Pattern Recognition 161 (2025) 111205

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Predicting pedestrian intentions with multimodal IntentFormer: A Co-learning approach

Neha Sharma, Chhavi Dhiman *, Sreedevi Indu

Department of Electronics and Communication Engineering, Delhi Technological University, India

ARTICLE INFO

Keywords:
Pedestrians intention prediction
Co-learning
Weight sharing intent transformer
Autonomous driving

ABSTRACT

The prediction of pedestrian crossing intention is a crucial task in the context of autonomous driving to ensure traffic safety and reduce the risk of accidents without human intervention. Nevertheless, the complexity of pedestrian behaviour, which is influenced by numerous contextual factors in conjunction with visual appearance cues and past trajectory, poses a significant challenge. Several state-of-the-art approaches have recently emerged that incorporate multiple modalities. Nonetheless, the suboptimal modality integration techniques in these approaches fail to capture the intricate intermodal relationships and robustly represent pedestrian-environment interactions in challenging scenarios. To address these issues, a novel Multimodal IntentFormer architecture is presented. It works with three transformer encoders $\{TE_I, TE_{II}, TE_{III}\}$ which learn RGB, segmentation maps, and trajectory paths in a co-learning environment controlled by a Co-learning module. A novel Co-learning Adaptive Composite (CAC) loss function is also proposed, which penalizes different stages of the architecture, regularizes the model, and mitigates the risk of overfitting. Each encoder $\{TE_{II}\}$ applies the concept of the Multi-Head Shared Weight Attention (MHSWA) mechanism while learning three modalities in the proposed co-learning approach. The proposed architecture outperforms existing state-of-the-art approaches on benchmark datasets, PIE and JAAD, with 93 % and 92 % accuracy, respectively. Furthermore, extensive ablation studies demonstrate the efficiency and robustness of the architecture, even under varying Time-to-event (TTE) and observation lengths. The code is available at https://github.com/neha013/IntentFormer

1. Introduction

As of 2021, the global market size for autonomous vehicles(AVs) is estimated to reach $125.67 billion by 2030, according to a recent report published by Market Research Future [1]. The market is expected to witness substantial growth in the coming years, with key players such as Alphabet, Uber, Tesla, and General Motors investing heavily in the development of autonomous vehicle technology. This growth is attributed to the increasing demand for advanced technologies in the automotive industry to increase safety, improve efficiency, and provide comfort and accessibility to all, irrespective of age, disability, or other factors. Furthermore, the unfortunate COVID-19 outbreak that curbed mobility for the risk of infection transmission highlighted the relevance of AVs as they can effectively reduce any human interaction. These could transport anti-epidemic food items, medicines, and other essentials without posing a risk of death due to virus infection.

Nonetheless, the full realization of these benefits is still several years away, as the challenges posed by the complexity of urban traffic environments can impede the reliable operation of AVs. The pedestrian-vehicle conflicts in a crowded urban road environment are one of the most crucial problems that have recently elicited enormous attention from the AV research community. Pedestrians crossing the street are susceptible to vehicle conflicts, leading to safety concerns. Therefore, a comprehensive understanding of pedestrian crossing behaviour cues can assist in interpreting pedestrian intentions and expected crossing actions, leading to improved road safety and fewer conflicts between pedestrians and vehicles.

A combination of visual, dynamic, and motion cues is exhibited by pedestrians when they intend to cross the road, offering valuable clues to their crossing behaviour. For instance, a pedestrian may cross the road if he/she is approaching the crosswalk and looking at the incoming vehicle to ask for a passage. On the other hand, a person standing still at the curb, showing no signs of motion or visual gait towards the crossing action, is less likely to cross the street in a short while. Hence, the pedestrian's positive crossing intent refers to observable behaviour and cues exhibited by a pedestrian, indicating a deliberate intention to cross

* Corresponding author.
E-mail addresses: nehashrm013@gmail.com (N. Sharma), chhavi.dhiman@dtu.ac.in (C. Dhiman), s.indu@dce.ac.in (S. Indu).

https://doi.org/10.1016/j.patcog.2024.111205
Received 7 November 2023; Received in revised form 6 September 2024; Accepted 15 November 2024
Available online 26 November 2024
0031-3203/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

a road or street. This intent is manifested through various actions, such as standing or approaching marked crosswalks, waiting at traffic lights designated for pedestrians, making eye contact with drivers, standing at or approaching zebra crossings, and raising a hand or arm as a signalling gesture to drivers. This kind of behaviour signifies a conscious decision by the pedestrian to engage in the act of crossing, contributing to overall road safety awareness. Contextual factors, including co-pedestrians' behaviour and traffic signals or signs, may further influence the perception of positive crossing intent. Crossing intention confidence is a numeric score estimated from human reference data [2].

Prior research works [3–5] in this area have primarily focussed on the pedestrian's visual appearance and motion features, considering it pivotal in deciphering the underlying intention to cross the road. Furthermore, the environmental context, which includes the presence of different road elements in a traffic scene surrounding pedestrians, has also been proven to influence pedestrian crossing decisions immensely.

In the early days of pedestrian intention anticipation, researchers used dynamical or goal-driven approaches [6,7], which did not scale well to more complex scenarios. With the rise of deep learning, researchers started using deep neural networks (DNNs) to analyse large amounts of data and automatically learn features indicative of pedestrian intention. These models were trained on large datasets of pedestrian behaviour and showed improved accuracy over traditional rule-based methods [2]. Recently, there has been a shift towards more sophisticated models that take into account not only the pedestrian's current behaviour but also the context in which they are operating. This has led to the development of end-to-end models involving convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants that process raw sensory data, such as camera images or lidar point clouds, to make predictions [3]. These approaches are seen as more robust and capable of handling complex scenarios where single-modality approaches may fail, as they can learn the relationships between different modalities and make predictions in a more integrated manner.

Lately, transformers and their variant Vision transformers have proven successful in several natural language processing and computer vision tasks, outperforming traditional RNNs and CNNs in their respective applications. The ability of transformers to process input globally and dynamically evaluate the relevance of different input segments through their attention mechanism has contributed significantly to their success in achieving accurate predictions. This allows transformers to effectively capture long-range dependencies in the data, such as the relationships between objects in an image, which is challenging for CNNs and RNNs together [8,9]. In addition, transformers are well-suited to parallel computation, which makes them more computationally efficient than CNNs and RNNs [10].

1.1. Key objectives

Prior methods [2,11] relying solely on trajectory data or bounding boxes have proven insufficient for accurate pedestrian intention prediction. Our recent work [5] demonstrated that visual features achieved the highest performance, followed by trajectory data when modelling temporal relationships of pedestrian interactions, with pose contributing the least. These findings suggest that integrating RGB images, segmentation maps, and trajectory details capture visual appearance, environmental context, and historical motion patterns to achieve better, more reliable and accurate intention predictions. The effective fusion of these diverse multimodal features necessitates advanced techniques to enhance prediction accuracy [3,4,12,13]. Traditional approaches [12,13], such as feature concatenation, often fail to capture complex intermodal relationships, while multi-stream architectures [3,4] may struggle with intermodal dependencies. In contrast, advanced methods [8,9] like self-attention mechanisms and cross-modal Transformers dynamically select and weigh features, enabling the capture of intricate and dynamic interactions across modalities.

Inspired by these insights, this paper introduces a transformer-based multimodal architecture for pedestrian intention prediction. This model is designed to incorporate and effectively integrate sequential and spatiotemporal information related to pedestrian behaviour, ultimately generating a probability estimate for crossing intention with smaller memory footprint. The following are the salient contributions of this paper:

1) A novel Multimodal IntentFormer architecture is proposed for pedestrian intention prediction leveraging multiple modalities-pedestrian's RGB features, segmentation maps of the surrounding scene, and the trajectory followed in consecutive frames.
2) It learns in three encoder stages: $\{TE_I,\ TE_{II},\ TE_{III}\}$ controlled by a Co-learning module to learn effective inter-modality relationships. A Co-learning Adaptive Composite (CAC) loss is also designed to fine-tune network training and penalize different stages adaptively and efficiently.
3) Each Transformer Encoder (TE_{η}), utilises the proposed Multi Head Shared Weight Attention based modality fusion approach. It learns attention matrices for both modalities simultaneously and generates rich representations.
4) Extensive evaluations demonstrate the architecture's competitive performance, achieving state-of-the-art results on benchmark datasets JAAD and PIE with confident prediction scores even for longer TTE (upto 3.5 secs).

The paper is structured as follows: Section II provides an overview of the current state-of-the-art pedestrian intention prediction and related works, giving context to the proposed framework; Section III delves into the details of the proposed framework, describing the three-tier Transformer-based architecture, co-learning module, the multi-head shared weight attention module, and the custom loss function; Section IV presents a thorough experimental analysis of the proposed framework, highlighting its performance in comparison to existing methods and demonstrating its effectiveness in pedestrian intention prediction tasks. Finally, Section V concludes the paper by summarizing the significant contributions of the proposed work and outlining its potential for future research and development.

2. Related works

This section provides an overview of multimodal approaches for pedestrian intention prediction, focusing on key aspects such as spatio-temporal modelling, multimodal feature integration, and fusion strategies within neural networks. It examines how these methods capture dynamic interactions between pedestrians and their environment, enhance prediction accuracy by combining different data types, and optimize model performance through advanced fusion techniques. Additionally, it discusses binary classification loss functions, highlighting recent innovations that address the limitations of traditional methods.

2.1. Spatio-temporal modelling techniques

Recently, spatiotemporal modelling has been widely used in pedestrian intention prediction, particularly with the development of deep learning models capable of handling spatial and temporal information. Spatio-temporal modelling is a crucial aspect of pedestrian intention prediction as it allows for modelling both the spatial and temporal dimensions of pedestrian behaviour. Spatial modelling refers to the modelling of the physical space in which the pedestrian is operating, including the location and orientation of the pedestrian in the environment. This information is vital for understanding the pedestrian's surroundings, potential obstacles, and interactions with other objects. On the other hand, temporal modelling refers to modelling the time aspect of pedestrian behaviour. This information is essential for

identifying sudden behavioural changes that may indicate an intention to cross the street.

Leveraging the unprecedented success of RNNs and CNNs in several computer vision applications, the last decade has witnessed an increase in their usage in modelling sequential behaviour of pedestrians over time. RNNs help capture their motion patterns by allowing the network to maintain information about the pedestrian's motion over time. CNNs learn to identify significant features, such as the shape and movement of the pedestrian and the fully connected layers, and then use these features to make a prediction. Hamed et al. [12] employed a combination of CNN and Time-Distributed Layers (TDL) to visually represent pedestrians, with the LSTM layer learning the temporal context. Rasouli et al. [14], introduced an RNN encoder-decoder architecture that captures a visual representation of the image surrounding pedestrians, concatenated with pedestrian dynamics. Inspired by this, Yao et al. [15] utilized an encoder-decoder architecture and a novel Attention Relation Network (ARN) to induce a spatiotemporal understanding for anticipating pedestrian crossing intentions. Other groundbreaking works [3, 16] integrated a hybrid combination of CNNs and RNNs for spatiotemporal encoding. However, RNNs and CNNs are challenging to train when there is sparse data, which could be the case in most pedestrian datasets. Furthermore, the vanishing gradient issue in RNNs for longer sequences and inefficiency in capturing the global relationship of the pedestrian with scene objects by CNNs make the overall performance of the CNN-RNN-based architectures suffer in the long run [17].

Quite a few approaches [18–22] have also explored Graph Neural Networks (GNNs) to capture the interactions between pedestrians and their environment. These approaches depict each pedestrian as a node in the graph, and edges are added between nodes to model pedestrian relationships. Liu et al. [23] utilised graph convolution to understand the intricate spatiotemporal relationships in a scene, incorporating both pedestrian-centric and location-centric perspectives. Chen et al. [18] advanced this concept further by employing graph autoencoders to comprehend the impact of the surroundings on pedestrian crossing decisions. Zhang et al. [20] integrated Graph Attention Networks(GAT) to Graph Convolutional Networks(GCNs) to strengthen further the ability to model complex social interactions. In another interesting work, Riaz et al. [21] proposed GNN-GRU-based architecture PedGNN that takes a sequence of pedestrian skeletons as input to predict crossing intentions. Ling et al. [22] utilised GCN(Graph Convolutional Network) with spatial, temporal and channel attention to strengthen feature extraction for more accurate and fast prediction. However, GNNs can struggle to generalize to unseen graphs, as they are heavily dependent on the graph structure and node features. This can be a limitation for anticipating pedestrian intention where the graph structure is subject to change over time [24].

To the extent of our knowledge, the examination of Transformers in pedestrian intention prediction is a novel and under-researched area, with only a handful of works that have addressed it [8,9,11,25]. Achaji et al. [11] proposed a Transformer model with bounding boxes as the only required input. However, it relies solely on bounding box information, which fails to capture the road context and may misinterpret movements similar to crossing behaviour. The PIT framework [9] incorporated a sophisticated integration of a temporal fusion block and a self-attention mechanism, enabling the modelling of the dynamic relationships between the pedestrian, ego-vehicle, and environment. This progressive processing of temporal information enables the capture of dynamic interactions between elements in a manner that is more congruent with human-like behaviour. Additionally, Osman et al. [8] introduced a novel adaptive mechanism that dynamically assigns weights to the significance of current and previous frames, utilizing an attention mask within the Transformer, thereby promoting dynamic spatiotemporal modelling. In another seminal work, Zhang et al. [25] captures temporal correlations within pedestrian video sequences using a Transformer module and addresses the uncertainty of complex pedestrian crossing scenes.

Given these advancements in transformer encoders in capturing dynamic and context-aware pedestrian behaviours, the proposed work leverages Transformer encoders in three distinct encoder stages to facilitate comprehensive learning and integrating multimodal information. It models complex interactions and spatiotemporal relationships across various modalities, including RGB features of pedestrians, segmentation maps of the surrounding scene, and trajectories observed across consecutive frames.

2.2. Multimodal features

Recent advancements in predicting pedestrian intent have predominantly relied on trajectory or historical motion data, as demonstrated in numerous studies [2,11,26]. For example, Achaji et al. [11] proposed a Transformer model utilizing bounding boxes as the sole input, while Bai et al. [26] employed FlowNet to capture motion information. However, these approaches exhibit suboptimal performance, rendering trajectory data or bounding box coordinates alone unreliable for accurate crossing intention estimation [2].

Several methods [19,21,22] have attempted to capture the intricate motion patterns of pedestrians by leveraging pose features derived from limb movements. However, these methods rely on off-the-shelf pose prediction models like OpenPose [27], with their performance heavily dependent on the accuracy of these models. The accuracy of pose predictions, in turn, is contingent upon the resolution of the input images; poor image quality can lead to erroneous pose estimations, further degrading the overall predictive accuracy. To address these challenges, subsequent works [3,4,28] have shown that incorporating visual appearance features provide significant cues regarding pedestrian intent and future actions. Building on this, recent pioneering research endeavours [3,5,8,14,26] enhanced contextual understanding by extracting regions around the pedestrian through the expansion of 2D bounding box coordinates to capture local or global context, often graying out the area within the target pedestrian's bounding box. However, semantic maps, which segregate scene objects, offer a more refined contextual understanding than merely using enlarged images, as evidenced by the success of recent seminal works [4,29] that employ segmentation maps for global context.

