment of the multi-label classifier and the examination of the complete pipeline. Initially, the classifier underwent training and testing procedures using the locally acquired DDD dataset, resulting in a commendable testing accuracy of 80% and an inference time of 20 ms. Subsequently, to evaluate the entire pipeline, a subset of 8 videos from the JAAD dataset was employed, utilizing the YOLOv4 detection module [6]. These videos were segregated into three categories: naturally flawed videos captured under challenging conditions such as nighttime or foggy weather, artificially flawed videos created by applying blur filters and reducing image resolution, and clear videos. It is important to note that the detection module was specifically calibrated to identify pedestrians exclusively. During testing, the detection frame (DF) along with the model confidence score (C) is recorded before and after en-

hancement. Two metrics are employed to assess the impact of utilizing the image enhancement pipeline: The average improvement in the detection frame, which is calculated for each category of videos using equation 4.1

$$IDF = \frac{\sum_{n=1}^{m} DF_{be} - DF_{ae}}{m} \tag{4.1}$$

where m is the number of videos in each category, $DF_{be}$is the detection frame number before enhancement and DF_{ae} is the detection frame number after enhancement. The other metric is the improvement in the confidence score calculated using equation 4.2

$$IC = \frac{\sum_{n=1}^{m} C_{ae} - C_{be}}{m} \tag{4.2}$$

where C_{be} is the confidence score at the detection frame before enhancement and C_{ae} is the confidence score at the same frame after enhancement. Results shown in table 4.1 indicate that the image enhancement pipeline improves both the detection frame and the confidence score. On average, the pipeline achieves a reduction of 3 frames in the detection process while simultaneously improving the confidence score of the detection module. Notably, the pipeline exhibits superior performance on artificially flawed videos, where the noise follows a consistent pattern. Additionally, an important observation is that the pipeline efficiently identifies clear videos, leading to time savings in the processing stage.

Table 4.1: Results of testing the image enhancement pipeline.

Video Type	**Improvement in detection frame**	**Improvement in confidence score**
Naturally flawed	2.5	4%
Artificially flawed	3	12.5%
Clear	0	0%

Table 4.2: Results after some videos feed into the image enhancement pipeline

Video number	**Description**	**Detection frame in original video (confidence score)**	**Detection frame in enhanced video (confidence score)**	**Multi-label classifier output**	**Videos Type**
124	Driving at low speed at a cloudy afternoon with pedestrians alongside the road	1 (95%)	3 (75%)	[1,0,0]	Naturally contaminated
128	Driving on a snow road on a cloudy afternoon with pedestrians alongside the road	14 (32%)	14 (32%)	[0,0,0]	Clear Video
26	Driving at night with pedestrians crossing the road	1 (93%)	1 (93%)	[1,0,1]	Naturally contaminated
281	Driving on a snow road on a cloudy afternoon with pedestrians alongside the road	10 (38%)	1 (26%)	[1,0,0]	Naturally contaminated
34	Driving in traffic during mid-day but stopped for a while in traffic with pedestrians alongside the road	473 (27%)	473 (27%)	[1,0,0]	Artificially contaminated
346	Driving at dawn with pedestrian crossing the road	4 (31%)	1 (27%)	[0,1,0]	Naturally contaminated
51	Driving in a parking lot at noon with people walking by	21 (25%)	21 (38%)	[0,0,0]	Clear Video
58	Driving a car at noon with other cars crossing in front of the camera to go to the store's parking lot and pedestrians walking on the side walk	11 (36%)	5 (37%)	[1,0,0]	Artificially contaminated

From Table 4.2 we observe that the multi-label classifier performed well with both artificially contaminated and naturally contaminated videos which shows the robustness of the model.

Some videos such as 128, 34, 51 had no difference in the number of frames

it took the YOLO model to detect pedestrians. This happened because these videos' frames were clear so the YOLO model could detect a person walking on the same number of frames. This result is also detected by the multi-label classifier which produced an output of zeros for videos 51,128. Although the multi-label classifier miss-labeled video 34. We can observe that the effect of the blurring filter is non-destructive on the input frames which is also a good feature in classical image enhancement algorithms. Furthermore, by using classical image enhancement techniques, the reduced number of frames was not huge, but, by enhancing those techniques, we can make the required number of frames to detect pedestrians much less.

4.3 Pedestrian Intention Prediction

The testing phase of our proposed model was conducted on a local server equipped with an A6000 RTX GPU. Figures 4.1,A.16,4.3,4.4 show the confusion matrices and the ROC curves for the proposed model on JAAD behavior and JAAD all datasets. The confusion matrix gives a good indication of the model's performance in terms of accuracy. The ROC curve shows the relation between the true positive rate and the false positive rate with different decision thresholds. This relation provides an insight into which threshold is the best and the area under the ROC curve is the AUC metric which indicates if a model generally tends to have a large true positive rate with different thresholds or not. A model with this characteristic tends to perform better on unseen data. The confusion matrices and ROC curves for other state-of-the-art models are included in the appendix. Table 4.3 and Table 4.4 provide a quantitative comparison between our model and several existing models, namely Single RNN [16], SF-GRU [28], PCPA [17], and Mask_PCPA [39].

We utilized default classification testing metrics to evaluate the performance of the models. Additionally, we measured the testing time, which represents the total duration required by a model to complete the prediction of all sequences in the testing set. It serves as an indicator of the inference time needed by the model. Since our model was evaluated using both the $JAAD_{behavior}$ dataset and the $JAAD_{all}$ dataset, there are two testing sets and two corresponding testing times. The results demonstrate that our proposed model outperforms the other models in terms of accuracy, AUC, F1 score, recall, and testing time on the $JAAD_{behavior}$ dataset. Moreover, our model maintains high accuracy and recall on the $JAAD_{all}$ dataset. One important

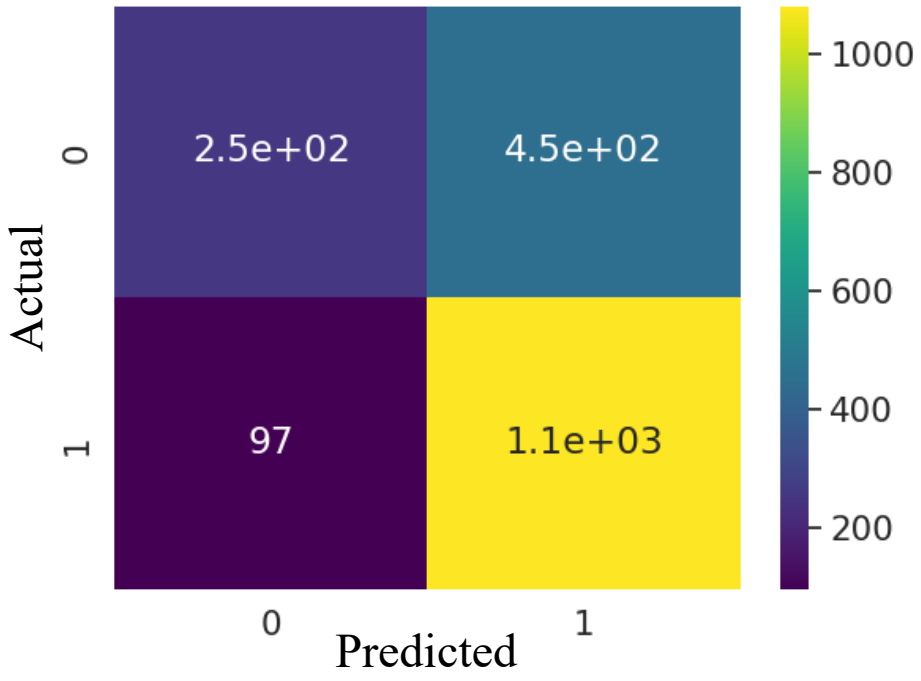

Figure 4.1: Confisuion matrix of the proposed prediction model on JAAD behavior

note is the comparison between precision and recall in this task. Precision reflects the accuracy of capturing a crossing behavior which means that high precision values reflect a low occurrence of false alarms. On the other hand, recall reflects the accuracy of capturing all the crossing samples in our testing sets. This means that a high recall value results in a low occurrence of missing a crossing behavior. From this note, we can agree that high recall values are crucial for the deployment of this framework in automated vehicles. However, precision receives more attention when deploying this framework in an advanced driver-assistance system as the driver can take over in a crossing misclassification event. To optimize the performance of our model, we

