Predicting Pedestrian Crossing Intention With Feature Fusion and Spatio-Temporal Attention

Dongfang Yang, *Member, IEEE*, Haolin Zhang, Ekim Yurtsever, *Member, IEEE*, Keith A. Redmill, *Senior Member, IEEE*, and Ümit Özgüner, *Life Fellow, IEEE*

***Abstract*—Predicting vulnerable road user behavior is an essential prerequisite for deploying Automated Driving Systems (ADS) in the real-world. Pedestrian crossing intention should be recognized in real-time, especially for urban driving. Recent works have shown the potential of using vision-based deep neural network models for this task. However, these models are not robust and certain issues still need to be resolved. First, the global spatio-temporal context that accounts for the interaction between the target pedestrian and the scene has not been properly utilized. Second, the optimal strategy for fusing different sensor data has not been thoroughly investigated. This work addresses the above limitations by introducing a novel neural network architecture to fuse inherently different spatio-temporal features for pedestrian crossing intention prediction. We fuse different phenomena such as sequences of RGB imagery, semantic segmentation masks, and ego-vehicle speed in an optimal way using attention mechanisms and a stack of recurrent neural networks. The optimal architecture was obtained through exhaustive ablation and comparison studies. Extensive comparative experiments on the JAAD and PIE pedestrian action prediction benchmarks demonstrate the effectiveness of the proposed method, where state-of-the-art performance was achieved. Our code is open-source and publicly available: https://github.com/OSU-Haolin/Pedestrian_Crossing_Intention_Prediction.**

***Index Terms*—Pedestrian intention, autonomous driving, spatial-temporal fusion.**

I. Introduction

AUTONOMOUS driving technology has progressed significantly in the past few years. However, to develop vehicle intelligence that is comparable to human drivers, understanding and predicting the behaviors of traffic agents is indispensable. This work aims to develop behavior understanding algorithms for vulnerable road users. Specifically, a vision-based pedestrian crossing intention prediction algorithm is proposed.

Manuscript received October 6, 2021; revised January 31, 2022; accepted March 9, 2022. Date of publication March 28, 2022; date of current version July 12, 2022. This work was supported by the United States Department of Transportation under Grant 69A3551747111 for Mobility 21 University Transportation Center. *(Dongfang Yang and Haolin Zhang contributed equally to this work.) (Corresponding author: Dongfang Yang.)*

Dongfang Yang is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA, and also with the Chongqnig Chang'an Automobile Company, Ltd., Chongqing 400020, China (e-mail: yang.3455@osu.edu).

Haolin Zhang, Ekim Yurtsever, Keith A. Redmill, and Ümit Özgüner are with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: zhang.10749@osu.edu; ekimyurtsever@gmail.com; redmill.1@osu.edu; ozguner.1@osu.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2022.3162719.

Digital Object Identifier 10.1109/TIV.2022.3162719

Behavior understanding plays an crucial role in autonomous driving. It establishes the trust between people and autonomous driving systems. By explicitly showing passengers how the system makes its decisions, people will be more willing to accept this technology.

In level 4 autonomy's driving, pedestrian crossing behavior is one of the most important behaviors that needs to be studied urgently. In urban scenarios, vehicles frequently interact with crossing pedestrians. If the autonomous system fails to handle vehicle-pedestrian interactions appropriately, casualties will most likely occur. With accurate intention prediction, the decision-making and planning modules in autonomous driving systems can access additional meaningful information, hence generating safer and more efficient maneuvers.

Nowadays, visual sensors such as front-facing cameras are becoming the standard configuration for autonomous driving systems. In the tasks of object detection and tracking, both the software and hardware of vision components are mature and ready for mass production. This provides a perfect platform on which vision-based behavior prediction algorithms can be deployed. Researchers and engineers in the prediction field can just focus on algorithm design. When the algorithm is ready, deployment becomes relatively trivial. The proposed algorithm is based on pure vision, it can be easily deployed. As long as the prediction algorithm is appropriately tested and verified, mass deployment becomes straightforward.