Inspired by these advancements, the proposed work integrates RGB images, segmentation maps, and trajectory data as primary modalities. This multimodal approach leverages the complementary strengths of each modality: trajectory data provides insights into past movements, RGB images capture critical visual appearance features such as body posture and gaze direction, and segmentation maps deliver essential environmental context by delineating areas such as roads and sidewalks. By integrating these modalities, our approach aims to enhance the accuracy and reliability of pedestrian intent prediction systems, addressing the limitations of previous methods and advancing the field toward more comprehensive solutions.

2.3. Different multimodal fusion strategies

In deep learning architectures, fusion techniques are pivotal for integrating information across multiple modalities to enhance prediction accuracy, thereby influencing the effectiveness of intention prediction tasks. Early feature fusion methods, such as straightforward or weighted concatenation of features before the final classification network, are employed in notable works [12,13]. However, these approaches may not fully capture the complex intermodal relationships essential for optimal performance, potentially limiting the integration of diverse modal information. Several pioneering works [3,4] employ Multi-Stream Architecture, processing each modality separately within network branches and combining their outputs later. While this allows for learning modality-specific representations and weighting each modality's importance in predictions, it may hinder capturing critical intermodal dependencies and interactions.

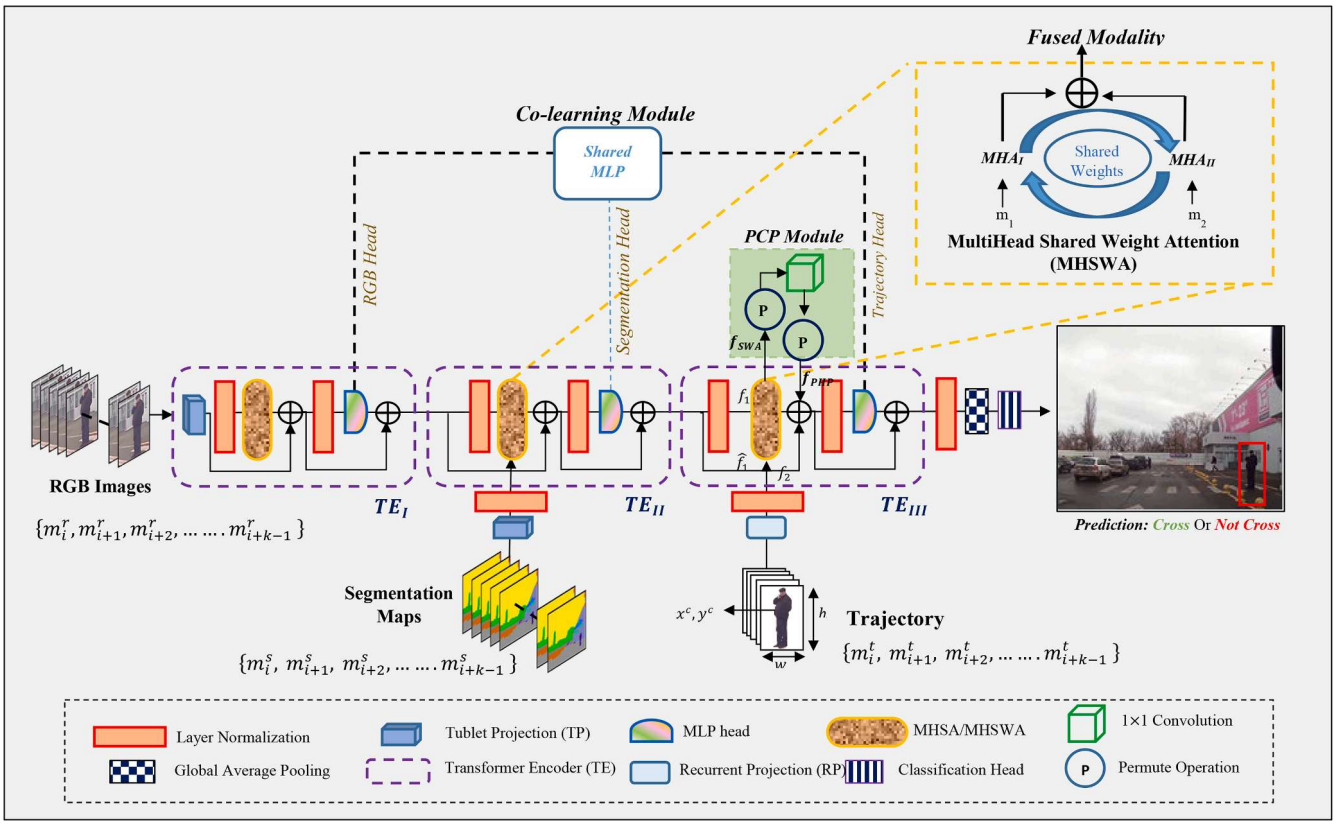


Fig. 1. Illustration of proposed IntentFormer architecture for pedestrian crossing intention prediction. The IntentFormer integrates RGB images capturing temporal variations of appearance, segmentation maps providing global context, and trajectory data representing movement of the pedestrian. There are three transformer encoder stages sequentially process modalities in a Co-learning framework sharing a unified MLP head across layers for efficient feature extraction. The Permutation-Convolution-Permutation (PCP) Module maintains feature patterns while adapting to tensor dimensions. The Multi-Head Shared Weight Attention Module (MHSWA) fosters cohesive modality fusion, enhancing cross-modal relationship leverage for improved intention prediction.

In contrast, advanced fusion techniques like self-attention mechanisms, as seen in noteworthy works [5,26,30], enhance pedestrian intention prediction by emphasizing relevant factors and dynamically selecting multimodal features. For instance, Bai et al. [26] introduces a progressive feature fusion module using a self-attention mechanism to select useful multimodal features from global to local perspectives for pedestrian crossing prediction. Sharma et al. [5] proposes an adaptive fusion module to dynamically weigh all the visual, motion and interaction features, enhancing performance. Additionally, cross-modal Transformer architectures, as explored in another notable study [29], capture dependencies between data types and model interactions between pedestrians and traffic agents, considering both pedestrian and ego-vehicle dynamics.

Despite recent advancements, current methodologies often face challenges in effectively interpreting correlations across different modalities, limiting their generalizability to unseen cases. Addressing these limitations, this paper proposes a novel multi-head shared weight attention (MHSWA) module to facilitate simultaneous learning of attention matrices across heterogeneous modalities. This module enhances the model's capability to integrate diverse sources of information robustly. Furthermore, integrating a co-learning module ensures efficient feature integration and fosters consistency in learned representations across modalities. The MHSWA and co-learning modules represent a significant advancement in multimodal fusion techniques, contributing to a deeper understanding of pedestrian behaviour and more accurate prediction of intentions in dynamic settings.

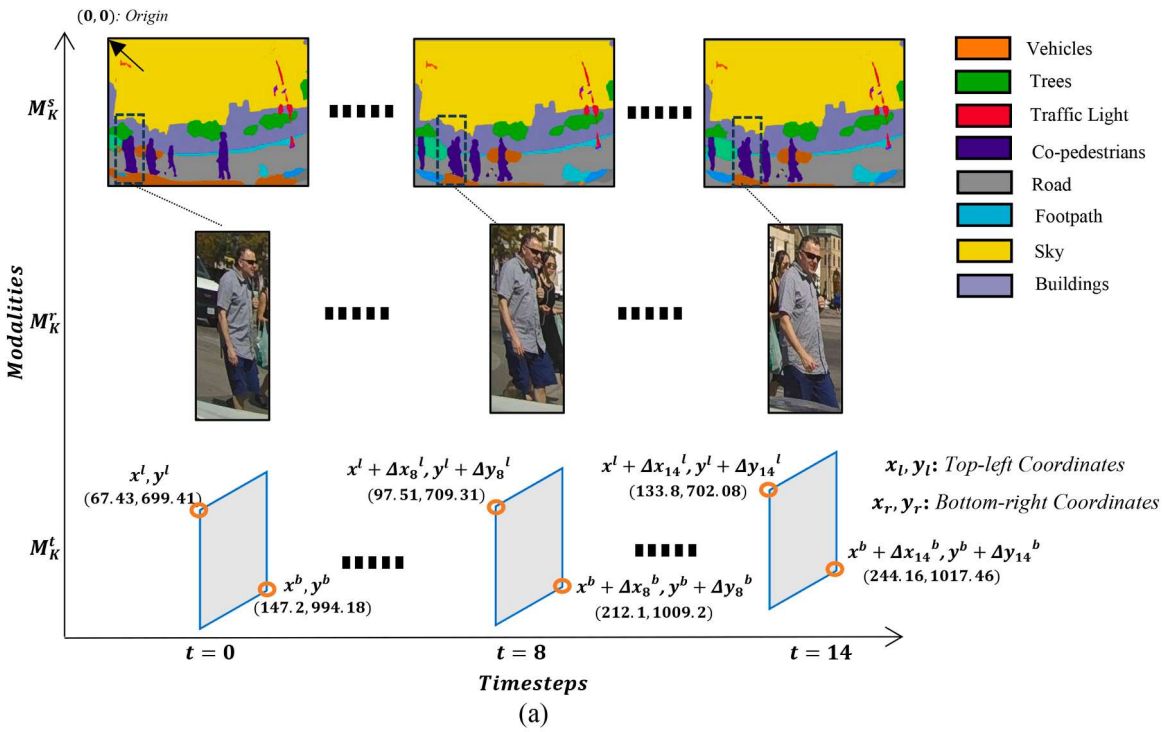
2.4. Binary classification loss functions

The binary cross-entropy (BCE) loss function, while essential for binary classification, faces challenges in complex multi-modal architectures due to its limited capacity to address generalization and overfitting [31–33]. Zhou et al. [32] sought to overcome these limitations by incorporating weighted cross entropy. This represented a significant shift from static loss functions, allowing for more flexible and context-sensitive training processes. Similarly, Lu et al. [31] chose to dynamically adjust the contribution of various loss terms, allowing the loss function to adapt to the specific requirements of each training stage. This approach enhanced the training dynamics, thereby improving the model's robustness and generalization capabilities.

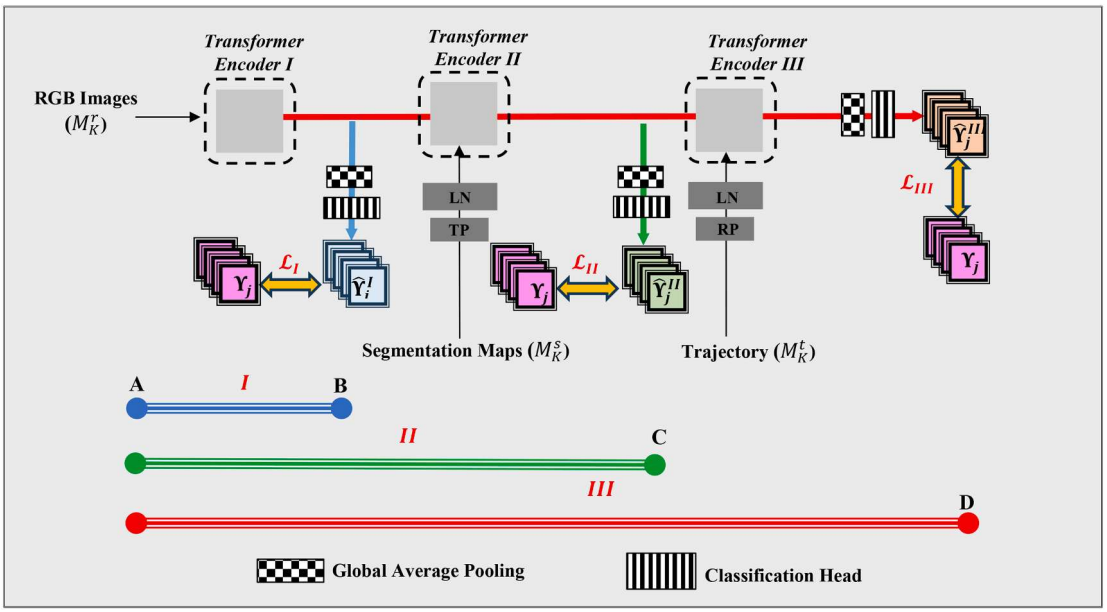
Building upon these insights, the Co-learning Adaptive Composite (CAC) loss function introduced in this work, represents a novel approach to addressing BCE's limitations. The CAC loss function integrates adaptive summation for different network stages—RGB head, segmentation head, and trajectory head. By applying dynamic weights $(\lambda,\ \mu,\ \nu)$ to each stage's loss, this function enables the network to adjust based on performance during training. This adaptive approach not only improves generalization and mitigates overfitting but also optimizes the model's ability to handle multi-modal inputs effectively, thereby enhancing overall training dynamics and performance.

3. Methodology

Predicting pedestrian crossing intention is a challenging task that has significant implications for pedestrian safety and the development of advanced driver assistance systems. In this work, a brief window of 'K' timesteps is analysed from the perspective of the ego vehicle, considering the RGB frames and trajectory coordinates of the pedestrian. The objective is to accurately ascertain the probability $\rho \in (0,1)$ of the pedestrian's intention to cross the road and, thus, classify the pedestrian as



(a)



(b)

Fig. 2. (a) Visualization of various input modalities (M_K^r, M_K^s and M_K^t) for a sample input utilised in the model architecture across timesteps that are processed and combined at different stages of the proposed IntentFormer to guide the learning process (b) Co-learning Composite (CAC) Loss Function. The proposed CAC loss introduces adaptive penalties for binary cross entropy loss for distinct architecture stages η – – – RGB ($\mathscr{L}_I$), segmentation ($\mathscr{L}_{II}$), and trajectory heads ($\mathscr{L}_{III}$), optimizing model's performance. The probabilities Y_j represent ground truth, while Y_j^η signifies predicted probabilities for the respective stages η, contributing to the calculation of loss terms $\mathscr{L}_I$, $\mathscr{L}_{II}$, and $\mathscr{L}_{III}$.
*Note: TP: Tubelet Projection, RP: Recurrent Projection, LN: Layer Normalisation.

a crossing "1″ or non-crossing "0″ entity.

To predict pedestrian crossing intention in traffic scenes, it is crucial to leverage a variety of modalities that can provide a comprehensive understanding of the pedestrian's surroundings. Therefore, the proposed approach combines three distinct modalities: RGB images, segmentation maps, and trajectory data.