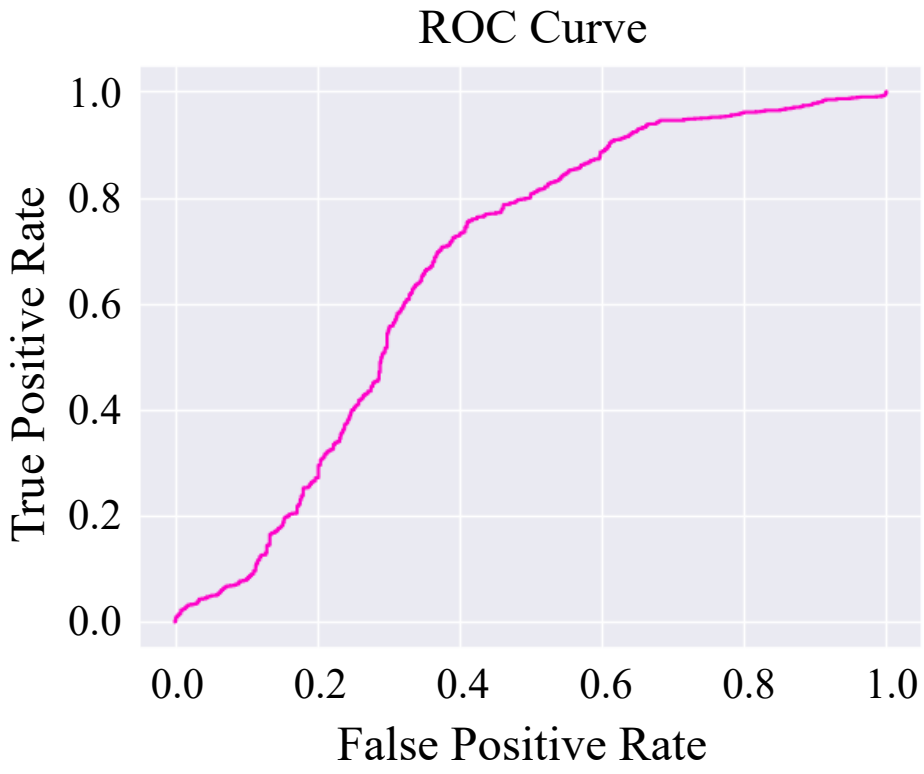

Figure 4.2: ROC curve of the proposed prediction model on JAAD behavior

Table 4.3: Performance comparison between the proposed model and previous models on JAAD behavior.

Model	Used Blocks	Accuracy	AUC	F1	Preci	Recall	Time (s)
SingleRNN	VGG + GRU	0.59	0.52	0.71	0.64	0.80	11.91
SF-GRU	VGG + GRU	0.58	0.56	0.65	**0.68**	0.62	13.71
PCPA	3D CNN	0.53	0.53	0.59	0.66	0.53	15.2
Mask_PCPA	VGG + GRU	0.62	0.54	0.74	0.65	0.85	12.66
Ours(7)	T.Encoder + ViT	**0.67**	**0.60**	**0.77**	0.68	**0.90**	**11.62**

conducted a study to tune different hyper-parameters such as the number of attention heads and the number of hidden neurons in FCLs (ffhn). The results are presented in Table 4.5. It's observed that increasing the number of attention heads had minimal impact on the model's accuracy and could even lead to decreased performance. This can be attributed to the increased number of

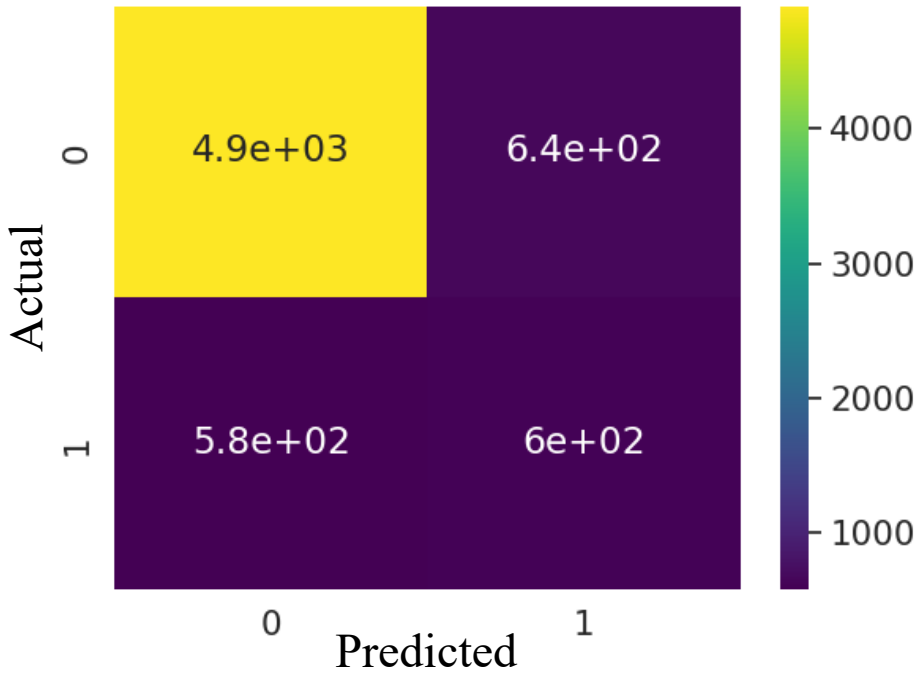

Figure 4.3: Confisuion matrix of the proposed prediction model on JAAD all

Table 4.4: Performance comparison between the proposed model and previous models on Jaad all.

Model	Used Blocks	Accuracy	AUC	F1	Precision	Recall	Time (s)
SingleRNN	VGG + GRU	0.79	0.76	0.54	0.44	0.71	30.11
SF-GRU	VGG + GRU	0.76	0.77	0.53	0.4	0.79	31.6
PCPA	3D CNN	0.76	0.79	0.55	0.41	0.83	75.2
Mask_PCPA	VGG + GRU	**0.83**	**0.82**	**0.63**	**0.51**	0.81	38.56
Ours(7)	T.Encoder + ViT	**0.83**	0.80	0.62	0.47	**0.82**	**30**

trainable parameters associated with additional attention heads, which may cause the model to struggle in finding a local minimum due to limited training data. For these reasons the usage of one attention head is preferred for our prediction model. Similarly, increasing the number of epochs beyond a certain point may result in overfitting. Increasing the number of hidden neurons in

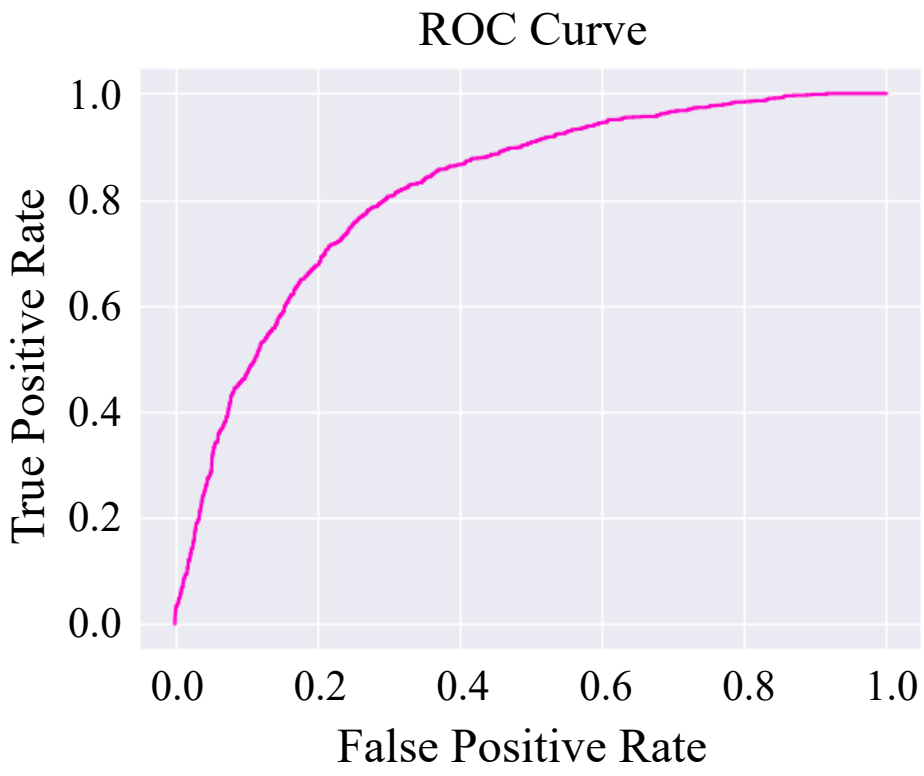

Figure 4.4: ROC curve of the proposed prediction model on JAAD all

the FCL of the transformer encoder leads to better accuracy but increases the total number of floating point operations (FLOPS) needed to calculate the prediction which increases the model testing time. Table 4.6 shows different

Table 4.5: Results of tuning different hyper-parameters of the proposed model.

num_heads	ffhn	Accuracy	AUC	F1	Precision	Recall	Time (s)	MFLOPS
1	2048	0.66	0.57	**0.78**	0.66	0.94	13.2	111
3	2048	0.66	0.57	0.77	0.66	0.91	13.97	112
5	2048	0.64	0.55	0.76	0.65	0.91	18.23	112
1	1024	0.64	0.55	0.76	0.65	0.91	**12.99**	56.1
1	4096	**0.67**	**0.58**	**0.78**	**0.67**	**0.95**	14.29	222

variations of the proposed architecture, in the first variation the FCLs in each transformer encoder are removed and only one FCL is used after concatenat-

Table 4.6: Results of different variations of the proposed intention prediction model.