Vision-based pedestrian crossing intention prediction has been explored for several years. Early works [1] usually utilized a single frame as input to a convolutional neural network (CNN) based prediction system. This approach ignores the temporal aspect of image frames, which plays a critical role in the intention prediction task. Later on, with the maturity of recurrent neural networks (RNNs), pedestrian crossing intention was predicted by considering both the spatial and temporal information [2]–[4]. This led to different ways of fusing different features, e.g., the detected pedestrian bounding boxes, poses, appearance, and even the ego-vehicle information [5]–[9]. The most recent benchmark of pedestrian intention prediction was released by [10], in which the PCPA model achieved the state-of-the-art in the most popular dataset JAAD [1]. However, PCPA does not consider global contexts such as road geometry and other road users, factors we believe are nonnegligible in pedestrian crossing intention prediction. Furthermore, the existing fusion strategies may not be optimal.

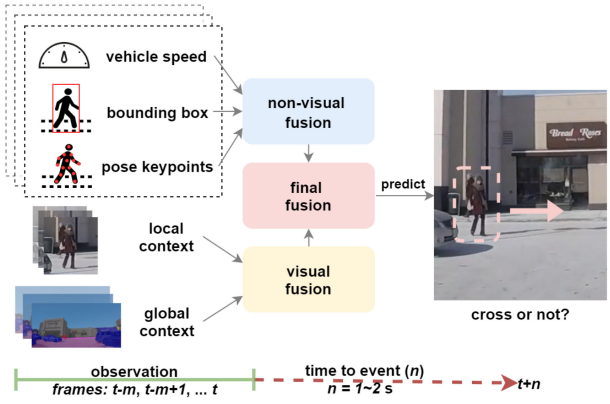


Fig. 1. Predicting pedestrian crossing intention is a multi-modal spatio-temporal problem. Our method fuses inherently different spatio-temporal phenomena with CNN-based visual encoders, RNN stacks, and attention mechanisms to achieve state-of-the-art performance.

In this work, we focus on improving the performance of vision-based prediction of pedestrian crossing intention, i.e., whether a pedestrian detected by a front-facing camera will cross the road or not in a short time horizon (1-2 s). Our work leverages the power of deep neural networks and fuses the features from different channels. As shown in Fig. 1, the proposed model considers both non-visual and visual information. They are extracted from a sequence of video frames 1-2 s before the crossing / not crossing (C/NC) event. Non-visual information includes the pedestrian's bounding box, pose keypoints, and ego-vehicle speed. Visual information contains local context and global context. Local context is the enlarged pedestrian appearance based on the bounding box position. Global context is the semantic segmentation of the road, pedestrians (all pedestrians in the scene), and vehicles. They are used because they significantly affect the target pedestrian's crossing decision. We propose a hybrid method of fusing the the non-visual and visual features, which is justified by comparing different strategies of feature fusion.

Our main contributions are as follows:

- A novel vision-based pedestrian intention prediction framework for ADSs and ADASs. The proposed method employs a novel neural network architecture for utilizing different spatio-temporal features with a hybrid fusion strategy.
- Extensive ablation studies on different feature fusion strategies (early, later, hierarchical, or hybrid), input configurations (adding/removing input channels, using semantic segmentation masks as explicit global context), and visual encoder options (3D CNN or 2D convolution with RNN + attention) to identify the best model layout.
- Demonstrating the efficiency of the proposed method on the commonly used JAAD [1] and PIE [4] datasets, and achieving state-of-the-art performance on the most recent pedestrian action prediction benchmark [10].

II. Related Work

Vision-based pedestrian crossing prediction traces back to the works [11] that utilize the Caltech Pedestrian Detection Benchmark [12]. However, the Caltech dataset does not explicitly annotate the crossing behavior of the pedestrians. This gap was later filled by the introduction of the JAAD dataset [1] that offers high-resolution videos and explicit crossing behavior annotations. With the release of the JAAD dataset, a simple baseline was also created that uses a 2D convolutional neural network (CNN) to encode the features in a given previous frame and then uses a linear support vector machine (SVM) to predict the C/NC event.