1. *RGB images* capture the temporal variations of pedestrian appearance using a sequence of images cropped to the bounding box coordinates provided in the dataset. By analysing a sequence of images, changes in the pedestrian's pose, facial expression, and other visual cues can be tracked, which may indicate crossing intention.
2. *Segmentation maps* provide a global context of the traffic scene surrounding the pedestrian. This facilitates the identification of areas

Algorithm 1
PCP Module

Input:
-Tensor X ($batch_{size}$: $\boldsymbol{b}$; *feature*_dim : $\boldsymbol{N}$; *embed_dim* : $\boldsymbol{M}$**)**
Hyperparameters:
- N°, Convolution filter: 1 × 1 kernel
Output:
-Tensor $\boldsymbol{Y_{PCP}}$
Step 1: Permutation (P_1**) Operation**
Perform permutation P_1 on the input tensor X, mathematically given as follows:
$X_{permuted} = P_1(X)$, where $X \in \mathbb{R}^{b\times N\times M}$, $X_{permuted} \in \mathbb{R}^{b\times M\times N}$
Step 2: Convolution filtering
Apply convolution operation with kernel size [1 × 1] using trainable filter weight as $W_{1\times 1}$ on the permuted tensor $X_{permuted}$, obtained from step 1.
$Y_{conv} = Conv(X_{permuted}, W)$, where tensor $Y_{conv} \in \mathbb{R}^{b\times M\times N'}$
Step 3: Permutation (P_2**) Operation**
Perform permutation operation P_2 on the convolved tensor Y_{conv}, obtained from step 2.
$Y_{P_2} = P_2(Y_{conv})$, where final tensor $Y_{P_2} \in \mathbb{R}^{b\times N'\times M}$
return $Y_{PCP} = Y_{P_2}$

that affect the pedestrian's crossing intention by segmenting the scene into distinct regions based on their visual characteristics. SegFormer [34] generates segmentation maps of the scene that encode different pixel regions in the road scene, including buildings, roads, vehicles, and pedestrians, where each region is assigned a distinct label. SegFormer is pre-trained using the ADE20k dataset with 150 distinct classes, enabling effective segmentation of various road scene elements. Visualisations are also provided in Fig. 2(a) for a better understanding of segmentation maps.

3. *Trajectory* provides the pedestrian's location in a 2D coordinate space, denoted by top-left (x^l, y^l) and bottom-right (x^b, y^b) pixel coordinates, enabling the tracking of their movement and predicting their future paths. Each coordinate is measured in the image frame with reference to the origin corner. Any amount of change in the top-left corner and bottom-right corner coordinates are measured as $(\Delta x_k{}^l, \Delta y_k{}^l)$, and $(\Delta x_k{}^b, \Delta y_k{}^b)$, at k^{th} timestep. The coordinates at the new time step k' are given by $(x^l + \Delta x_{k'}{}^l, y^l + \Delta y_{k'}{}^l)$and $(x^b + \Delta x_{k'}{}^b, y^b + \Delta y_{k'}{}^b)$. Fig. 2(a) illustrates the trajectory coordinates of a pedestrian in a sample trajectory.

Together, these modalities provide a comprehensive representation of the pedestrian and their surroundings, enabling the proposed architecture to accurately predict their crossing intention and ultimately enhance pedestrian safety in traffic scenes. The visualisation of the above-mentioned modalities for input pedestrian samples is provided in Fig. 2(a). The mathematical representation of the modalities is as follows:

$$M_K^r = \{m_i^r, m_{i+1}^r, m_{i+2}^r, \ldots\ldots m_{i+k-1}^r\} \quad (1)$$

$$M_K^s = \{m_i^s, m_{i+1}^s, m_{i+2}^s, \ldots\ldots m_{i+k-1}^s\} \quad (2)$$

$$M_K^t = \{m_i^t, m_{i+1}^t, m_{i+2}^t, \ldots\ldots m_{i+k-1}^t\} \quad (3)$$

where M_K^r, M_K^s and M_K^t are RGB images, segmentation maps and trajectory data for a total of $'K'$ consecutive frames, respectively. Each modality is taken from i^{th} index to $i+k-1^{th}$ frames where $'i'$ is the starting index number.

The architecture of the proposed Multimodal IntentFormer is illustrated in Fig. 1. The proposed architecture harnesses the power of three transformer encoder stages $\{TE_I, TE_{II}, TE_{III}\}$ in order to process a heterogeneous array of input modalities. The inputs are diligently sequentially fed to the encoder stages, conforming to the order in which they are presented. Notably, each encoder stage is endowed with Projection, Layer Normalization, Multi-head Attention (MHA), Multi-head Shared Weights Attention(MHSWA), and Multi-layer Perceptron layers that operate seamlessly in tandem to process the corresponding modality as represented in the Eqs. (4)-(17) as follows:

***TE_I*:**

$$\mathrm{PE}^{rgb} = Positional\_Encoder(Conv3d(M_K^r)) \quad (4)$$

$$\mathrm{Att}^{rgb} = MHA(LN(\mathrm{PE}^{rgb})) + \mathrm{PE}^{rgb} \quad (5)$$

$$Features^I = MLP_{shared}(LN(\mathrm{Att}^{rgb})) + \mathrm{Att}^{rgb} \quad (6)$$

***TE_{II}*:**

$$\mathrm{PE}^{seg} = Positional\_Encoder(Conv3d(M_K^s)) \quad (7)$$

$$LN^{seg} = Layer\_Normalization(\mathrm{PE}^{seg}) \quad (8)$$

$$LN^I = Layer\_Normalization(Features^I) \quad (9)$$

$$Att^{seg,\ I} = MHSWA(LN^{seg}, LN^I) + LN^{seg} \quad (10)$$

$$Features^{II} = MLP_{shared}(LN(Att^{seg,\ I})) + Att^{seg,\ I} \quad (11)$$

***TE_{III}*:**

$$\mathrm{PE}^{traj} = Positional\_Encoder(GRU(M_K^t)) \quad (12)$$

$$LN^{traj} = Layer\_Normalization(\mathrm{PE}^{traj}) \quad (13)$$

$$LN^{II} = Layer\_Normalization(Features^{II}) \quad (14)$$

$$Att^{traj,\ II} = PCP(MHSWA(LN^{traj}, LN^{II})) + LN^{traj} \quad (15)$$

$$Features^{III} = MLP_{shared}(LN(Att^{traj,\ II})) + Att^{traj,\ II} \quad (16)$$

Final output,

$$\hat{Y} = sigmoid(GAP(Layer\_Normalization(Features^{III}))) \quad (17)$$

Where PE^{rgb}, PE^{seg}, PE^{traj} represent positional encodings for RGB images, segmentation maps and trajectories, respectively. It is essential to underscore that two distinct types of projections are leveraged in this architecture: Tubelet projections $Conv3d(M_K^r)$ and recurrent projections $GRU(M_K^t)$. Firstly, the tubelet projections (TP) [35] given by $Conv3d(M_K^r)$ are deployed to assimilate both RGB pedestrian crops and segmentation maps, as utilised in Eq. (4) and Eq. (7). Secondly, recurrent projections (RP) given as $GRU(M_K^t)$, serve as a pivotal tool in processing complex trajectory data, as shown in Eq. (12). $Features^I$, $Features^{II}$ and $Features^{III}$ represent the output feature vectors coming from the transformer encoder stages: TE_I, TE_{II} and TE_{III}. The objective of the Projection layer is to transform the input data into a latent representation space. The Layer Normalization layer is utilized to normalize the activations of the neurons in each layer, thereby facilitating the optimization process.

Figure 3. Diverse Data Augmentations applied to sample pedestrian crops from the dataset (a) Original image (b) Rotation by an angle ± 15 (c) Horizontal Flip (d) Gaussian blur with a kernel value 0.9 (e) Addition of 50, (f) Subtraction of 50, and (g) Multiplication by 2 to pixel intensities.

The mid-level fusion of different modalities commences with the second stage encoder inspired by [36]. The proposed architecture builds on it by employing a novel shared weight attention mechanism for cohesive learning of parameters. The following section explores the technical intricacies of the IntentFormer, shedding light on the PCP (Permutation Convolution Permutation) module, Shared MLP (Multi-Layer Perceptron) layers, and Multi-Head Shared Weight Attention Module (MHSWA), and expounds on their functions and workings:

1. *Co-learning Module*: It enables the integration of different modalities, i.e. RGB images, segmentation maps and trajectory in a unified framework, as illustrated in Fig. 1. This module is designed to share the MLP head across different layers, which helps to reduce the complexity of the framework while preserving the cross-modality relationships. In practice, this means that the module can simultaneously learn to map the input features to the correct pedestrian class using different modalities. It ensures that the learned representations are consistent across modalities, thus producing multi-modality enriched models for predicting pedestrian crossing intention.
2. *Permutation-Convolution-Permutation (PCP) Module*: The PCP module, as shown in Fig. 1, facilitates the establishment of skip connections between two transformer layers despite the different dimensions of the output tensors: $f_{SWA(1,2)}$, and f_2. It performs a sequence of permutation operations, a 1×1 convolution operation followed by a permutation operation again. This sequence of operations ensures that the pattern of features stays unaltered without any parameter overhead, as observed when reshaping after dense operation. The steps of the algorithm are provided in Algorithm 1.
3. *Multi-Head Shared Weight Attention Module (MHSWA)*: The proposed multi-head shared weight attention (MHSWA) module enables the simultaneous learning of attention matrices for heterogeneous modalities, fostering a more cohesive approach to modality fusion, as shown in Fig. 1. This module uses multiple instances of the same multi-head attention layer for different modalities, eliminating the need for separate attention layers and promoting efficient parameter usage. It employs key, query, and value matrices, which are computed by linearly projecting the inputs for each modality. These matrices are used to compute the attention weights for each modality. When multiple instances of the same multi-head attention layer are called for different modalities, the weights for each modality are adjusted simultaneously in the shared weight attention mechanism.

4. Experimental results

The proposed method is evaluated using two commonly used benchmark datasets, JAAD [37] and PIE [14]. The JAAD dataset consists of 346 high-resolution video clips depicting various driving scenarios in an urban setting, with pedestrians performing activities such as crossing the road, walking along the road, and waiting on the side. The dataset is split into two subsets, JAAD_{all} and JAAD_{beh}, with the former containing 2100 visible pedestrians who are not crossing or near the end, and the latter comprising 495 crossings and 191 non-crossings. The PIE dataset offers a more extensive pedestrian data collection than JAAD, with 1, 842 sections of the roadside annotated across different street structures and population densities. The dataset includes 1,842 behaviourally annotated pedestrians, with 519 crossings and 1323 non-crossings, as well as ego-vehicle speed annotations. Both datasets follow the same recommended training/validation/test split configuration for a thorough evaluation [14,38]. Standard classification metrics such as Accuracy, AUC, F1 score, Precision, and Recall are employed to assess the proposed method's performance.

4.1. Implementation details

Loss Function: The most commonly used loss function for binary classification is the binary cross-entropy loss that measures the difference between predicted and true probability distributions. To achieve the goal of fine-tuning the training process and optimizing the model's performance in case of multiple modalities, this work presents a Co-learning Adaptive Composite (CAC) loss function to penalize different stages of the network's architecture, where $'\eta'$ denotes the stages of the architecture, namely RGB head, segmentation head and trajectory head, as described in Fig. 2(b). Y_j and $\hat{Y}_j^{\eta}$ represents the ground truth values and predicted probabilities at stage $'\eta'$ respectively for $'j^{th'}$ pedestrian sample. The loss computations for the stages *I*, *II* and *III* follow a path $A \rightarrow B$, $A \rightarrow C$ and $A \rightarrow D$, respectively, as depicted in Fig. 2(b). The final loss function is an adaptive summation of individual binary cross entropy loss terms calculated from various stages of the architecture for a total of $'m'$ samples in the dataset, as represented in Eqs. (18) and (19) as

follows:

$$\mathscr{L}^{\eta}_{BCE} = -\sum_{j=1}^{m} Y_j \log\left(\widehat{Y}^{\eta}_j\right) + (1 - Y_j)\log\left(1 - \widehat{Y}^{\eta}_j\right) \tag{18}$$

$$\mathscr{L}_{final} = \lambda\mathscr{L}^{I} + \mu\mathscr{L}^{II} + \nu\mathscr{L}^{III} \tag{19}$$

Here, λ, μ and ν represent adaptive weights assigned for each stage, which allow the network to dynamically adjust the composite loss based on the performance of the corresponding stage during the training process.

Data Augmentation: Numerous pixel and geometric transformation techniques have been implemented to augment pedestrian crops to counteract overfitting. Fig. 3 showcases several data augmentation techniques applied to a subset of pedestrian crops from the dataset, including rotation by an angle of $\pm\theta$, horizontal flip, Gaussian blur, with a kernel σ, addition/subtraction by $\in$, and multiplication by a δ to pixel intensities.

Training Specifications: The proposed architecture is trained on a Google Colab Pro instance with access to a high-performance NVIDIA Tesla T4 GPU equipped with 16 GB of memory, running on the CUDA 12.0 platform. The model architecture is built using the TensorFlow 2.10.1 framework. The training regimen involves executing 28 *epochs* and utilizing a batch size of 2 in conjunction with a tuning phase incorporating the $L2$ regularizer with a regularization factor of $1e^{-6}$. The ADAM optimizer is employed in these experiments, with learning rates $1e^{-4}$and $1e^{-5}$ for the PIE and JAAD datasets, respectively, that decay by 0.1 every 10 *epochs*. Early stopping callback is also employed to prevent overfitting by monitoring validation loss improvement and halting the training if no improvement is observed for the next 7 *epochs*. The benchmark protocol is followed to address the dataset imbalance, which involves adding flipped versions of underrepresented sequences and subsampling from the overrepresented samples to balance the number of samples [3].

The computation of segmentation maps using Segformer [30] for the whole dataset has been executed before training. As reported in Table 7, Section 4.2.7, each transformer block is configured to include 4 heads, a projection dimension of 64, and a shared MLP head consisting of 64 $\times$4 and 64 MLP heads. The patch size for inputting RGB and segmentation maps is set to $(2,8,8)$. The Tubelet Projection (TP), implemented as a 3D convolutional layer, efficiently extracts features by aligning the number of filters with the projection dimension, using a kernel size matching the specified patch size, and employing strides and padding configurations. The Recurrent Projection (RP), realized through a GRU layer with the number of hidden units equivalent to the projection dimension, is crucial in capturing temporal dependencies and patterns within the input data. The MLP layers are initialized using the HeNormal initializer, including a 50 % dropout rate between layers to mitigate overfitting. The entire experiment is initialized with a random seed to ensure the reproducibility of results. Through empirical analysis, it has been determined that an observation length of 0.5 seconds and a time-to-event of 2.5 seconds represents an optimal configuration, as shown in the ablation study, Section 4.2.1. Thus, the IntentFormer is trained with the number of observation frames fixed at 15, i.e. 0.5-second observation length at a frame rate of 30fps.

Prior Knowledge: In the context of pedestrian crossing intention detection, two fundamental temporal parameters shape the foundation of predictive systems: observation length and time-to-event (TTE). These parameters intricately influence the accuracy and responsiveness of intention predictions by determining the historical context and temporal proximity to the crossing event. Understanding their roles is pivotal for designing efficient and contextually aware systems that enhance pedestrian safety and optimize interactions with autonomous technologies.