Model variation	Accuracy	AUC	F1	Precision	Recall	Time (s)	MFLOPS
(Ours) Hierarchical fusion	0.67	0.58	0.78	0.67	0.94	13.24	108
(Ours1) One Feedforward Layer	0.66	0.57	0.77	0.66	0.93	**9.90**	**50.3**
(Ours2) Transformers + VGG-19	0.68	0.58	**0.78**	0.68	0.93	22.18	151
(Ours3) Removing GC	**0.70**	**0.68**	0.77	**0.74**	0.89	10.34	66.5
(Ours4) Using local attention	0.68	0.61	0.77	0.69	0.87	17.76	117
(Ours5) Later fusion	0.67	0.57	0.78	0.66	**0.95**	13.068	107
(Ours6) Early fusion	0.66	0.55	0.78	0.66	0.89	12.86	105
(Ours7) Total fusion	0.67	0.60	0.77	0.68	0.90	11.62	101

ing all the non-visual features, this variation resulted in the best testing time with a negligible loss in accuracy. This modification reduces the computational complexity of the model while maintaining reasonable accuracy. Using VGG-19 as a visual encoder instead of ViT improved the accuracy of the model but significantly increased the testing time. This can be attributed to the fact that ViT performs better than CNN-based architectures when trained on a large amount of data. Removing the global context (GC) from the model inputs led to improved testing time while maintaining high accuracy. The impact of the global context becomes more prominent with a larger training dataset. In the case of testing the same model on the $JAAD_{all}$ dataset, where the amount of training data is larger, the model achieved lower accuracy without the GC. However, removing the GC still provides a quick and reasonably accurate decision. Also, when replacing the global self-attention unit in the multi-head attention of the transformer with a local self-attention unit the model accuracy increases as local self-attention considers attention weights in a specific window of frames. This moving window is set to 4 frames in this study this means that for the input sequence of frames, we calculate the attention weights using the first 4 frames from the input and then calculate the weights for the next 4 frames and so on till the end of the input sequence then concatenate these

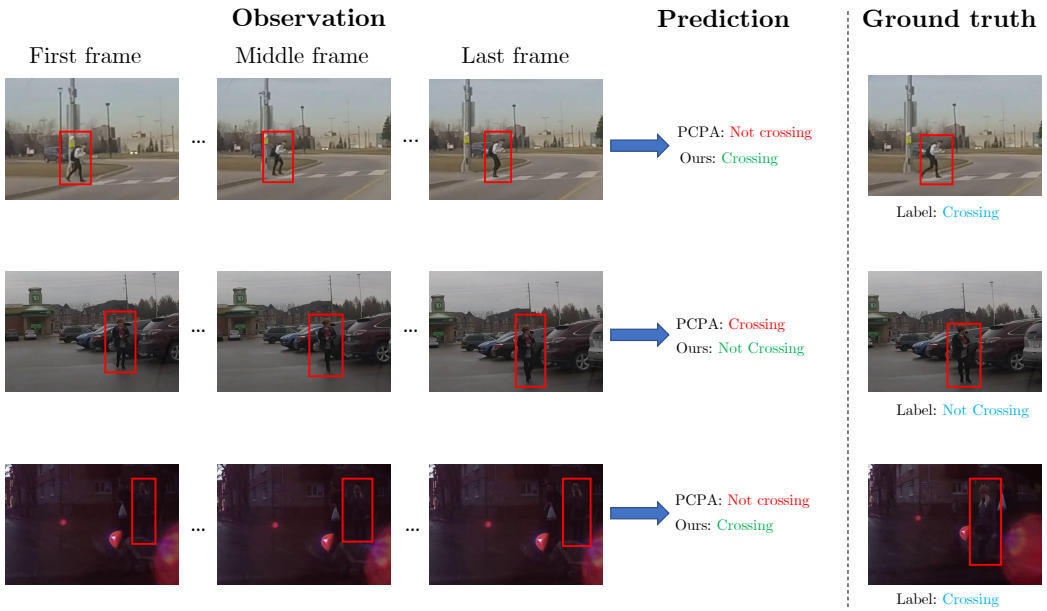

Figure 4.5: A qualitative comparison between our model and PCPA model: The comparison highlights the robustness of our model and the ability to predict the intention of pedestrians even in bad lighting conditions.

weights together. This method improves the accuracy when long-term dependencies between the input sequence don't exist. Moreover, the proposed total fusion method provided decent accuracy and also a reasonable testing time. Figure 4.5 shows a qualitative comparison between our model and PCPA [17].

4.4 End To End Deployment

The final step is to test the entire framework, to deploy the framework in real-time we use MoveNet [3] model to fetch the pose key points and the pedestrian bounding box, the advantage of using this model over openpose [8] is that MoveNet is much faster and can provide the bounding box of the pedestrian which removes the need to use a dedicated model for pedestrian detection. The downside of this model is that it can only provide poses for up to 6 pedestrians

in a single frame. DeepSort [36] is used to track the detected pedestrian, and Deeplabv3 [10] is utilized to extract the segmentation mask of the scene. To test the inference time of the prediction model and the end-to-end inference time which includes the time taken by the other models to fetch the required inputs a single video not included in the JAAD dataset with one pedestrian crossing the street is used. This video resembles the first case of the Euro NCAP pedestrian safety tests. The speed of the ego vehicle in the test video is slow so the input of the vehicle speed is set to 1. Table 4.7 shows the inference time of each prediction model and the total end-to-end time from the moment of detection till the final prediction, the table also includes whether the model was able to successfully classify the target pedestrian or not. Results show that removing the GC improves the end-to-end time significantly while being able to successfully identify the pedestrian intention. The low inference time and end-to-end time achieved by our models indicate that a perception model that utilizes our framework will most likely pass the Euro NCAP pedestrian safety tests. In general, transformer-based models performed better than RNN and CNN-based architecture in terms of identifying the intention of the pedestrian

Table 4.7: Results of testing the proposed framework in real time environment.

Model name	Inference time	End-to-end time	Prediction result
(Ours1) One Feedforward Layer	21	165	Correct
(Ours2) Transformers + VGG-19	22	182	Correct
(Ours3) Removing GC	18	72	Correct
SingleRNN	**17**	175	Incorrect
(Ours4) Using local attention	25	190	Correct
SF-GRU	18	178	Incorrect
PCPA	25	**70**	Incorrect
Mask_PCPA	23	185	Correct
(Ours5) Later fusion	20	180	Correct
(Ours6) Early fusion	20	179	Correct
(Ours7) Total fusion	20	130	Correct

as transformers are less prone to domain adaptation problems and generalize better than CNNs and RNNs. From the results, we recommend the usage of a switching mechanism when deploying this model in a real-world vehicle, taking into consideration the ego vehicle speed and estimating the distance between the ego vehicle and the target pedestrian using a depth camera the perception model can switch between using a fast variation of the proposed model such as Ours3 and a slower but more accurate version such as Ours2 or Ours4.

4.4.1 Summary

This chapter presents an overview of the findings obtained from comprehensive testing conducted on each constituent element of the framework. The evaluation begins by examining the performance of the image enhancement pipeline. The impact of the pipeline on reducing detection time is demonstrated, specifically with the implementation of the YOLOv4 detector. The chapter highlights the improvements achieved through the pipeline, emphasizing its contribution to efficient and effective detection.

Next, the outcomes of evaluating the intention prediction model are presented. Two datasets, namely the JAAD behavior dataset and the JAAD all dataset, are utilized for this evaluation. The chapter discusses the results of these evaluations, providing insights into the accuracy and performance of the intention prediction model. This analysis helps to assess the model's ability to predict pedestrian intentions accurately and reliably.

Furthermore, the chapter conducts an assessment of the overall framework's effectiveness by deploying it on a local server. The accuracy and real-time performance of the framework are thoroughly evaluated in this deployment scenario. The chapter presents the findings of this assessment, discussing the

strengths and limitations of the framework and providing a comprehensive understanding of its practical viability.

All in all, this chapter focuses on the evaluation of the framework components and its overall effectiveness. It begins by examining the performance of the image enhancement pipeline, emphasizing its impact on reducing detection time. The outcomes of evaluating the intention prediction model using different datasets are presented, shedding light on its accuracy and performance. Lastly, the chapter assesses the overall effectiveness of the framework by deploying it on a local server, providing insights into its real-time performance and practical viability.

Chapter 5

Conclusions

5.1 Summary and conclusions

In this study, we proposed a comprehensive framework for pedestrian crossing intention prediction. The framework combines a prediction model utilizing self-attention and vision transformer with an image enhancement pipeline incorporating CNNs and classical image enhancement techniques. The primary objective of the framework is to maintain prediction accuracy, particularly during adverse weather conditions, by enhancing input frames before the prediction stage. Our framework was evaluated using a locally collected and annotated dataset, and it achieved a testing accuracy of 80% with the multilabel classifier.

The prediction model, which employed self-attention and vision transformer, demonstrated state-of-the-art performance on the JAAD behavior dataset. Through extensive examination and experimentation, several variants of the prediction model were developed, and an ablation study was conducted to identify the most effective components. This rigorous evaluation helped to refine the prediction model and understand the impact of different architectural choices.

The image enhancement pipeline, consisting of CNNs and classical image enhancement techniques, played a crucial role in improving the input frames' quality before the prediction stage. By leveraging these techniques, the pipeline successfully enhanced the frames and mitigated the effects of bad weather conditions, ensuring accurate and reliable pedestrian crossing intention prediction.