Spatio-temporal modeling: Instead of using a single image, most recent works use image sequences as input to the prediction model due to the importance of temporal information in the prediction task. This leads to spatio-temporal modeling.

Spatio-temporal modeling can be achieved by first extracting visual (spatial) features per frame via 2D CNNs [13] or graph convolution networks (GCNs) [14], and then feeding these features into RNNs such as the long-short term memory (LSTM) model [15] and the gate recurrent unit (GRU) model [16]. For example, [2]–[4] use 2D convolution to extract the visual features from image sequence and RNNs to encode the temporal information among these features. The encoded sequential visual features are fed into a fully-connected layer to obtain the final intention prediction. [14] uses a graph representation to encode the spatial relationship among the target pedestrian and surrounding agents. The prediction task was evaluated from two different perspectives, a pedestrian-centric setting and a location-centric setting. However, ego vehicle motion and explicit visual features are not modeled in this work.

Another way of extracting the sequential visual features is utilizing a 3D CNN [17]. It directly captures the spatio-temporal features by replacing the 2D kernels of the convolution and the pooling layers in the 2D CNN with 3D counterparts. For example, [18], [19] use a 3D CNN based framework (3D DenseNet) to directly extract the sequential visual features from the pedestrian image sequence. The final prediction is achieved by using a fully-connected layer.

The crossing intention prediction task can also be combined with scene prediction. A couple of works [20], [21] attempted to decompose the prediction task into two stages. In the first stage, the model predicts a sequence of future scenes using an encoder/decoder network. Then, pedestrian actions are predicted based on the generated future scenes using a binary classifier.

Feature fusion: Instead of end-to-end modeling of visual features, information such as pedestrian's bounding box, body-pose keypoints, vehicle motion, and the explicit global scene context can also be modeled as separate channels as inputs to the prediction model. This requires a proper way of fusing the above information.

For example, [5], [6], [22]–[24] introduced human poses/skeletons in pedestrian crossing prediction tasks since the human pose can be considered as a good indicator of human behaviors. By extracting the pose keypoints from cropped pedestrian images, crossing behavior classifiers were built based on the human pose feature vectors. Improvement in prediction accuracy shows the effectiveness of using pose features. However, these methods either only rely on human pose features without considering other important features or pay less attention to feature fusion.

Some other methods focused on novel fusion architectures. For instance, [7] proposed SF-GRU, a stacked RNN-based architecture, to hierarchically fuse five feature sources (pedestrian appearance, surrounding context, pose, bounding box, and ego-vehicle speed) for pedestrian crossing intention prediction. Nevertheless, this method does not take global context into account. [8] proposed a multi-modal based prediction system that integrates four feature sources (local scene, semantic map, pedestrian motion, and ego-motion). The global context (semantic map) is utilized, but it lacks other important features such as human pose. [25] proposed a multi-task based prediction framework to take advantages of feature sharing and multi-task learning. It fuses four feature sources (semantic map, pedestrians' trajectory, grid locations, and ego-motion). However, local context and human pose are not considered in the model.

Very recently, more datasets such as PIE [4] and PePScenes [26] provide annotations for fusing different features. A benchmark was also released with the PCPA model [10]. These create more room for researchers to explore the task of vision-based pedestrian crossing intention prediction.

III. Proposed Method

A. Problem Formulation

The task of vision-based pedestrian crossing intention prediction is formulated as follows. Given a sequence of observed video frames from the vehicle's front view and the relevant information of ego-vehicle motion, the goal is to design a model that can estimate the probability of the target pedestrian i's action $A_i^{t+n} \in \{0, 1\}$ of crossing the road, where t is the specific time of the last observed frame and n is the number of frames from the last observed frame to the crossing / not crossing (C/NC) event.

It is worth noting that the existing literature usually uses the terms pedestrian action, behavior, and intention interchangeably. What is being predicted in this work is whether a pedestrian will cross or not in a short time horizon. Here we use pedestrian crossing intention as surrogates for crossing action or behavior. We assume that the action of crossing is equivalent to the intention of crossing.