Table 1
Evaluation of the Proposed Architecture in Comparison to Other Methods on the PIE Dataset

Methods	Year	PIE				
		Acc	AUC	F1	Prec	Rec
PIE_traj [14]	2019	0.79	–	0.87	–	–
SF-GRU [3]	2020	0.87	0.85	0.78	0.74	0.64
PCPA [28]	2021	0.87	0.86	0.77	–	–
TED [11]	2021	0.91	0.91	0.83	–	–
PG+ [19]	2022	0.89	0.90	0.81	0.83	0.79
TAMFORMER [8]	2022	0.87	0.84	0.76	–	–
V-PedCross [26]	2022	0.89	0.88	0.67	0.74	0.84
MFFN [30]	2023	0.88	0.89	0.81	0.79	–
PedGNN [21]	2023	0.71	–	0.79	0.75	0.83
TrEP [25]	2023	**0.93**	**0.94**	0.87	**0.89**	–
PedFormer [29]	2023	**0.93**	0.90	0.87	**0.89**	–
VMI [5]	2023	0.92	0.91	0.87	0.86	0.88
IntentFormer(*Ours*)	–	**0.93**	0.90	**0.88**	0.86	**0.89**

Table 2
Evaluation of the Proposed Architecture in Comparison to Other Methods on the $JAAD_{beh}$ Dataset

Methods	Year	$JAAD_{beh}$				
		Acc	AUC	F1	Prec	Rec
PCPA [28]	2021	0.58	0.5	0.71	–	–
FFSTA [4]	2022	0.62	0.54	0.74	0.65	0.85
PG+ [19]	2022	0.70	**0.70**	0.76	**0.77**	0.75
TAMFORMER [8]	2022	0.73	0.70	0.79	–	–
V-PedCross [26]	2022	0.64	0.66	0.76	0.70	0.89
STMA-GCN PedCross [22]	2023	0.69	0.58	0.80	0.68	**0.97**
IntentFormer(*Ours*)	–	**0.75**	**0.70**	**0.82**	0.74	0.88

i) **Observation Length**: Observation length refers to the number of consecutive time steps for which historical pedestrian data is considered during the training process of a pedestrian crossing intention detection system. In other words, it is the duration of the past behaviour and cues that are taken into account for making predictions about a pedestrian's intention to cross the road.
ii) **Time-to-Event (TTE):** Time-to-event (TTE) is the temporal difference between the last time step of the observation length and the occurrence of the actual crossing event. It quantifies the interval from when the system last observes the pedestrian's behaviour to when the pedestrian starts crossing the road.

The observation length and TTE are interconnected parameters crucial for designing effective pedestrian crossing intention detection systems. Striking the right balance between these factors is essential to ensure timely, accurate, and context-aware predictions, contributing to enhanced pedestrian safety and smoother interactions between pedestrians and autonomous systems.

Comparison with State-of-the-art Methods: The proposed architecture is evaluated against state-of-the-art methods as follows: PIE_traj [14], SF-GRU [3], PCPA [28], TED [11], PG+ [19], TAMFORMER [8], V-PedCross [26], MFFN [30], PedGNN [21], TrEP [25], PedFormer [29], FFSTA [4], STMA-GCN PedCross [22] and VMI [5]. Tables 1 and 3 illustrate that the proposed architecture, IntentFormer, achieves performance levels comparable to PedFormer [29] and TrEP [25]. This can be primarily attributed to the integration of the Transformer encoder, a fundamental architectural component common to all these methods. Nonetheless, IntentFormer outperforms these methodologies [25,29] on the $JAAD_{all}$ dataset, with a substantial improvement ranging from 14 % to 54 % in AUC, F1 score, precision, and recall. Moreover, while prior methodologies [19,25,29] typically confine time-to-event (TTE) predictions to 1–2 seconds, IntentFormer attains superior results with the highest reported TTE of 2.5 seconds. On the $JAAD_{all}$ dataset, PedGNN [21] achieves the highest recall of 0.96; however, our proposed method

Table 3
Evaluation of the Proposed Architecture in Comparison to Other Methods on the $JAAD_{all}$ Dataset

Methods	Year	$JAAD_{all}$				
		Acc	AUC	F1	Prec	Rec
SF-GRU [3]	2020	0.84	0.80	0.62	0.54	0.73
PCPA [28]	2021	0.85	0.86	0.68	–	–
FFSTA [4]	2022	0.83	0.82	0.63	0.51	0.81
PG+ [19]	2022	0.86	0.88	0.65	0.58	0.75
TAMFORMER [8]	2022	0.89	0.82	0.68	–	–
V-PedCross [26]	2022	0.86	0.81	0.77	0.74	0.81
MFFN [30]	2023	0.91	**0.90**	0.81	0.80	–
PedGNN [21]	2023	0.86	**–**	**0.86**	0.77	**0.96**
TrEP [25]	2023	0.91	0.86	0.69	0.71	–
PedFormer [29]	2023	**0.93**	0.76	0.54	0.65	–
VMI [5]	2023	0.89	**0.90**	0.81	0.79	0.83
IntentFormer(*Ours*)	–	0.92	**0.90**	0.83	**0.81**	0.85

outperforms with superior accuracy and precision of 0.92 and 0.81, respectively, compared to 0.86 and 0.77 of PedGNN [21]. Furthermore, Table 2 demonstrates that IntentFormer exhibits the highest performance among methods evaluated on the $JAAD_{beh}$ dataset.

These findings indicate enhanced generalizability of the proposed IntentFormer across diverse datasets. This is attributed to an enriched understanding of pedestrian intentions facilitated by co-learning-induced shared training of the MLP layer. Incorporating Co-learning Adaptive Composite (CAC) loss has contributed to the model's generalizability by providing regularization. Moreover, deploying the Multi-Head Shared Weight Attention (MHSWA) module has effectively modelled intermodal relationships, further bolstering the model's superior performance.

4.2. Ablation study

This section presents an ablation study of a proposed work for pedestrian intention prediction employing quantitative and qualitative approaches to investigate different aspects of the architecture. These include evaluating the impact of observation length and time-to-event information on prediction performance, assessing various modality fusion approaches, analysing the effectiveness of a custom loss function for binary classification, and comparing the performance of the proposed transformer model with shared MLP head to the standard vanilla transformer model.

4.2.1. Effect of time-to-event (TTE) and observation length

The influence of time-to-event (TTE) and observation length on predictive performance is examined by considering various TTE points and observation lengths along the timeline of the crossing event. TTE points, ranging from 0 to 4 seconds, are sampled at intervals of 0.5 seconds, while observation lengths from 0 to 2 seconds are taken at intervals of 0.25 seconds, as depicted in Fig. 4.

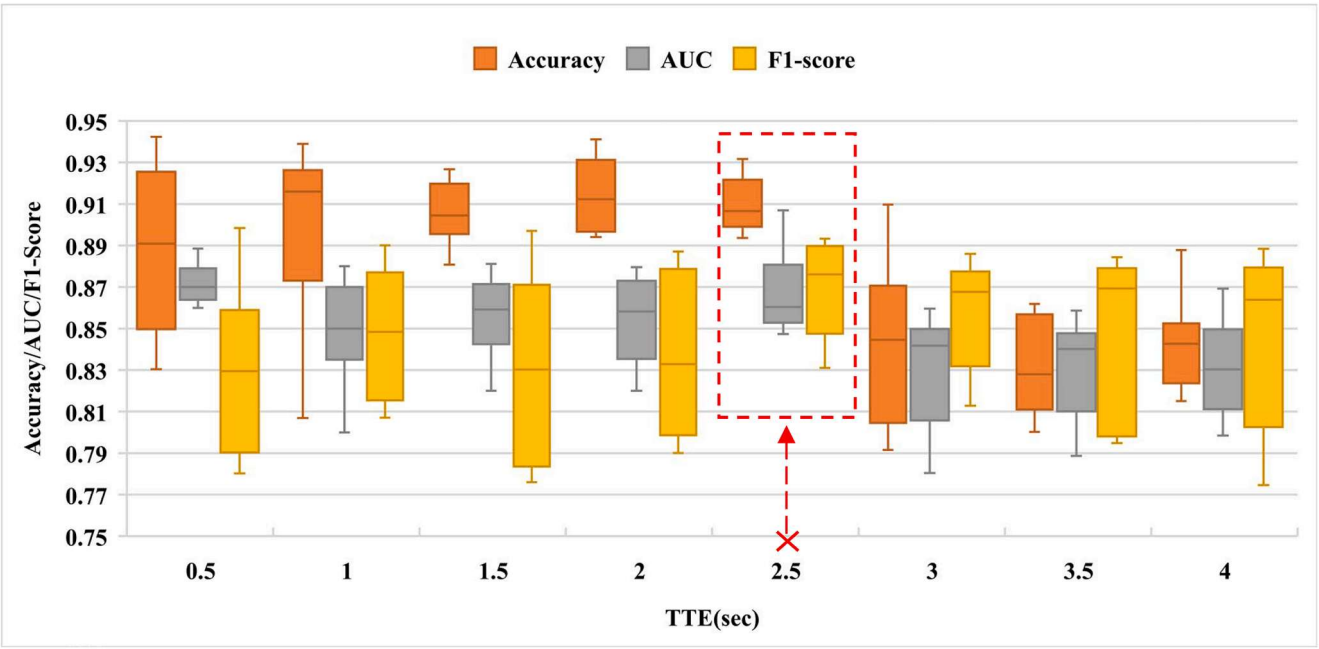
TTE=0 represents the time of the crossing event. Performance improves as TTE approaches 0 seconds, indicating increased confidence in predicting crossing events. However, the variability in performance is also high, indicating that the performance at these timesteps does not consistently ensure high accuracy. For an efficient intention prediction model, the prediction confidence score should be high right before the crossing event, i.e., TTE>0. At 2.5 seconds, the statistical measures of accuracy, AUC, and F1 score demonstrate high and relatively stable values with varying observation lengths, as depicted in Fig. 4(a). Beyond 2.5 seconds, there is a notable decline in overall performance, with accuracy decreasing by up to 6.5 %.

Fig. 4(b) demonstrated that the optimal performance is observed within the 0.5–1.25 seconds observation length range, exhibiting minimal variation with changing TTE. The performance metrics peak at an observation length of 0.5 seconds and show minimal fluctuation. Hence, this observation length is ideal for achieving optimal performance, as the accuracy, AUC, and F1 scores remain consistently high within this range. Moreover, accuracy, the area under the curve (AUC), and the F1 score show a modest gain up to an observation length of 1.25 seconds since such a prolonged duration leads to higher information acquisition. However, beyond that, the performance drops as prolonged observation periods may contain irrelevant details about the scene dynamics that can undermine the prediction accuracy. Larger observation lengths signify a more significant number of frames required for analysing crossing intention, resulting in high computational demands. Therefore, an efficient intention prediction model should make confident predictions with the least possible observation length. The proposed model demonstrates robustness by achieving optimal performance with an observation length of just 0.5 seconds, thereby minimizing computational demands and ensuring efficient prediction. These results highlight the efficiency of the proposed architecture in predicting crossing events even with fewer frames and high TTE (upto 3.5 secs), with performance metrics dropping by no more than 12–14 %. This starkly contrasts the SF-GRU [3] method, which exhibited a substantial decline in performance metrics, reaching up to 33 % when TTE is increased beyond 3 seconds. Furthermore, the PG+ [19] approach restricts TTE to 1–2 seconds, limiting its suitability for real-time scenarios. Notably, the proposed approach achieves superior accuracy compared to VMI [5] and comparable metrics, with the highest reported TTE to date while maintaining a significantly reduced computational footprint and inference time, as outlined in Section 4.2.7. *Quantitative Analysis*.

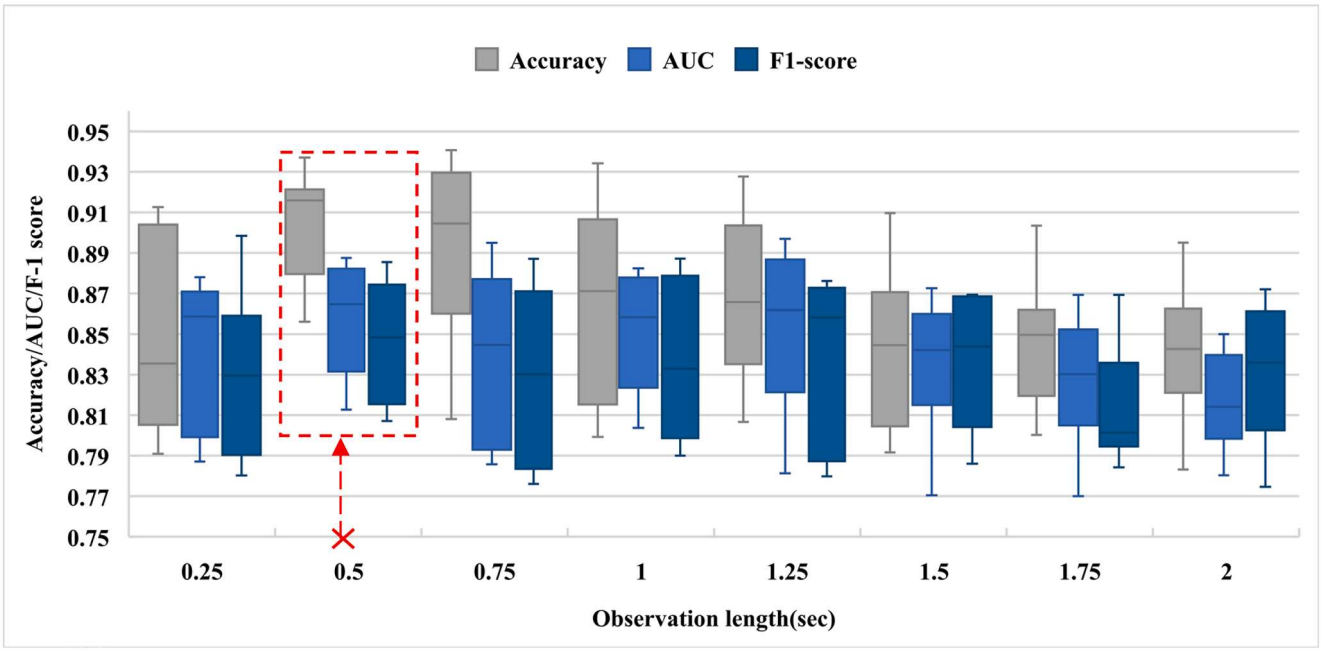
4.2.2. Analysis of modality fusion approaches

In the field of multimodal deep learning, multi-head cross-modal attention (MHCMA) and multi-head multimodal attention (MHMMA) based fusion techniques have emerged as popular mid-level transformer-based approaches [36]. These attention mechanisms have unique characteristics and functionalities that may cater to specific application domains. The proposed model employs a multi-head shared weight attention (MHSWA) mechanism to facilitate the synergistic fusion of information across distinct modalities. The shared weight attribute capitalizes on the synergy of attention weights from various heads to exploit cross-modal correlations efficiently. It comprises two scaled dot product attention instances tailored to specific modalities. The first instance, trained on the initial modality data, captures intricate interdependencies among elements. Subsequently, the second instance, initialized with learned weights from the first modality, refines its training on the subsequent modality, fostering a sequential, contextual understanding enriched by prior knowledge.