To assess the real-time performance of the entire framework, it was deployed on a local server, and its inference time and end-to-end time were

compared against various models. Remarkably, our framework outperformed other models, achieving the best inference time and end-to-end time. This result highlights the efficiency and practical viability of our proposed framework for real-time applications.

Overall, our framework offers a holistic solution for pedestrian crossing intention prediction, addressing the challenges posed by adverse weather conditions. By integrating a prediction model with self-attention and vision transformer and an image enhancement pipeline using CNNs and classical image enhancement techniques, we achieved high prediction accuracy and real-time performance. The extensive examination of the prediction model and the deployment of the framework on a local server underscore the robustness and effectiveness of our approach.

In conclusion, our framework showcases the potential of combining self-attention, vision transformer, CNNs, and classical image enhancement techniques to address the challenge of pedestrian crossing intention prediction under adverse weather conditions. With its high accuracy, real-time performance, and robustness, our framework lays the foundation for safer and more reliable pedestrian crossing assistance systems in various real-world scenarios.

5.2 Limitations and Future work

The limitations of our framework are mentioned as follows:

1. Our first limitation is the limited amount of data used to train our multilabel classifier which may affect the ability of our classifier to generalize on different sources of data. Our multilabel classifier was trained and tested on the DDD dataset which contains 75 dashcam videos, expanding the dataset will enable us to improve the performance of our classifier.

2. Our image enhancement pipeline mainly relies on the usage of classical image enhancement modules to perform the enhancement task, although these modules offer an extremely fast processing time that reaches 5ms, they are only capable of producing moderate performance in terms of the quality of the output frame.

3. Finally, our framework was trained and tested on input frames extracted from front dashboard cameras which restricts its usage to vehicles. Some changes can be done to enable the usage of our framework in the traffic infrastructure.

Our Future work can be summarized in these bullet points :

1. Expand the dataset: Consider augmenting and diversifying the dataset used for training the multilabel classifier. This could involve collecting additional data under various weather conditions, at different time of day, and in different geographical locations. A larger and more diverse dataset would provide more comprehensive training for the classifier and improve its generalization capabilities.

2. Explore advanced image enhancement techniques: Investigate the use of more advanced image enhancement techniques, including deep learning-based methods. Techniques such as generative adversarial networks (GANs) or attention mechanisms specifically designed for image enhancement could be explored to further improve the quality and clarity of input frames, particularly in challenging weather conditions.

3. Fine-tune for overhead camera inputs: Adapt and fine-tune the prediction model and the multilabel classifier to handle inputs from overhead cameras instead of dashcam footage. Smart intersections often employ overhead cameras for pedestrian monitoring and safety. Fine-tuning the framework to work with such camera inputs would enable its deployment in smart intersections, enhancing pedestrian safety in real-world settings.

Appendix A

Appendix

In this appendix, we present ROC curves and confusion matrices of different intention prediction models for the purpose of comparison. ROC curves demonstrate the relation between the true positive rate and the false positive rate and the area under the ROC curve is the AUC metric which is commonly used to evaluate classification models where high AUC represents a good model that is not biased to a certain class.

The confusion matrix presents the number of true positives, true negatives, false positives, and false negatives for a classification model on the test dataset, these numbers give an indication of the model performance and can be used to calculate popular metrics such as precision and recall. All of our testing was done on both JAAD behaviour and JAAD all datasets so every model has two ROC curves and two confusion matrices.