The distinct design characteristics of these three attention-based fusion strategies are elucidated in Fig. 5. An ablation study is conducted using the Precision-Recall curve, as depicted in Fig. 6, to assess the impact of the various fusion strategies on performance. The study's results revealed that the MHSWA method's precision-recall curve is notably closer to the ideal curve compared to the other two approaches. The varying behaviour of attention coefficients across the different stages of the proposed shared weight attention model is illustrated in Fig. 7. At stage *I*, Fig. 7(a), a high range of attention coefficients indicates that the model assigns varying levels of importance to different embeddings within the RGB data. This stage focuses on capturing fine-grained details and relationships specific to the RGB input, as it is the primary modality. A slight decrease in the attention coefficient range at this stage *II* is observed in Fig. 7(b), suggesting that the model focuses on commonalities and interactions between RGB and segmentation embeddings. The shared weight attention mechanism allows the model to emphasize cross-modal correlations and jointly process features from both modalities. In the last stage *III*, Fig. 7(c) highlights attention coefficients distinguished by a much shorter range, indicating that the model is assigning more consistent attention across embeddings from different modalities. It is inferred that the model is integrating information from previous stages (RGB and segmentation) and trajectory



(a)



(b)

Fig. 4. Performance of the proposed architecture is evaluated under varying conditions of (a) Time-to-Event (TTE) and (b) Observation Length, sampled at intervals of 0.5 seconds and 0.25 seconds, respectively. Optimal performance is observed at TTE=2.5 seconds and an observation length of 0.5 seconds, highlighted by dashed lines.

more uniformly. The change in attention behaviour from varying ranges to more uniform attention signifies that the model progressively shifts its emphasis from capturing modality-specific details to integrating multi-modal information for decision-making. Hence, the observed behaviour aligns with the objective of multimodal learning: to learn robust representations that capture inter-modal relationships and produce consistent outputs despite the varied nature of the input sources.

In Fig. 8. Guided Integrated Gradient (IG) [39] Visualizations corresponding to individual attention map heads are presented for RGB sequences. It highlights the areas where Multi-head Shared Weights Attention (MHSWA) mechanisms positively influence the model's classification decision. This configuration comprises a total of four discrete attention map heads. The first attention map head primarily emphasizes the outline or shape of the target pedestrian. The second and third attention maps appear to capture details related to the target's immediate surroundings and the pedestrian's dynamic variations across the sequence of frames. The fourth attention map identifies contours and distinct patterns within the cropped image.

4.2.3. CAC vs BCE

This section explores the impact of the proposed Co-learning Adaptive Composite (CAC) loss function on validation performance and the

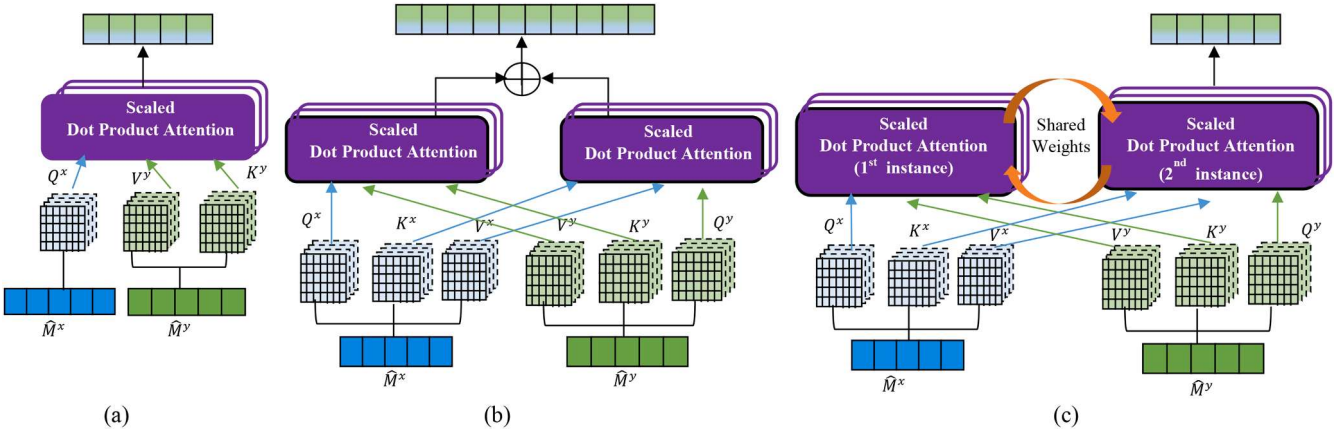


Fig. 5. Illustrates three distinct types of Multi-head attention, namely (a) Cross-Modal Attention (MHCMA) (b) Multimodal Attention (MHMMA), and (c) Shared-weights Attention(MHSWA). MHCMA targets interaction between distinct modalities to enhance alignment and comprehension, while MHMMA concurrently attends to multiple modalities for informed decision-making. The proposed MHSWA, efficiently captures cross-modal correlations while optimizing parameter utilization, with it's distinctive shared weight attention attribute.

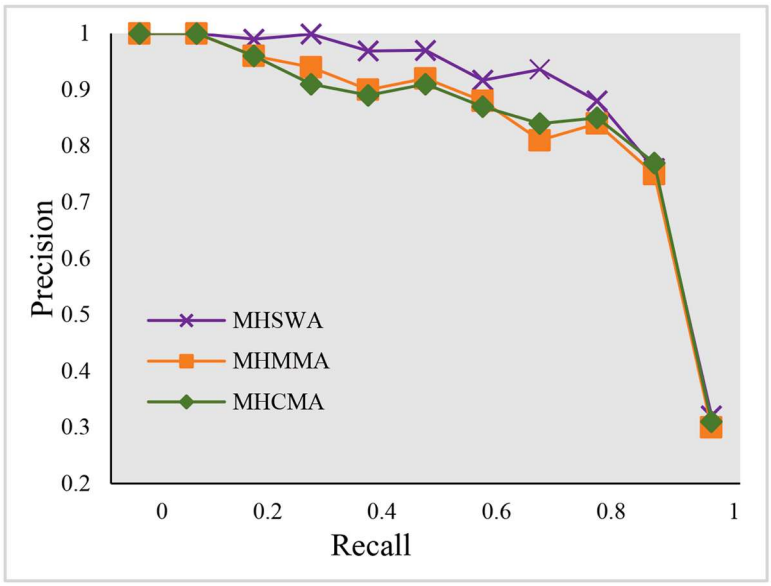


Fig. 6. Precision-Recall curves for different types of modality fusion attention mechanisms: MHCMA, MHMMA and MHSWA

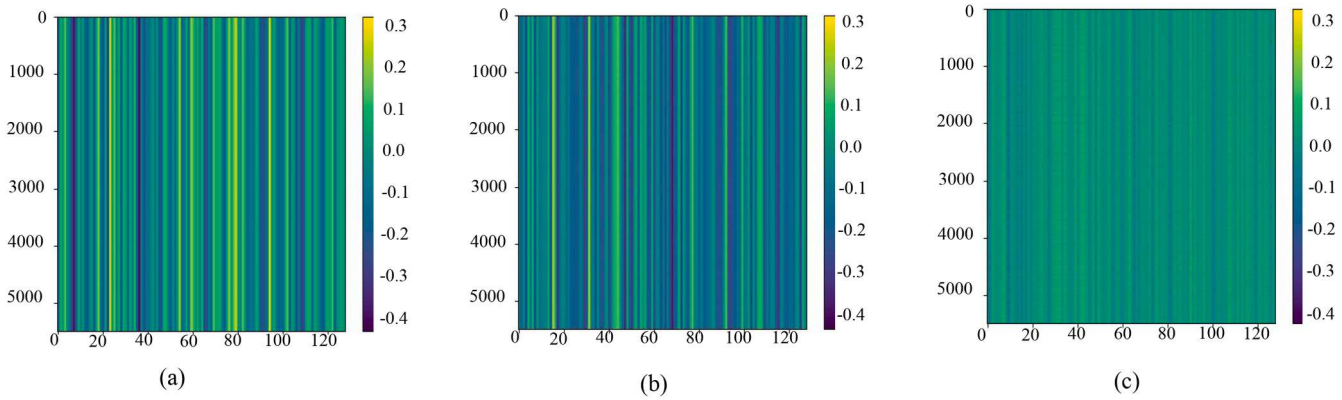


Fig. 7. Evolution of Attention Coefficients across Sequential Stages in the Proposed Shared Weight Attention Model. The attention coefficient distributions for different stages (a) *I*, (b) *II*, and (c) *III* are depicted, showcasing the model's dynamic learning process. In (a), attention coefficients exhibit a wide spectrum, indicating nuanced importance assignments to RGB embeddings. In (b), a refined attention range highlights cross-modal correlations between RGB and segmentation embeddings. (c) reveals a focused attention span, reflecting the harmonious integration of multi-modal insights from previous stages—RGB, segmentation, and trajectory—highlighting the model's progressive shift towards synthesizing diverse modalities for effective decision-making.

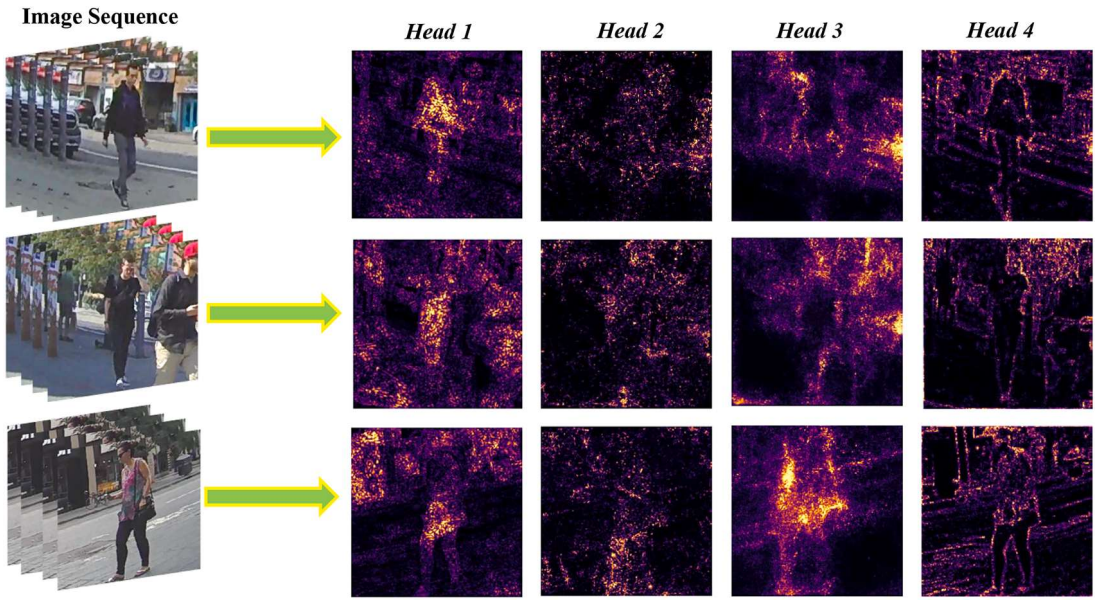


Fig. 8. Guided Integrated Gradient [39] Visualisation of IntentFormer MHSWA maps of sample sequence of RGB images. Each head highlighting diverse elements from the image sequence such as pedestrian contours, contextual details, motion dynamics and intricate patterns enabling a nuanced analysis.

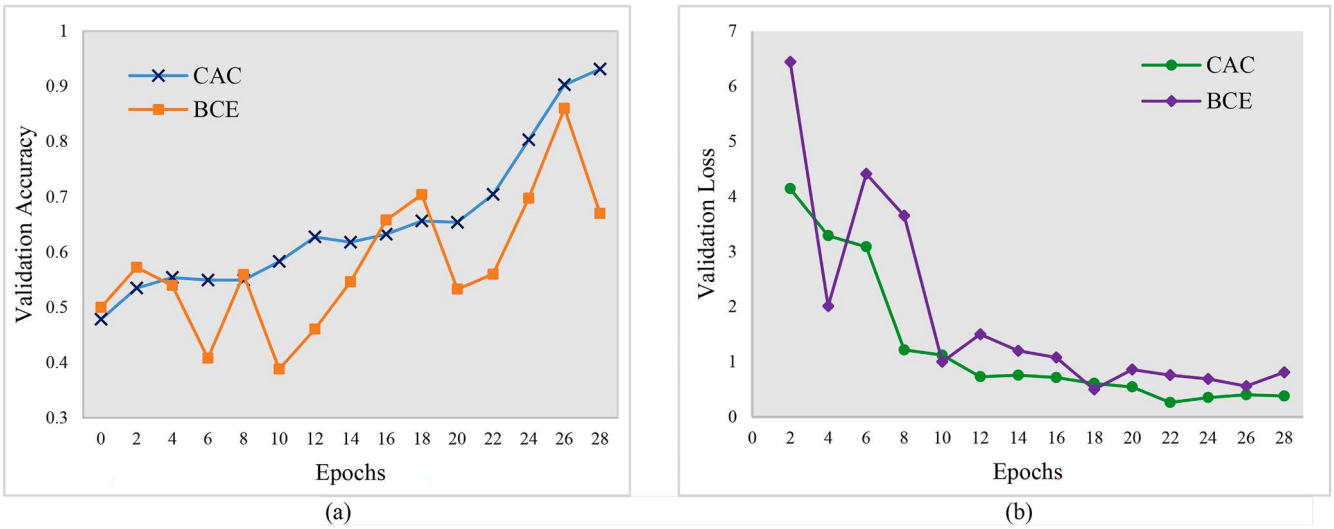


Fig. 9. Effect of Co-learning Adaptive Composite (CAC) loss function and standard BCE function illustrated through (a) Validation Accuracy, and (b) Validation Loss curves.

dynamic relationship between adaptive loss weights and training progress. Fig. 9(a) presents the validation accuracy curves for models trained using the standard Binary Cross-Entropy (BCE) and the proposed CAC loss function. The CAC loss function notably enhances the stability of validation accuracy throughout the training phase, reducing fluctuations compared to BCE and achieving superior validation accuracy. In Fig. 9(b), the validation loss curves show that BCE induces more frequent fluctuations than the CAC loss, leading to difficulties in convergence. In contrast, the CAC loss function achieves the lowest validation loss. These results indicate that the CAC loss function effectively mitigates overfitting during training, thereby enhancing the generalization capacity of the proposed architecture.

Fig. 10 illustrates the evolution of the training and validation loss alongside the changes in adaptive loss weights (w_1, w_2, w_3), throughout training epochs. The adaptive loss weights, initialised randomly, exhibit dynamic adjustments in response to the changing training landscape. Specifically, weights w_2 and w_3, exhibit a gradual and consistent increment throughout epochs culminating at respective maximal values of 0.46 and 0.60. Contrastingly, weight w_1 display more intricate behaviour, initially decreasing, followed by a gradual and consistent increase over epochs, reaching a maximum value of 0.35. This suggests that the w_1 loss term contributes significantly less to the overall loss as the model refines its representations. The vertical line denotes the epoch at which the minimum loss is attained, providing insight into the optimal point (w_1 : 0.29, w_2 : 0.36 and w_3 : 0.52) in the training process. This allows us to fine-tune training strategies and highlights the potential for adaptive loss weighting to enhance model training efficiency and performance.

The binary cross-entropy (BCE) loss function, although essential for binary classification tasks, proves inadequate for complex multi-modal architectures, leading to overfitting, as demonstrated by the ablation study. Recent advancements [31,32] have introduced dynamic and weighted loss functions to improve training stability and generalization. Thus, inspired by these advancements, the proposed work developed the

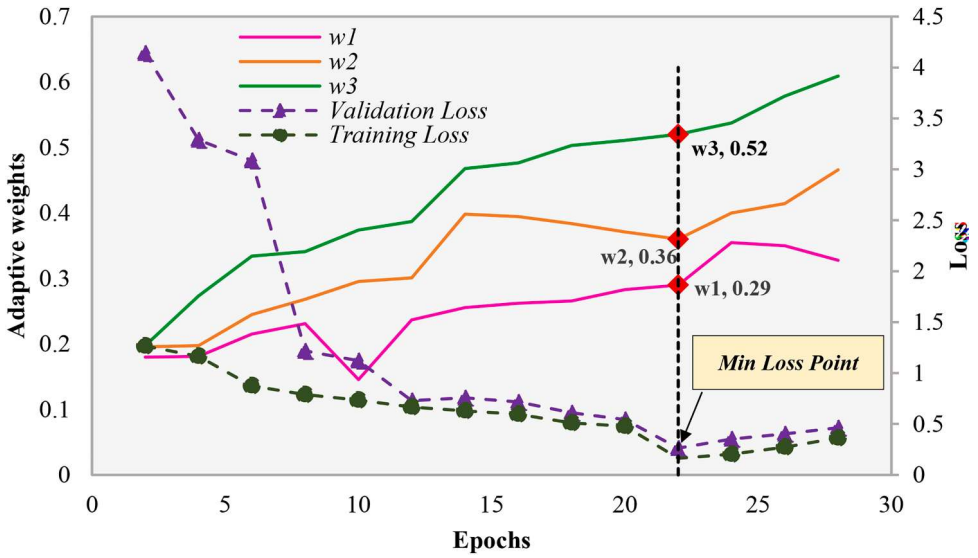


Fig. 10. Adaptive Loss Weights and Training Dynamics. The graph depicts the training and validation loss curves (left Y-axis) along with the evolution of adaptive loss weights (right Y-axis) across epochs. The Training Loss and Validation Loss exhibit a decreasing trend, indicating improved model performance. Concurrently, the adaptive weights ($w1$, $w2$, $w3$) adjust dynamically over epochs, reflecting the model's learned emphasis on different loss components.

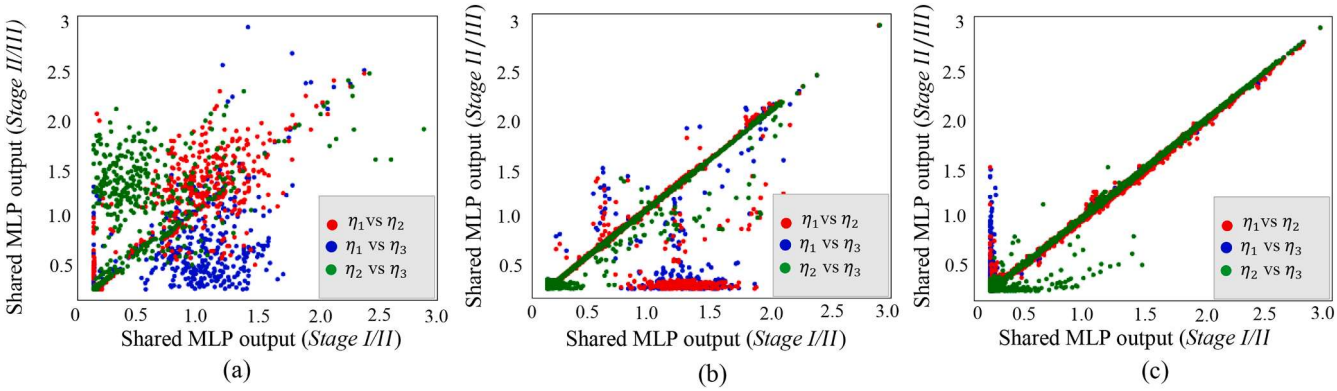


Fig. 11. Learned feature Representations using the proposed shared MLP layer in the Co-learning architecture. Pairwise scatter plots illustrate the alignment of learned representations for three distinct epochs, (a)Epoch -3, (b) Epoch -15, and (c) Epoch – 22 for three shared MLP layer stages (*I*, *II* and *III*) of the architecture. *Stage I*: RGB, *Stage II*: RGB with segmentation, Stage *III:* RGB, Segmentation with trajectory. Each scatter plot compares the shared MLP layer outputs of two stages, revealing how the representations transform and establish coherent relationships over time with epochs. The figures also demonstrate a dynamic convergence of representations as epochs progress. The increasing linear alignment signifies the architecture's capacity to establish a unified encoding that is shared between modalities as training progresses.

Co-learning Adaptive Composite (CAC) loss function. This novel loss function has demonstrated superior performance compared to the standard BCE, significantly reducing accuracy and loss fluctuations during training and leading to more stable training dynamics, as illustrated in Fig. 9.

4.2.4. *Co-learning v/s Vanilla transformer architecture*

This section discusses the adaptive learning process of the proposed co-learning architecture that leverages shared MLP heads. The pairwise scatter plots in Fig. 11 illustrate a notable evolution in the alignment of learned representations across the three epochs (Epochs 3, 15, and 22) for the Co-learning multimodal architecture employing shared MLP heads. As training progresses, the shared MLP layers output at three stages increasingly converges along a linear trajectory. This highlights that the architecture effectively captures the shared semantics across modalities, allowing for improved feature extraction and cross-modal interaction. Furthermore, the dynamic alignment of representations over epochs suggests that the shared MLP layer effectively captures cross-modal relationships, allowing different modalities to learn and adapt coherently.

Furthermore, a comparative analysis of the proposed architecture with a vanilla transformer model without a shared MLP head is also carried out. The term "Vanilla transformer" here denotes a model variant in which the shared MLP in the co-learning architecture is substituted with three independent trainable MLPs, each assigned to a specific modality (RGB, segmentation, and trajectory). This modification facilitates a comparative analysis between the co-learning architecture utilizing shared MLPs and an alternative configuration employing non-shared, individual MLPs for each modality. The goal is to evaluate the influence of shared semantics across modalities on the learning dynamics. Qualitative results for the few samples from the JAAD/PIE dataset are presented in Fig. 12. Each example demonstrates the model's focus across different time steps—specifically, past observed frames on the left and future frames on the right. In the first row of each example, the original image sequence captures the target pedestrian within the context of the surrounding scene. These sequences illustrate the pedestrian's movement trajectory as they approach or cross the street, providing a visual reference for the pedestrian's intent. The second row

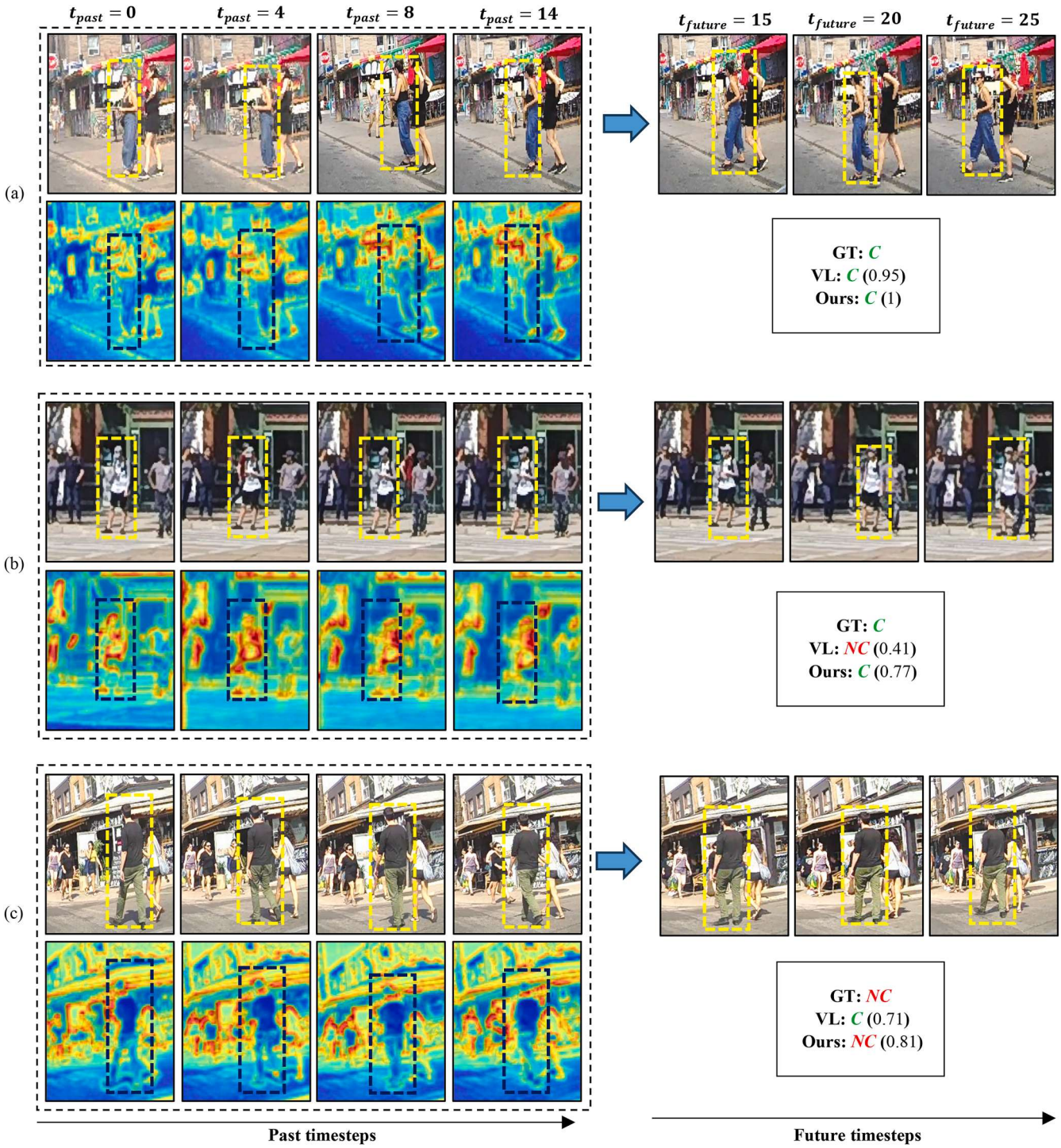


Fig. 12. Qualitative predictions on the PIE/JAAD datasets, where the proposed IntentFormer (**Ours**) accurately classifies pedestrian intentions that are either misclassified or less confidently predicted by the vanilla transformer (**VL**) architecture. ***Red*** indicates detected non-crossing intentions, while ***green*** represents crossing intentions. The left side displays past pedestrian frames, illustrating their trajectories, with corresponding cross-attention maps shown in the subsequent row. These maps highlight the temporal scene information utilized by the model to predict crossing intentions in the future frames shown on the right side. Past timesteps t_{past}; future timesteps t_{future}.

reveals the model's internal focus through cross-attention heatmaps generated at the transformer encoder in Stage II. These heatmaps highlight the regions of the image that the model deems important for predicting the pedestrian's crossing intent. A key insight from these heatmaps is the model's ability to attend to not just the target pedestrian, but also to surrounding co-pedestrians and other critical scene elements, such as co-pedestrians, vehicles, sidewalks, and road markings.

Notably, Fig. 12 (b)-(c) depicts instances of no eye contact between the pedestrian and the camera, resulting in uncertainty regarding the direction in which the pedestrian would move. For instance, Fig. 12 (b) shows a pedestrian looking at a phone, making it difficult for the model to interpret intention from visual appearance cues such as gaze. Similarly, in Fig. 12 (c), despite the pedestrian walking along the sidewalk,

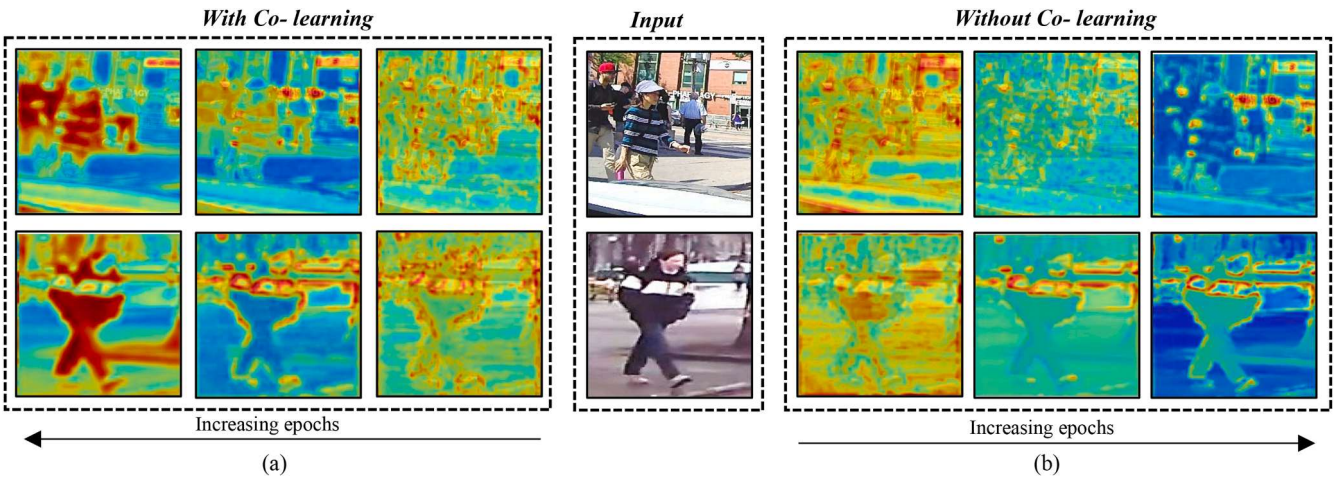


Fig. 13. Visualization of Grad-Cam maps for the proposed IntentFormer trained for 3, 15 and 22 epochs: (a) With the co-learning module (*right to left*), and (b) Without the co-learning module (*left to right*).

their constant motion leads the vanilla transformer architecture to inaccurately predict a crossing intention. These examples underscore the vanilla architecture's limitations in handling difficult classification scenarios. However, the correct predictions by the proposed model in such hard samples of crossing and non-crossing intent can be attributed to the fact that it caters to the cross-modal relationships among visual appearance, segmentation maps and trajectory with consistent learned representations as shown in the heatmaps. Thus, even if one representation fails to capture the pedestrian's intention correctly, its relationship with the other two modalities strives to decipher it correctly, albeit with less confidence.

The Grad-CAM visualizations for the IntentFormer with and without the co-learning module (Vanilla transformer) are depicted in Fig. 13(a) and (b), respectively. Analysis of Fig. 13 (a) reveals a progressive refinement in the Grad-CAM attention maps in the co-learning environment as the number of training epochs increases. Initially, at epoch 3, the Grad-CAM outputs are dispersed across the input image, lacking specific focus on any element. However, as training progresses, the importance weights become increasingly localized to image regions pertinent to classifying the pedestrian's intention. The attention maps become more precise, effectively highlighting the silhouette of the target pedestrian. Additionally, with the incorporation of segmentation maps and trajectory data in the second and third stages, respectively, it is observed that co- pedestrians and certain scene elements, such as road boundaries, also receive higher weightage as observed for models trained for epochs 15 and 22. This indicates an enhanced understanding of the context and contributing factors to pedestrian intention prediction.

Conversely, in Fig. 13(b), where IntentFormer is trained without the co-learning module, the pedestrian torso and some scene elements sparsely receive higher weights by the last training epoch. The input pixels are not highlighted precisely or comprehensively as in the co-learning training mode. This less effective localization of important features reduces the ability to identify the most relevant features for intention prediction.