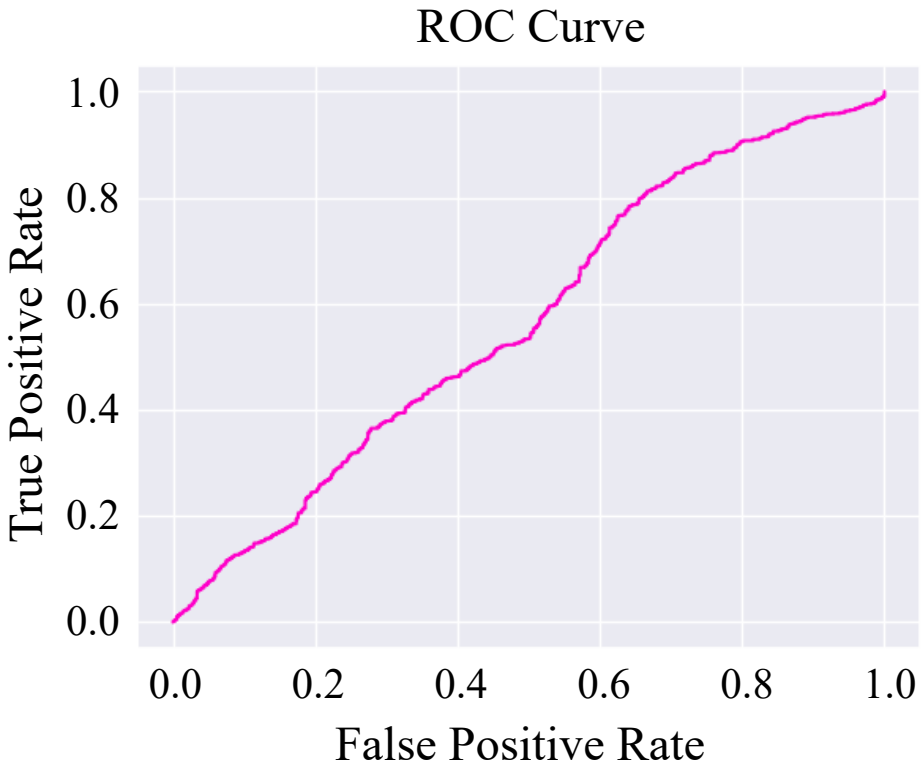

Figure A.1: ROC curve of Mask_PCPA model [39] on JAAD behavior

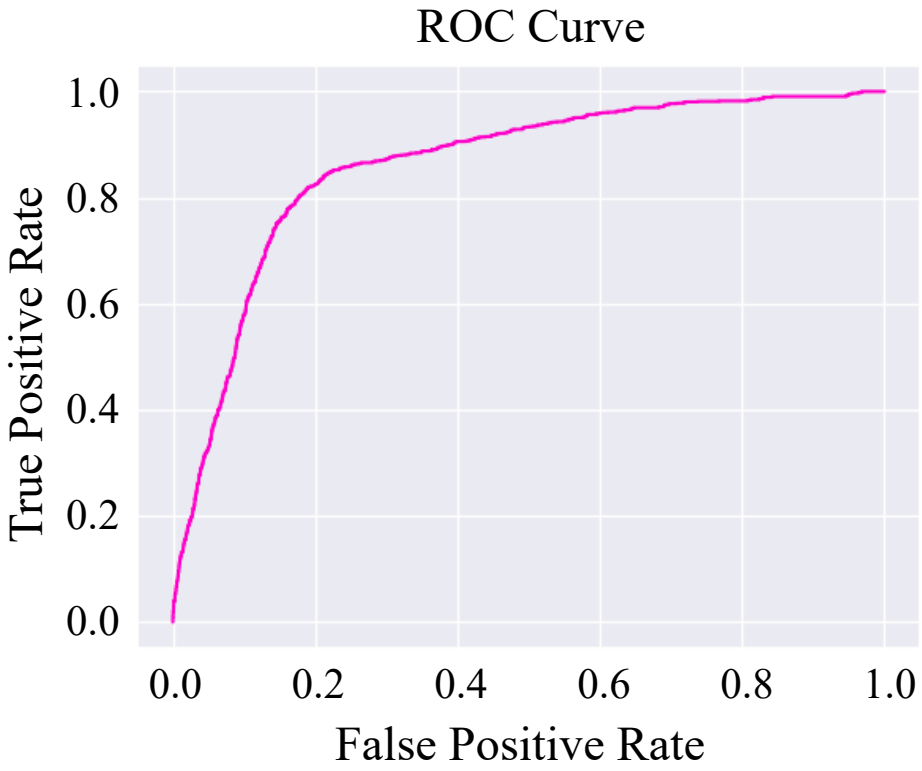

Figure A.2: ROC curve of Mask_PCPA model [39] on JAAD all

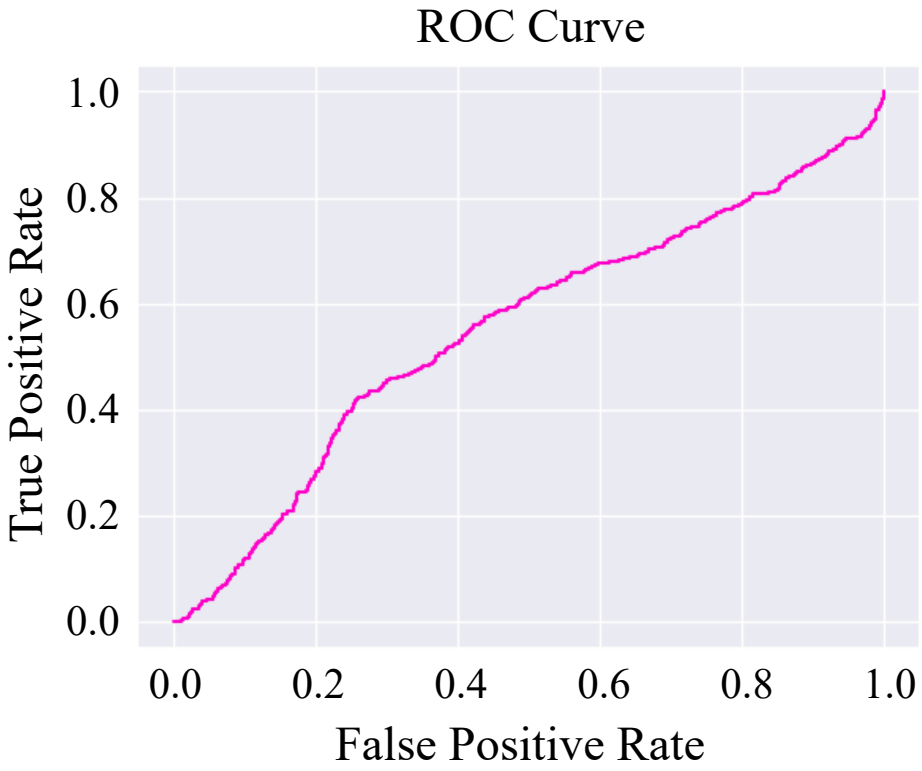

Figure A.3: ROC curve of SF_GRU model [22] on JAAD behavior

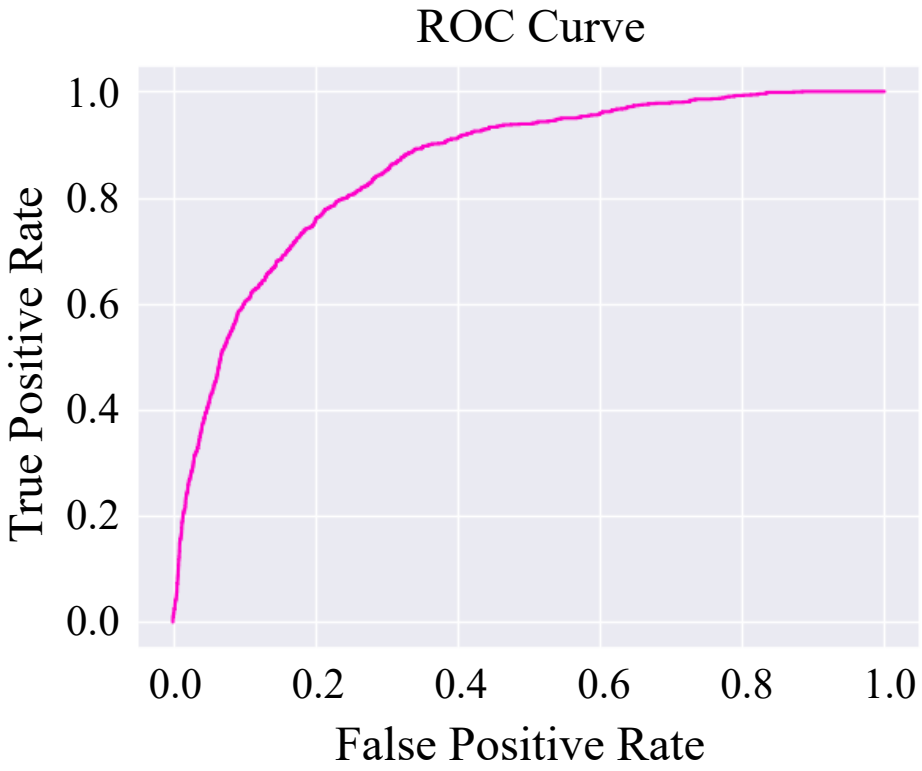

Figure A.4: ROC curve of SF_GRU model [22] on JAAD all

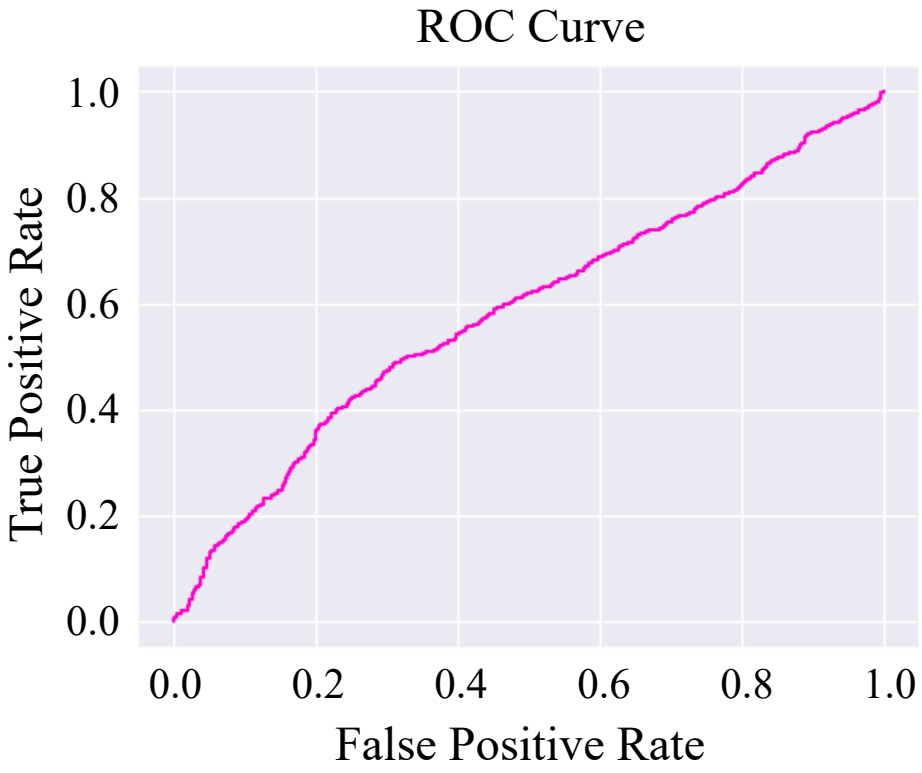

Figure A.5: ROC curve of single RNN model [16] on JAAD behavior

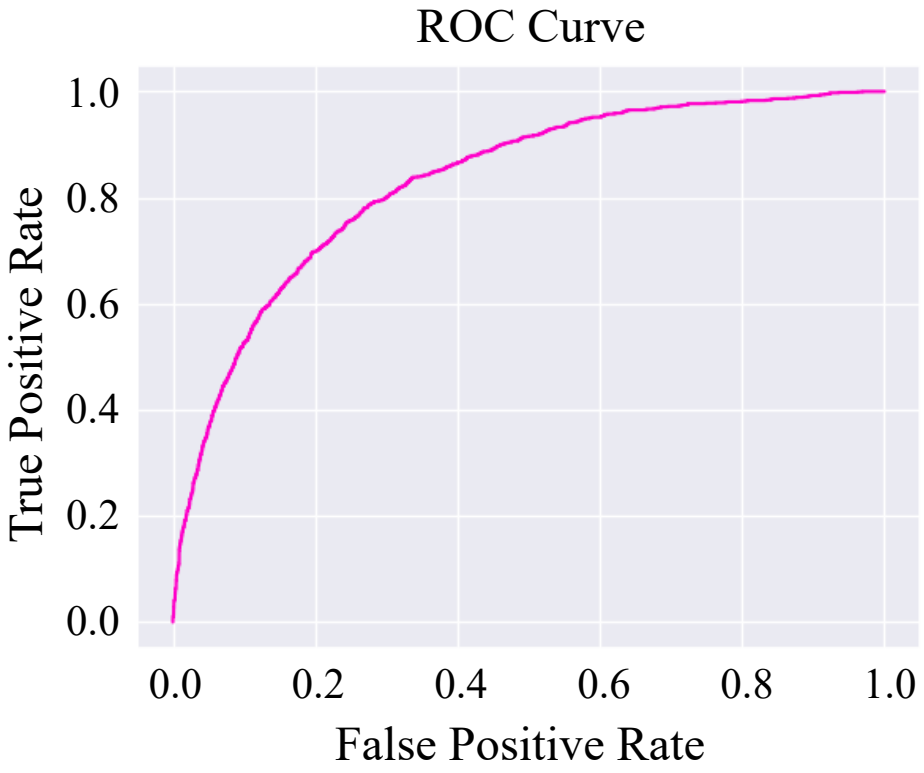

Figure A.6: ROC curve of single RNN model [16] on JAAD all

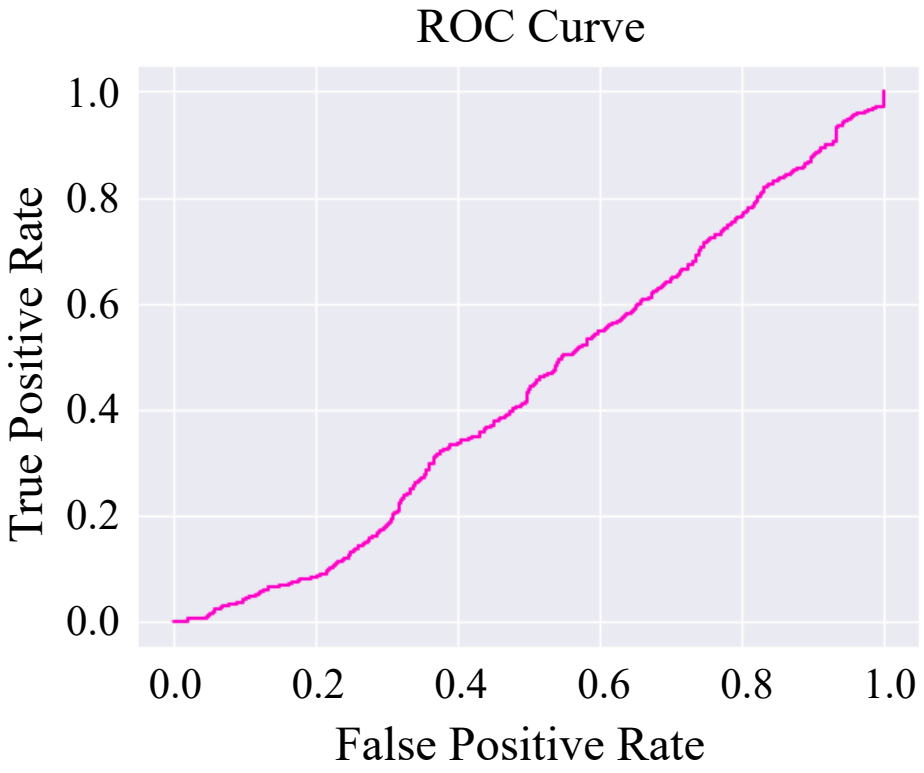

Figure A.7: ROC curve of PCPA model [17] on JAAD behavior

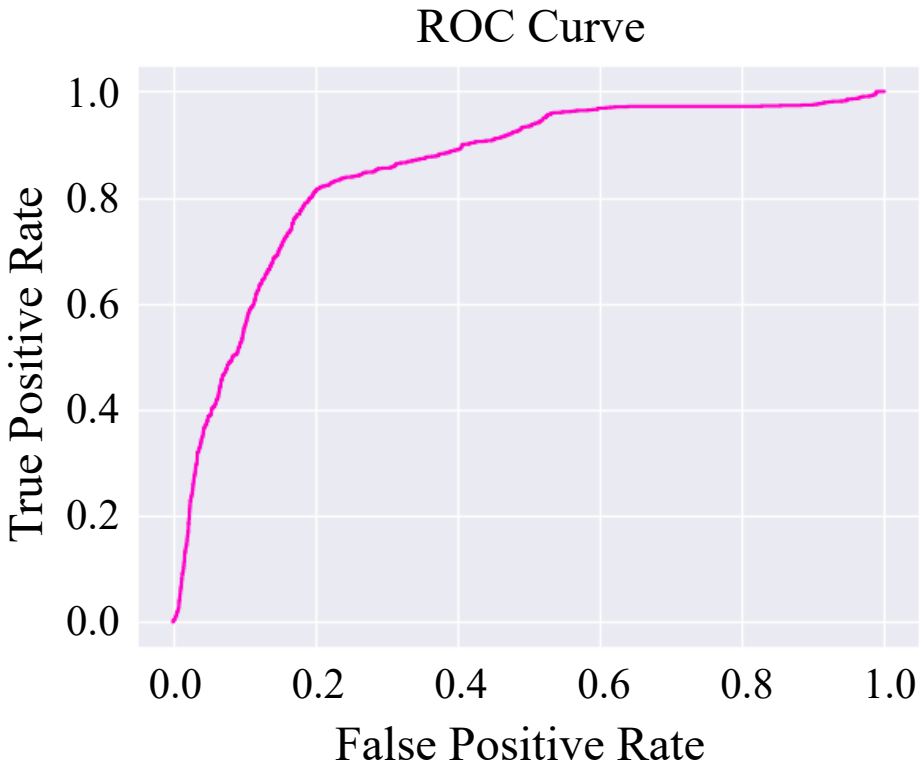

Figure A.8: ROC curve of PCPA model [17] on JAAD all

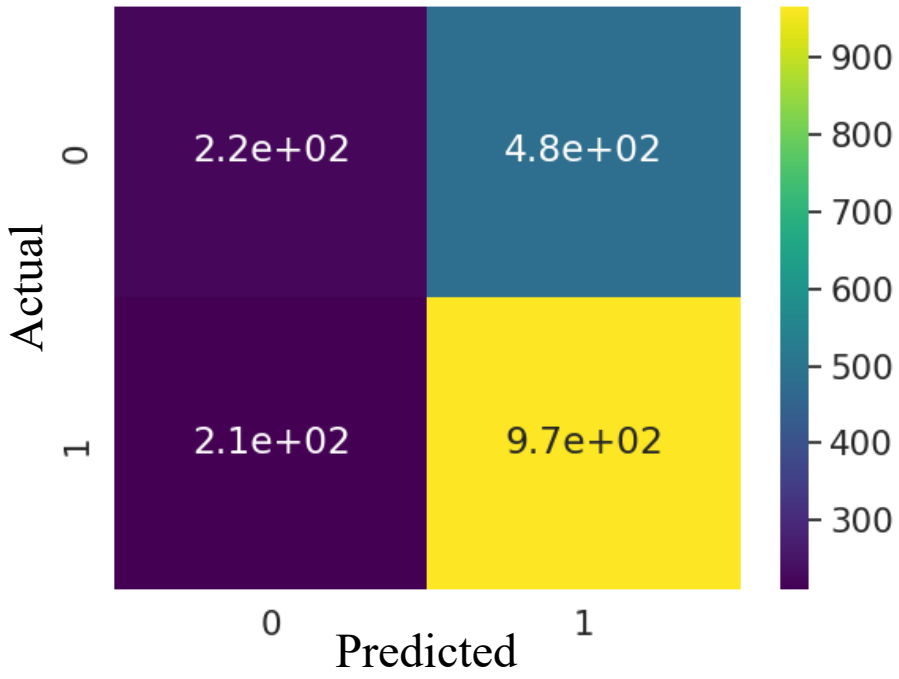

Figure A.9: Confisuion matrix of Mask_PCPA model [39] on JAAD behavior

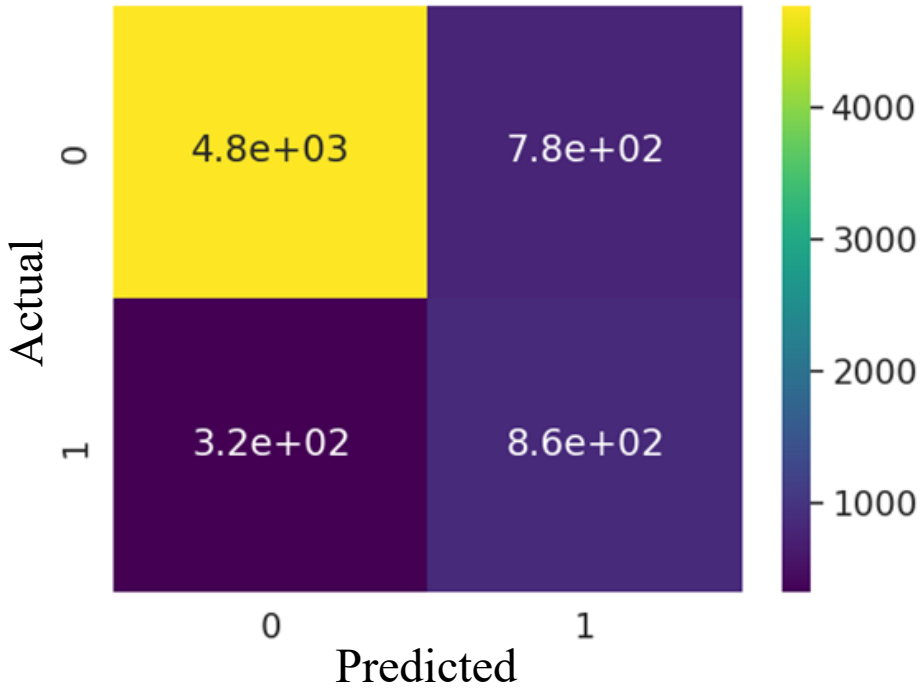

Figure A.10: Confisuion matrix of Mask_PCPA model [39] on JAAD all

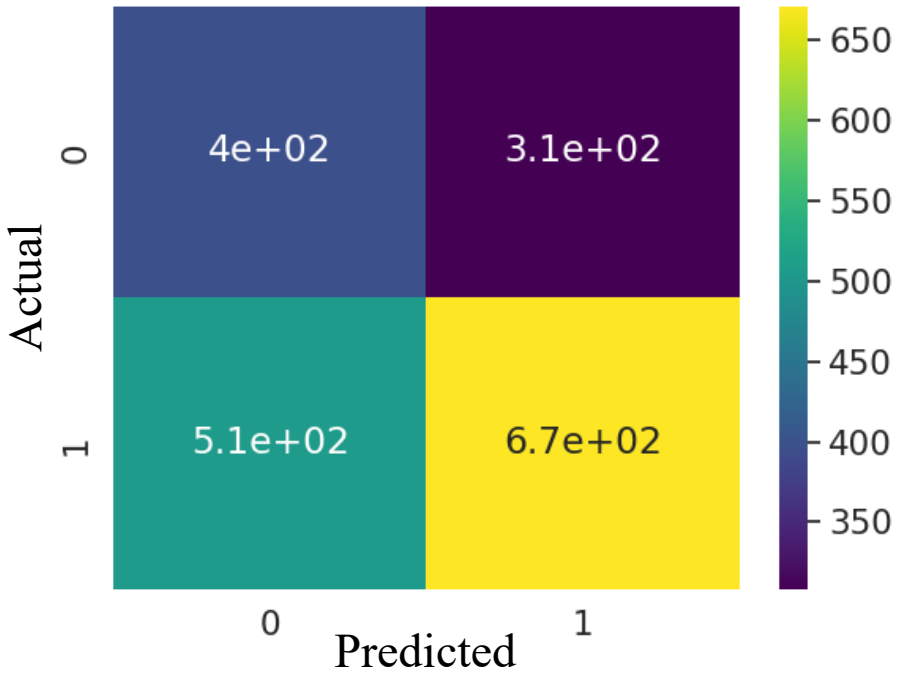

Figure A.11: Confisuion matrix of SF_GRU model [22] on JAAD behavior

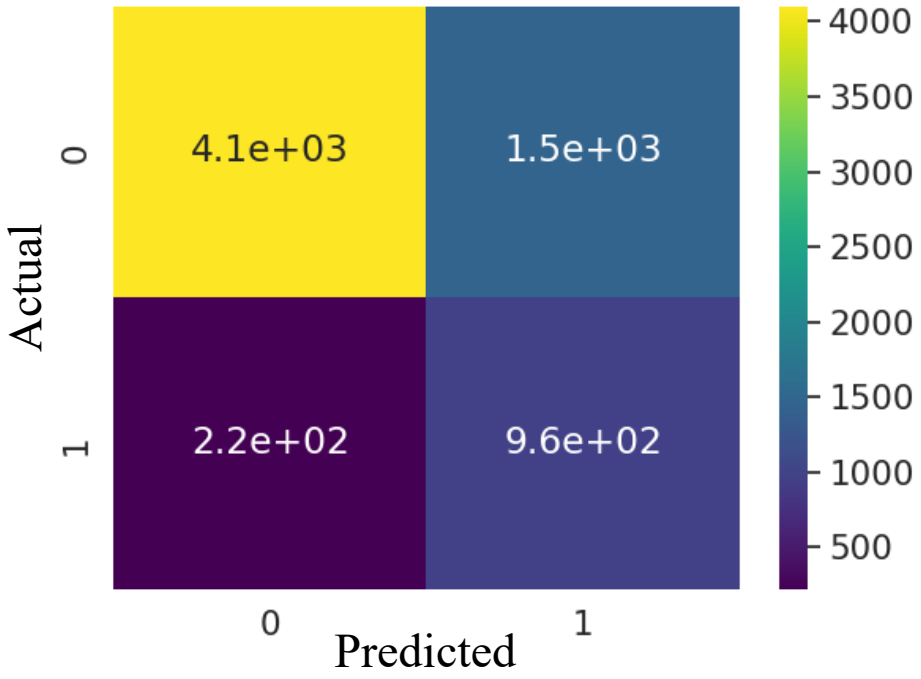

Figure A.12: Confisuion matrix of SF_GRU model [22] on JAAD all

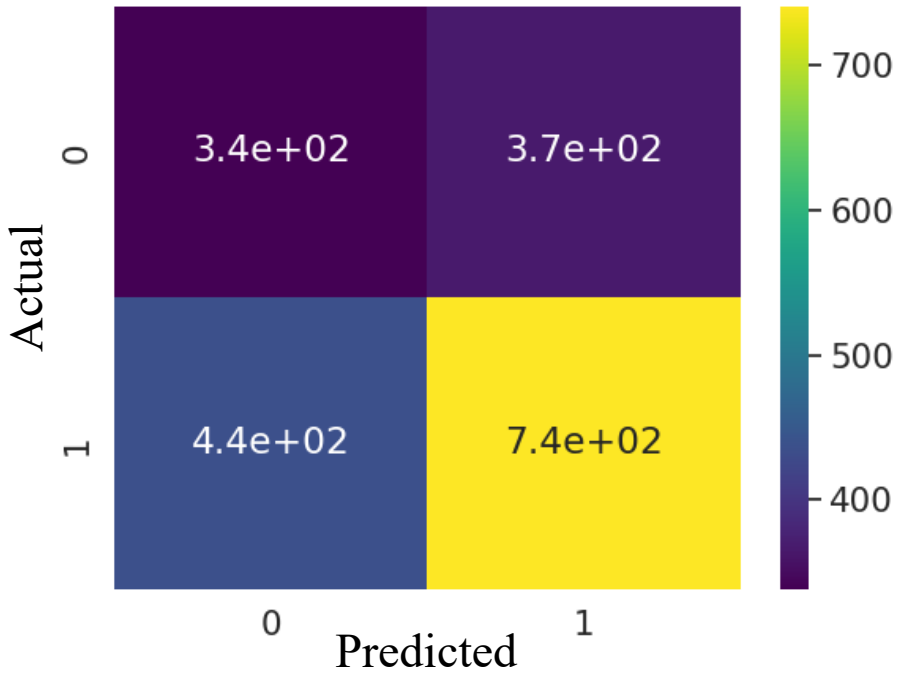

Figure A.13: Confisuion matrix of single RNN model [16] on JAAD behavior

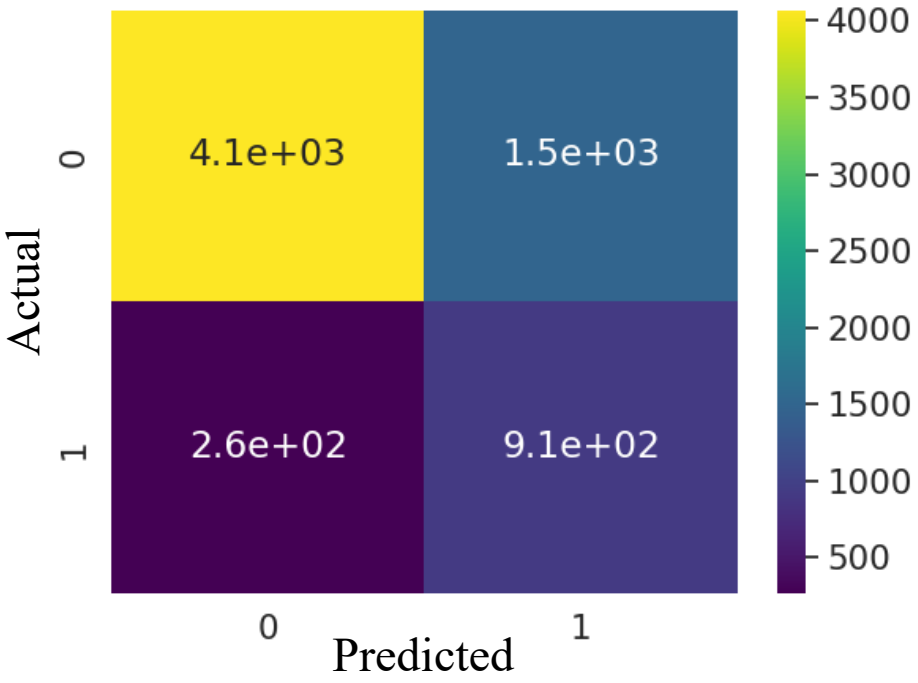

Figure A.14: Confisuion matrix of single RNN model [16] on JAAD all

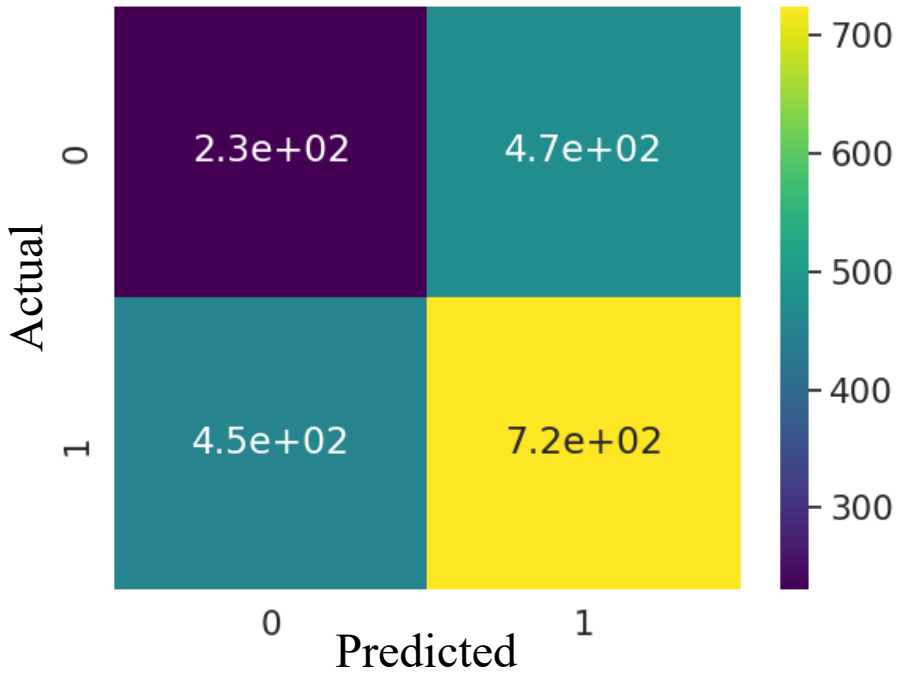

Figure A.15: Confisuion matrix of PCPA model [17] on JAAD behavior

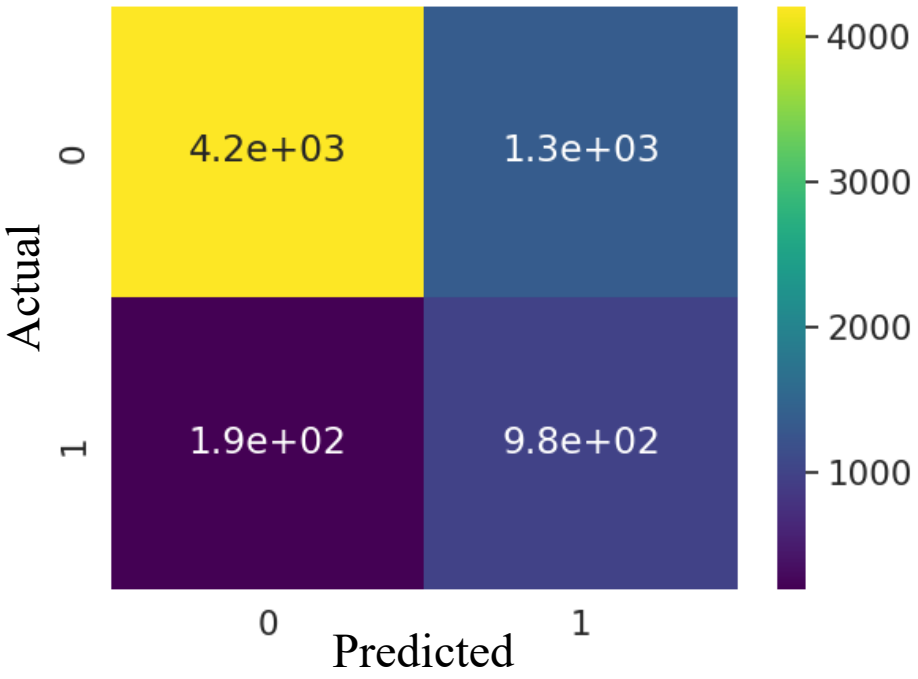

Figure A.16: Confisuion matrix of PCPA model [17] on JAAD all

Bibliography

[1] Al-Ameen, Z., Sulong, G., Johar, M. G. M., Verma, N., Kumar, R., Dachyar, M., Alkhawlani, M., Mohsen, A., Singh, H., Singh, S., et al. A comprehensive study on fast image deblurring techniques. *International Journal of Advanced Science and Technology 44* (2012).

[2] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate, 2016.