4.2.5. *Impact of individual modalities, their combinations and fusion order*

This section investigates the impact of different modalities and fusion order permutations on the overall performance of pedestrian intention prediction. In our recent work [5], pedestrian appearance, scene context, pose, trajectory, and ego-vehicle speed were utilised for pedestrian intention prediction. The analysis demonstrated that visual features achieved the highest performance metrics, followed by trajectory and pose features. In contrast, pose features contributed the least when utilized as graph node features to model the temporal relationships of pedestrian interactions. Based on these findings, the proposed work incorporates only RGB crops, trajectory, and segmentation maps for context as the primary modalities for the proposed intention prediction model. This approach effectively minimizes the additional memory footprint associated with pose features without significantly impacting the model's overall performance.

Table 4
Performance comparison of the IntentFormer model with different modalities, their combinations, and the order of fusion

Modalities	Accuracy		
	PIE	$JAAD_{beh}$	$JAAD_{all}$
T	0.56	0.41	0.55
R	0.59	0.45	0.60
S	0.43	0.39	0.40
T+R	0.63	0.52	0.64
T+S	0.58	0.48	0.61
R+S	0.66	0.54	0.69
T+S+R	0.78	0.68	0.82
T+R+S	0.76	0.67	0.80
S+T+R	0.88	0.69	0.86
S+R+T	0.89	0.70	0.88
R+T+S	0.90	0.69	0.85
R+S+T	**0.93**	**0.75**	**0.92**

**R: RGB Images, S=Segmentation Maps, T: Trajectory*

It can be observed from Table 4 that individual modalities (R: RGB pedestrian crops, S: Segmentation maps and T: Trajectory) achieve the lowest accuracy. When assessing single modality performance, input feed is given only through the first encoder stage; no other feed is given through subsequent encoder stages. In subsequent ablations involving combinations of two modalities, input feed is given through the first and second encoder stages. Combining these modalities leads to substantial performance improvements. For instance, combining T+R increases accuracy by 12.5 % on PIE, T+S increases accuracy by 3.6 %, and R+S increases accuracy by 16.4 %, considering accuracy with only T as baseline. The highest accuracy is obtained with the combination R+S+T, resulting in a 66.1 % increase in PIE, a 66.7 % increase in $JAAD_{beh}$, and a 53.3 % increase in $JAAD_{all}$ over baseline, demonstrating the effectiveness of integrating these modalities. In the case of a single modality in any of the encoder stages with no other modality feed, the MHSWA, designed for the fusion of two diverse modalities within the encoder, operates as standard MHA.

The experiments with different orders of fusion, as reported in Table 4, highlight that a noticeable dip in performance is observed when features such as RGB images and segmentation maps are integrated at

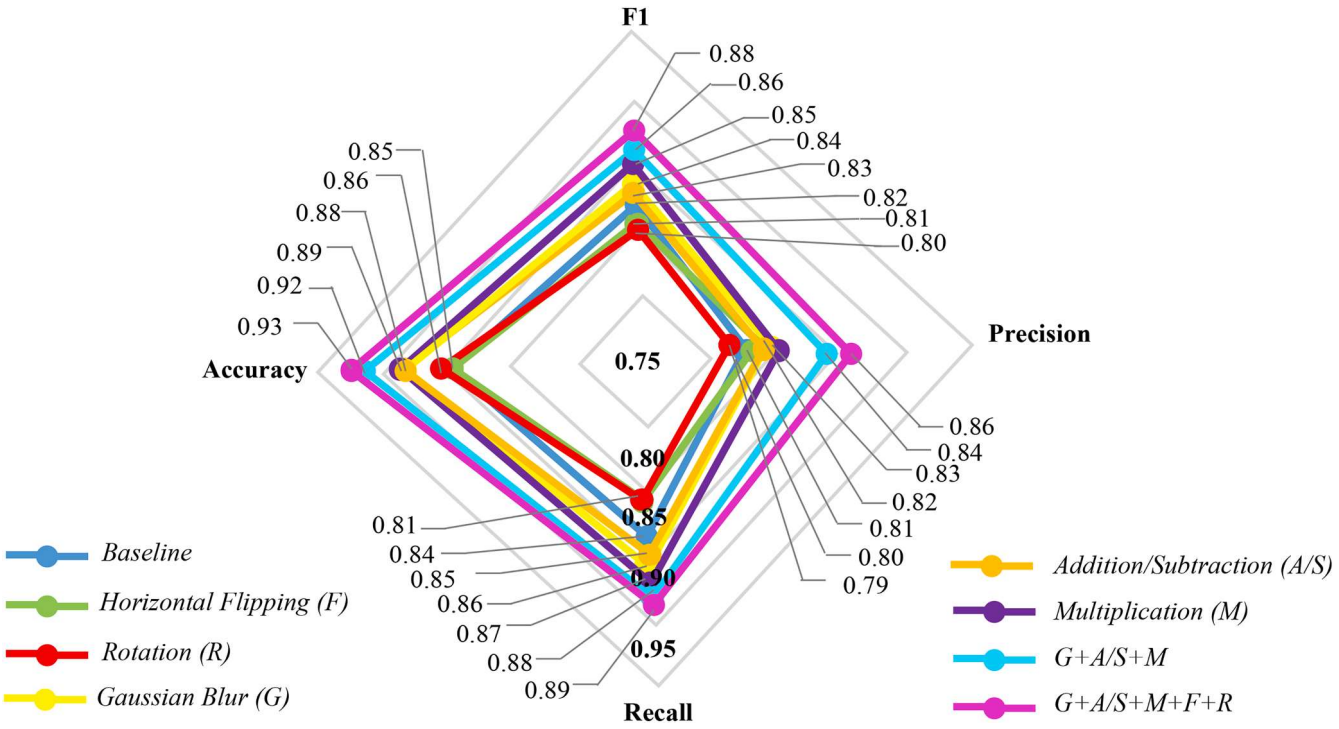


Fig. 14. A visual classification performance comparison of the IntentFormer model trained with various data augmentation techniques and their combinations. The combination of Gaussian blur, addition/subtraction, multiplication, horizontal flipping, and rotation (G + A/S + M + F + R) in ***pink*** achieves the highest performance metrics.

later stages of the network. By selecting the correct permutation by feeding trajectory at the last stage of the network, the accuracy performance improves by up to 9 % on PIE, 5 % on $JAAD_{beh,}$ and more than 8 % on $JAAD_{all}$. This observation can be attributed to the proposed architecture's ability to leverage visual features in the earlier network stages effectively. The subsequent integration of dynamic features like trajectory coordinates at later stages optimally takes advantage of the enriched contextual understanding constructed by prior modalities. By aligning the integration order with the intrinsic complexity of features, the architecture maximizes the information captured by each modality. These findings highlight the pivotal role of the chosen sequence of feature integration in enhancing prediction accuracy.

4.2.6. *Effect of data augmentation*

Fig. 14 illustrates the impact of various augmentation techniques on the performance of our pedestrian intention prediction model. Among the techniques evaluated, horizontal flipping (F) and rotation (R) provided minimal enhancements compared to the baseline without augmentation. Additionally, Gaussian blur (G), addition/subtraction (A/S), and multiplication (M) demonstrated notable improvements, increasing overall performance metrics by 2.71 %, 2.01 %, and 3.63 %, respectively, relative to the baseline. The combination of Gaussian blur, addition/subtraction, and multiplication (G + A/S + M) resulted in substantial enhancements, boosting accuracy by 8.24 %, F1 score by 4.88 %, precision by 5 %, and recall by 4.76 %. The inclusion of all five augmentations (G + A/S + M + F + R) yielded the highest overall improvements, with increases in accuracy by 9.41 %, F1 score by 7.32 %, precision by 7.50 %, and recall by 5.95 %.

These results demonstrate that complex augmentations such as Gaussian blur, addition/subtraction, and multiplication significantly enhance the model's ability to predict pedestrian intentions. Although primary augmentations like horizontal flipping and rotation are insufficient to capture the complexities of pedestrian movements and interactions, the synergistic effect observed from combining multiple

Table 5
Quantitative evaluation on the PIE/JAAD Dataset

Ablations	Model Variants							Accuracy		
	MLP Heads		Multi-Head Attention			Loss				
	MLP	MLP-shared	MHCMA	MHMMA	MHSWA	BCE	CAC	PIE	$JAAD_{beh}$	$JAAD_{all}$
1	✓	×	✓	×	×	✓	×	0.89	0.69	0.88
2	✓	×	×	✓	×	✓	×	0.89	0.70	0.87
3	✓	×	×	×	✓	✓	×	0.91	0.69	0.91
4	×	✓	✓	×	×	✓	×	0.90	0.70	0.90
5	×	✓	×	✓	×	✓	×	0.91	0.71	0.91
6	×	✓	×	×	✓	✓	×	0.90	0.70	0.90
7	✓	×	✓	×	×	×	✓	0.86	0.69	0.89
8	✓	×	×	✓	×	×	✓	0.87	0.65	0.88
9	✓	×	×	×	✓	×	✓	0.88	0.63	0.87
10	×	✓	✓	×	×	×	✓	0.91	0.70	0.88
11	×	✓	×	✓	×	×	✓	0.92	0.71	0.89
12	×	✓	×	×	✓	×	✓	**0.93**	**0.75**	**0.92**

Table 6
Comparison of IntentFormer with state-of-the-art models on the PIE, $JAAD_{beh}$, and $JAAD_{all}$ datasets, highlighting memory footprint, inference time, and highest achieved accuracy. Data have been extracted from the respective publications.

Model	Size (MB)	Inference time(ms)	Accuracy (PIE)	Accuracy ($JAAD_{beh}$)	Accuracy ($JAAD_{all}$)
PCPA [28]	118.8	38.6	86	50	70
FFSTA [4]	374.2	70.83	–	62	83
PG+ [19]	0.28	5.47	89	70	86
TED [11]	12.8	2.76	91	–	–
V-PedCross [26]	4.8	–	89	64	86
PedGNN [21]	**0.027**	**0.58**	70.52	–	86.22
VMI [5]	19.07	11.03	92	–	89
IntentFormer	2.13	3.8	**93**	**75**	**92**

Table 7
Model Architecture and Hyperparameter Configuration

Modules/Layers/ Encoders	Trainable Parameters			Hyperparameters
	Proposed IntentFormer		Vanilla Transformer	
	Non-Shared	Shared	Non-shared	
Tubelet/ Recurrent Projection (TP/ RP)	38K (*total*)	–	38K(*total*)	***TP**: Conv 3D: Filters-64, Kernel Size-(2,8,8)* ***RP:** GRU: Hidden units-64*
Positional Encoder_TP	351K	–	351K	*Embedding Layer Output Dimension- 64*
Positional Encoder_RP	896	–	896	*Embedding Layer Output Dimension- 64*
MHSA/MHSWA	16.6K		16.6K	*No. of heads (4), Size of each attention head (64), Dropout-50 %*
PCP	82K	–	82K	*Conv 1D:* 1×1
Shared MLP	–	33K	–	*Two sequential MLPS with* 64×4 *and 64 neurons, Dropout-50 %*
Layer Normalization (LN)	128			–
Classification Head	130	–	130	*Layer Normalization, GAP, Dropout-50 %, MLP with 2 neurons*
$TE_I + TE_{II} + TE_{III}$	132K	33K	231K	–
Total	522K	33K	621K	–

augmentations highlights that diverse and comprehensive augmentations can collectively enhance the model's robustness and accuracy in pedestrian intention prediction tasks.

4.2.7. Quantitative analysis

The analysis of model ablations in Table 5 reveals a notable 3-4 % increase in accuracy for shared MLP configurations compared to their non-shared MLP counterparts. The multi-head attention configurations (MHCMA, MHMMA, MHSWA) demonstrate a systematic rise in accuracy across all datasets, with MHCMA exhibiting the lowest accuracy and the proposed MHSWA achieving the highest levels. This validates the impact of shared weight attention among diverse modalities (RGB images, Segmentation maps and trajectory) in a co-learning framework. The proposed Co-learning Adaptive Composite (CAC) loss also shows comparable performance to the widely used Binary Cross-Entropy (BCE) loss. It also introduces a significant improvement in regularization, leading to reduced fluctuations in validation accuracy, as visually depicted in Fig. 9. These collective findings underscore the effectiveness and efficiency of the proposed IntentFormer architecture in capturing intricate relationships among modalities for robust pedestrian intention prediction.

The IntentFormer model achieves superior accuracy of 93 % on the PIE dataset, 75 % on the $JAAD_{beh}$ dataset, and 92 % on the $JAAD_{all}$ dataset while maintaining a competitive memory footprint of 2.13 MB and an inference time of 3.8 ms, as shown in Table 6. It consists of 555k parameters, showcasing a substantial decrease in parameters by approximately 11 % compared to the vanilla transformer with 621K parameters, suggesting more parameter-efficient learning (as observed in Table 7). This parameter reduction also results in a 10 % decrease in memory footprint. One of the key reasons behind the competitive memory footprint achieved for the proposed architecture is the co-learning module and the Multi-head Shared Weights Attention devised for model training that keeps the trainable parameters limited in numbers. Although the memory footprint is higher than that of PedGNN [21], IntentFormer offers a significant accuracy improvement, with a 27.62 % increase on the PIE dataset and a 3.22 % increase on the JAADall dataset compared to PedGNN [21]. Thus, despite PedGNN's minimal memory footprint of 0.027 MB, it fails to adequately address the complex dynamics of real-time scenes compared to the proposed IntentFormer. These results highlight the model's efficiency and effectiveness, making it well-suited for real-time applications in autonomous driving.

4.2.8. Failure cases

Although the proposed models outperform SOTA in predicting pedestrian crossing intentions, they face challenges with unpredictable human behaviour and adverse environmental conditions, underscoring the need for further refinement to enhance accuracy and reliability in real-world scenarios. For instance, in Fig. 15, the target pedestrian is observed walking on the sidewalk with their back turned, showing no intention of crossing during the initial frames ($TTE \geq 2$ secs). However, in the subsequent frames ($TTE < 2$ secs), an unexpected change in direction occurs, with the pedestrian now appearing to wait at the curb, indicating a positive crossing intention. Such complex cases, reflecting unpredictable human behaviour, can mislead the model into making inaccurate predictions of the pedestrian's intention as indicated in the crossing prediction and the confidence scores *w.r.t.* TTE in Fig. 15. Consequently, despite the model demonstrating optimal performance at $TTE = 2.5$ *secs* as highlighted in *Section 4.2.1*, in this case, the intention is not accurately predicted until the TTE reaches 1.5 *secs* or less. This example highlights the model's limitations in handling highly dynamic and variable pedestrian behaviours.

Furthermore, Fig. 16 illustrates several instances where the model underperforms due to challenging conditions, such as poor illumination, low resolution resulting from the pedestrian's distance from the autonomous vehicle (AV), and occlusion by vehicles or other scene objects. These factors compromise the quality of the RGB modality features required by the model, leading to suboptimal performance.