[3] Bajpai, R., and Joshi, D. Movenet: A deep neural network for joint profile prediction across variable walking speeds and slopes. *IEEE Transactions on Instrumentation and Measurement 70* (2021), 1–11.

[4] Bengio, Y., Frasconi, P., and Simard, P. Problem of learning long-term dependencies in recurrent networks. pp. 1183 – 1188 vol.3.

[5] Bhattacharyya, A., Fritz, M., and Schiele, B. Long-term on-board prediction of people in traffic scenes under uncertainty. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 4194–4202.

[6] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

[7] Bouhsain, S. A., Saadatnejad, S., and Alahi, A. Pedestrian intention prediction: A multi-task perspective. *arXiv preprint arXiv:2010.10270* (2020).

[8] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 1302–1310.

[9] Carreira, J., and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 4724–4733.

[10] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. Rethinking atrous convolution for semantic image segmentation, 2017.

[11] Chen, S.-D., and Ramli, A. Contrast enhancement using recursive mean-separate histogram equalization for scalable brightness preservation. *IEEE Transactions on Consumer Electronics 49*, 4 (2003), 1301–1309.

[12] Chen, S.-D., and Ramli, A. R. Contrast enhancement using recursive mean-separate histogram equalization for scalable brightness preservation. *IEEE Transactions on consumer Electronics 49*, 4 (2003), 1301–1309.

[13] Cheng, H.-D., and Shi, X. A simple and effective histogram equalization approach to image enhancement. *Digital signal processing 14*, 2 (2004), 158–170.

[14] Dong, C., Loy, C. C., He, K., and Tang, X. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision* (2014), Springer, pp. 184–199.

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[16] Kotseruba, I., Rasouli, A., and Tsotsos, J. K. Do they want to cross? understanding pedestrian intention for behavior prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)* (2020), pp. 1688–1693.

[17] Kotseruba, I., Rasouli, A., and Tsotsos, J. K. Benchmark for evaluating pedestrian action prediction. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2021), pp. 1257–1267.