5. Conclusions

The present study introduces a novel multimodal transformer-based architecture, '*IntentFormer*', that learns pedestrian road crossing intentions in a co-learning environment. Leveraging information from diverse modalities, including pedestrians' RGB features, segmentation maps of the surrounding scene, and the trajectory traced through consecutive frames, the proposed architecture stands out with three-fold contributions. The key contributions are a shared-MLP head for collaborative co-learning, multi-head shared weights attention (MHSWA) to capture inter-modal relationships efficiently, and the integration of Co-learning Adaptive Composite (CAC) loss, strategically mitigating overfitting risks by penalizing corresponding stages during training.

Several key findings elucidate the efficacy of the proposed IntentFormer in pedestrian intention prediction. The model excels with optimal performance in the 0.5 to 1.25 seconds observation range, showcasing efficiency with fewer frames and high TTE. Notably, the

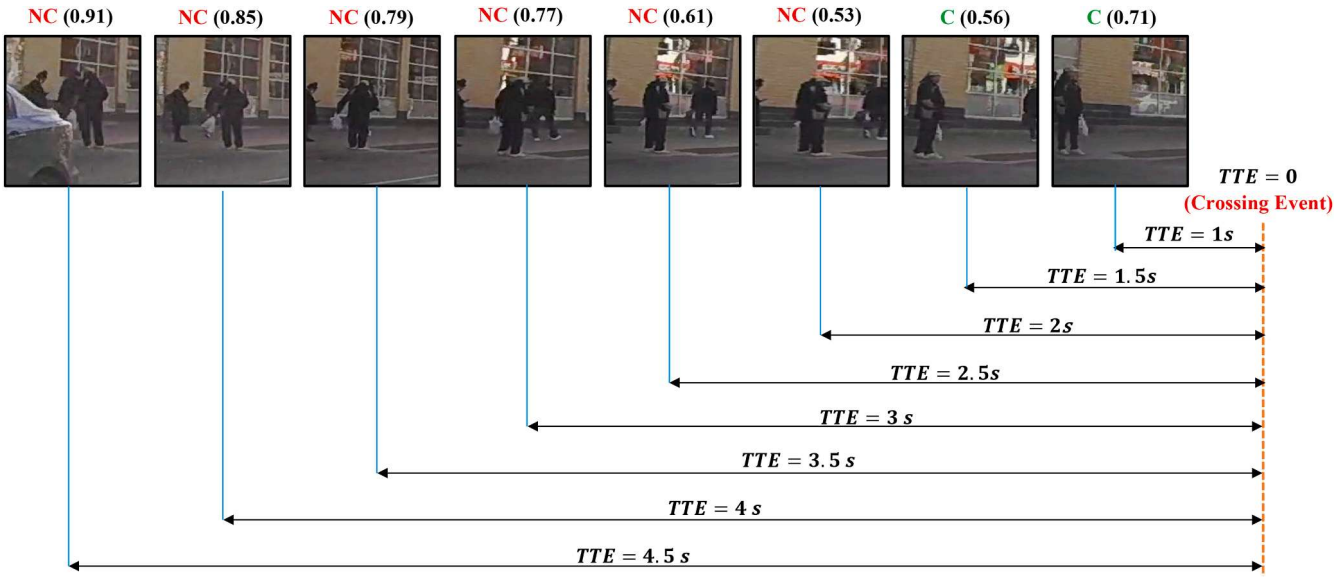


Fig. 15. Illustration of a pedestrian crossing sequence where the pedestrian initially displays no intention to cross until the time-to-event (TTE) is less than or equal to 2 secs. As the TTE decreases further, an unexpected change in direction observed in subsequent frames, indicating a positive crossing intention. The proposed model, however, fails to accurately predict this positive crossing intention until TTE is 1.5 seconds or less. This is reflected in the confidence scores at the corresponding prediction frame, where preceding frames are fed to the model as observed input frames.

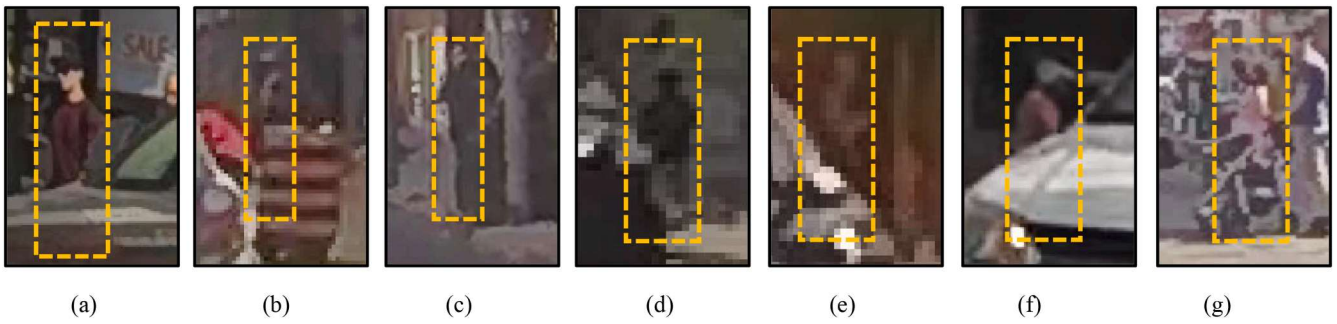


Fig. 16. Instances of failure cases under conditions of poor illumination, low resolution, and occlusion.

proposed approach outperforms state-of-the-art methods with a minor 12–14 % drop in performance metrics even at 3.5 seconds before the crossing event, highlighting its resilience against varying observation lengths compared to SF-GRU [3] and PG+ [21].

Furthermore, the study unveils the effectiveness of the Multi-Head Shared Weight Attention (MHSWA) method, which is evident in the precision-recall curve closely aligning with the ideal curve compared to alternative approaches. The dynamic shift in attention weights behaviour, evolving from varied ranges to a more uniform focus, indicates the model's progression toward integrating multimodal information for decision-making. The exploration of attention map heads within the proposed configuration emphasizes diverse aspects, from the outline and shape of the target pedestrian to details about the surroundings and dynamic variations in the pedestrian's movement. The fourth attention map stands out for its focus on identifying contours and distinct patterns within the cropped image. These findings underscore the model's ability to discern nuanced features across modalities, contributing to accurate intention predictions.

Lastly, the study emphasizes the critical role of feature integration order. Integrating features such as RGB images and segmentation maps at earlier stages proves advantageous, with trajectory integration at the last stage resulting in significant accuracy improvements. The proposed CAC loss also effectively regularizes the model, mitigating instability and overfitting issues during training. Moreover, the analysis of model ablations highlights a 3–4 % accuracy increase for shared-MLP configurations, coupled with a substantial 11 % parameter decrease compared to the vanilla transformer, indicating more efficient learning. These cumulative findings affirm the robustness and advancements the proposed multimodal architecture brings.

However, generating precise segmentation maps in a real-time egocentric perspective for dynamically changing traffic scenarios is challenging. Beyond this, the model's current architecture confronts hurdles in accurately predicting pedestrian intentions during unpredictable and abrupt changes in motion. Furthermore, the effectiveness of the proposed architecture may also be compromised in scenarios involving severe occlusion, poor illumination and low resolution. Moreover, the model's scalability to simultaneously manage a large number of pedestrians in crowded urban environments poses another challenge. Developing strategies to enhance the model's resilience to challenging scenarios and to better account for the inherent randomness in pedestrian motion is crucial. Future research should address these limitations to improve the model's applicability, robustness, and feasibility in diverse real-world settings.

CRediT authorship contribution statement

Neha Sharma: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis,

Conceptualization. **Chhavi Dhiman:** Writing – review & editing, Visualization, Supervision, Methodology, Formal analysis, Conceptualization. **Sreedevi Indu:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The link to access the implementation of the proposal is provided in abstract section.

References

[1] "Autonomous vehicles market size forecast 2030 | industry share report." [Online]. Available: https://www.marketresearchfuture.com/reports/autonomous-vehicles-market-1020.

[2] N. Sharma, C. Dhiman, S. Indu, Pedestrian intention prediction for autonomous vehicles: a comprehensive survey, Neurocomputing. 508 (Oct. 2022) 120–152.

[3] A. Rasouli, I. Kotseruba, J.K. Tsotsos, Pedestrian action anticipation using contextual feature fusion in stacked RNNs, in: 30th British Machine Vision Conference 2019, BMVC, 2019, pp. 1–13. May 2020.

[4] D. Yang, H. Zhang, E. Yurtsever, K.A. Redmill, Ü. Özguner, Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention, IEEE Trans. Intell. Vehicles 7 (2) (2022) 221–230. Apr.

[5] N. Sharma, C. Dhiman, S. Indu, Visual-motion-interaction-guided pedestrian intention prediction framework, IEEe Sens. J. 23 (22) (2023) 27540–27548. Nov.

[6] R.Q. Minguez, I.P. Alonso, D. Fernandez-Llorca, M.A. Sotelo, Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition, IEEE Trans. Intell. Transport. Syst. 20 (5) (2019) 1803–1814.

[7] C. Wang, Y. Wang, M. Xu, D.J. Crandall, Stepwise goal-driven networks for trajectory prediction, IEEe Robot. Autom. Lett. 7 (2) (2022) 2716–2723. Apr.

[8] N. Osman, G. Camporese, L. Ballan, TAMformer: multi-modal transformer with learned attention mask for early intent prediction, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Rhodes Island, Greece, 2023, pp. 1–5. Jun.

[9] Y. Zhou, G. Tan, R. Zhong, Y. Li, C. Gou, PIT: progressive interaction transformer for pedestrian crossing intention prediction, IEEE Trans. Intell. Transport. Syst. 24 (12) (2023).

[10] A. Vaswani, et al., Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017, pp. 6000–6010.

[11] L. Achaji, J. Moreau, T. Fouqueray, F. Aioun, F. Charpillet, Is attention to bounding boxes all you need for pedestrian action prediction?, in: Intelligent Vehicles Symposium (IV) IEEE, Aachen, Germany, 2022, pp. 895–902. Jun.

[12] O. Hamed, H.J. Steinhauer, Pedestrian intention recognition and action prediction using a feature fusion deep learning approach, in: Proceedings of The 18th International Conference on Modeling Decisions for Artificial Intelligence (MDAI), 2021, pp. 89–100.

[13] A. Singh, U. Suddamalla, Multi-input fusion for practical pedestrian intention prediction, in: International Conference on Computer Vision Workshops (ICCVW), IEEE, Montreal, BC, Canada, 2021, pp. 2304–2311. Oct.

[14] A. Rasouli, I. Kotseruba, T. Kunic, J. Tsotsos, PIE: a large-scale dataset and models for pedestrian intention estimation and trajectory prediction, in: International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 6261–6270. Oct.

[15] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, X. Du, Coupling intent and action for pedestrian crossing behavior prediction, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, California, 2021, pp. 1238–1244. Aug.

[16] A. Rasouli, T. Yau, M. Rohani, J. Luo, Multi-modal hybrid architecture for pedestrian action prediction, in: IEEE Intelligent Vehicles Symposium, Proceedings, Aachen, Germany, 2022.

[17] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780. Nov.

[18] T. Chen, R. Tian, Z. Ding, Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention, in: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 2021, pp. 3096–3102. Oct.

[19] P.R.G. Cadena, Y. Qian, C. Wang, M. Yang, Pedestrian graph +: a fast pedestrian crossing prediction model based on graph convolutional networks, IEEE Trans. Intell. Transport. Syst. 23 (11) (2022) 21050–21061.

[20] X. Zhang, P. Angeloudis, Y. Demiris, Dual-branch spatio-temporal graph neural networks for pedestrian trajectory prediction, Pattern. Recognit. 142 (2023) 109633. Oct.

[21] M.N. Riaz, M. Wielgosz, A.G. Romera, A.M. López, Synthetic data generation framework, dataset, and efficient deep model for pedestrian intention prediction, in: 26th International Conference on Intelligent Transportation Systems (ITSC), IEEE, Bilbao, Spain, 2023, pp. 2742–2749. Sep.

[22] Y. Ling, Q. Zhang, X. Weng, Z. Ma, STMA-GCN_PedCross: skeleton based spatial-temporal graph convolution networks with multiple attentions for fast pedestrian crossing intention prediction, in: 26th International Conference on Intelligent Transportation Systems (ITSC), IEEE, Bilbao, Spain, 2023, pp. 500–506. Sep.

[23] B. Liu, et al., Spatiotemporal relationship reasoning for pedestrian intent prediction, IEEe Robot. Autom. Lett. 5 (2) (2020) 3485–3492.

[24] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P.S. Yu, A comprehensive survey on graph neural networks, IEEe Trans. Neural Netw. Learn. Syst. 32 (1) (2019) 4–24. Jan.

[25] Z. Zhang, R. Tian, Z. Ding, TrEP: transformer-based evidential prediction for pedestrian intention with uncertainty, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 3534–3542. Jun.

[26] J. Bai, X. Fang, J. Fang, J. Xue, C. Yuan, Deep virtual-to-real distillation for pedestrian crossing prediction, in: 25th International Conference on Intelligent Transportation Systems (ITSC), IEEE, Macau, China, 2022, pp. 1586–1592. Oct.

[27] Z. Cao, G. Hidalgo, T. Simon, S.E. Wei, Y. Sheikh, OpenPose: realtime multi-person 2D pose estimation using part affinity fields, IEEe Trans. Pattern. Anal. Mach. Intell. 43 (1) (2021) 172–186, https://doi.org/10.1109/TPAMI.2019.2929257.

[28] I. Kotseruba, A. Rasouli, J.K. Tsotsos, Benchmark for evaluating pedestrian action prediction, in: Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 2021, pp. 1257–1267. Jan.

[29] A. Rasouli, I. Kotseruba, PedFormer: pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning, in: International Conference on Robotics and Automation (ICRA), IEEE, London, United Kingdom, 2023, pp. 9844–9851. May.

[30] R. Ni, B. Yang, Z. Wei, H. Hu, C. Yang, Pedestrians crossing intention anticipation based on dual-channel action recognition and hierarchical environmental context, IET Intell. Transport Syst. 17 (2022) 1–15.

[31] S. Lu, F. Gao, C. Piao, Y. Ma, Dynamic weighted cross entropy for semantic segmentation with extremely imbalanced data, in: 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), 2019, pp. 230–233, https://doi.org/10.1109/AIAM48774.2019.00053. Oct.

[32] Z. Zhou, H. Huang, B. Fang, Application of weighted cross-entropy loss function in intrusion detection, JCC 09 (11) (2021) 1–21, https://doi.org/10.4236/jcc.2021.911001.

[33] J.R. Layza, H. Pedrini, R. da S. Torres, 1-to-N Large Margin Classifier, in: 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2020, pp. 316–323, https://doi.org/10.1109/SIBGRAPI51738.2020.00050. Nov.

[34] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.

[35] R. Girdhar, J.J. Carreira, C. Doersch, A. Zisserman, Video action transformer network, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 244–253. Jun.

[36] Z. Zhong, D. Schneider, M. Voit, R. Stiefelhagen, J. Beyerer, Anticipative feature fusion transformer for multi-modal action anticipation, in: Proceeding of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2023.

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

[38] A. Rasouli, I. Kotseruba, J.K. Tsotsos, It's not all about size: on the role of data properties in pedestrian detection, in: Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2019, pp. 210–225.

[39] A. Kapishnikov, S. Venugopalan, B. Avci, B. Wedin, M. Terry, T. Bolukbasi, Guided integrated gradients: an adaptive path method for removing noise, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5048–5056. Jun.