[18] Liu, B., Adeli, E., Cao, Z., Lee, K.-H., Shenoi, A., Gaidon, A., and Niebles, J. C. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters 5*, 2 (2020), 3485–3492.

[19] Lore, K. G., Akintayo, A., and Sarkar, S. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition 61* (2017), 650–662.

[20] Lv, Z., Huang, X., and Cao, W. An improved gan with transformers for pedestrian trajectory prediction models. *International Journal of Intelligent Systems 37*, 8 (2022), 4417–4436.

[21] Nah, S., Hyun Kim, T., and Mu Lee, K. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 3883–3891.

[22] Ng, J. Y.-H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. Beyond short snippets: Deep networks for video classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 4694–4702.

[23] Palumbo, D., Yee, B., O'Dea, P., Leedy, S., Viswanath, S., and Madabhushi, A. Interplay between bias field correction, intensity standardization, and noise filtering for t2-weighted mri. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2011), IEEE, pp. 5080–5083.

[24] Qi, Y., Yang, Z., Sun, W., Lou, M., Lian, J., Zhao, W., Deng, X., and Ma, Y. A comprehensive overview of image enhancement techniques. *Archives of Computational Methods in Engineering 29*, 1 (2022), 583–607.

[25] Quan, R., Zhu, L., Wu, Y., and Yang, Y. Holistic lstm for pedestrian trajectory prediction. *IEEE transactions on image processing 30* (2021), 3229–3239.

[26] Rasouli, A., Kotseruba, I., Kunic, T., and Tsotsos, J. K. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 6262–6271.

[27] Rasouli, A., Kotseruba, I., and Tsotsos, J. K. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (2017), pp. 206–213.

[28] Rasouli, A., Kotseruba, I., and Tsotsos, J. K. Pedestrian action anticipation using contextual feature fusion in stacked rnns. *arXiv preprint arXiv:2005.06582* (2020).

[29] Schörkhuber, D., Pröll, M., and Gelautz, M. Feature selection and multi-task learning for pedestrian crossing prediction. In *2022 16th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)* (2022), pp. 439–444.

[30] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems 28* (2015).

[31] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[32] Smiley, L. 'i'm the operator': The aftermath of a self-driving tragedy, Mar 2022.

[33] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 4489–4497.

[34] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems* (2017), I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc.

[35] Wang, Z., and Bovik, A. A universal image quality index. *IEEE Signal Processing Letters 9*, 3 (2002), 81–84.

[36] Wojke, N., Bewley, A., and Paulus, D. Simple online and realtime tracking with a deep association metric. *2017 IEEE International Conference on Image Processing (ICIP)* (2017), 3645–3649.

[37] Xue, H., Huynh, D. Q., and Reynolds, M. Poppl: Pedestrian trajectory prediction by lstm with automatic route class clustering. *IEEE transactions on neural networks and learning systems 32*, 1 (2020), 77–90.

[38] Yang, C.-Y., Ma, C., and Yang, M.-H. Single-image superresolution: A benchmark. In *European conference on computer vision* (2014), Springer, pp. 372–386.

[39] Yang, D., Zhang, H., Yurtsever, E., Redmill, K. A., and Özgüner, Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles 7*, 2 (2022), 221–230.

[40] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5728–5739.

[41] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. Learning enriched features for real image restoration and enhancement. In *European Conference on Computer Vision* (2020), Springer, pp. 492–511.

[42] Zhang, X., Angeloudis, P., and Demiris, Y. St crossingpose: A spatial-temporal graph convolutional network for skeleton-based pedestrian crossing intention prediction. *IEEE Transactions on Intelligent Transportation Systems 23*, 11 (2022), 20773–20782.

[43] Zhang, Y. World health organization releases global status report on road safety 2018. *Chin. Disaster Relief Med 7* (2019), 100.

[44] Zhao, J., Xu, H., Wu, J., Zheng, Y., and Liu, H. Trajectory tracking and prediction of pedestrian's crossing intention using roadside lidar. *IET Intelligent Transport Systems 13*, 5 (2019), 789–795.