In the proposed model, explicit features such as pedestrian's bounding box, pose keypoints, local context (cropped image around the pedestrian), and global context (semantic segmentation) are first extracted. They are then used together with the vehicle's speed as separate channels that serve as the input to the prediction model. Our model has the following inputs:

- The sequential local context around pedestrian i:
$$C_{li} = \{c_{li}^{t-m}, c_{li}^{t-m+1}, \ldots, c_{li}^{t}\};$$
- The 2D location trajectory of pedestrian i denoted by bounding box coordinates (top-left points and bottom-right points):
$$L_i = \{l_i^{t-m}, l_i^{t-m+1}, \ldots, l_i^{t}\};$$
- Pose keypoints of pedestrian i:
$$P_i = \{p_i^{t-m}, p_i^{t-m+1}, \ldots, p_i^{t}\};$$
- Speed of ego-vehicle:
$$S = \{s^{t-m}, s^{t-m+1}, \ldots, s^{t}\};$$
- The sequential global context denoted by the mask of semantic segmentation:
$$C_g = \{c_g^{t-m}, c_g^{t-m+1}, \ldots, c_g^{t}\}.$$

Each source has a sequence of length $m + 1$. The input sources are illustrated in Fig. 2 and further described below.

B. Input Acquisition

Local context and 2D location trajectory: The local context C_{li} provides visual features of the target pedestrian. The 2D location trajectory L_i gives the position change of the target pedestrian in the image. They can be extracted by a detection (e.g. YOLO [27]) and tracking (e.g. Deep-SORT [28]) system. At present, the detection and tracking algorithms are good enough to generate near ground-truth results. Therefore, in this work, we directly use the ground truth C_{li} and L_i from the dataset. The main reason is that pedestrian detection and tracking are not the primary focus of this work. We would like to focus on the model architecture design and remove the noise from the detection and tracking. This also follows the configurations in most related works. A small part of the work of [6] considers the impact of 2D detection in the prediction task. However, their innovation and focus are on how to build the overall pipeline. Another reason is that by using ground truth we can fairly compare our method with most related works. Specifically, the local context $C_{li} = \{c_{li}^{t-m}, c_{li}^{t-m+1}, \ldots, c_{li}^{t}\}$ consists of a sequence of RGB images of size [224,224] pixels around the target pedestrian. The 2D location trajectory $L_i = \{l_i^{t-m}, l_i^{t-m+1}, \ldots, l_i^{t}\}$ consists of target pedestrian's bounding box coordinates, i.e.,

$$l_i^{t-m} = \{x_{it}^{t-m}, y_{it}^{t-m}, x_{ib}^{t-m}, y_{ib}^{t-m}\},$$

where $x_{it}^{t-m}, y_{it}^{t-m}$ denotes the top-left point and $x_{ib}^{t-m}, y_{ib}^{t-m}$ bottom-right point.

Pedestrian pose keypoints: Pedestrian pose keypoints represent the target pedestrian's detailed motion, i.e., the posture at each frame while moving. They can be obtained by applying a pose estimation algorithm on the local context C_{li}. Since the applied JAAD dataset does not provide ground truth pose keypoints, we utilize the pre-trained OpenPose model [29] to extract the pedestrian pose keypoints $P_i = \{p_i^{t-m}, p_i^{t-m+1}, \ldots, p_i^{t}\}$, where p is a 36D vector of 2D coordinates that contain 18 pose joints, i.e.,

$$p_i^{t-m} = \{x_{i1}^{t-m}, y_{i1}^{t-m}, x_{i2}^{t-m}, y_{i2}^{t-m}, \ldots, x_{i18}^{t-m}, y_{i18}^{t-m}\}.$$

Ego-vehicle speed: Ego-vehicle speed S is a major factor that affects the pedestrian's crossing decision. It can be directly read from the ego-vehicle's system. Since the dataset contains the annotation of ego-vehicle's speed, we directly use the ground truth labels for the vehicle speed $S = \{s^{t-m}, s^{t-m+1}, \ldots, s^{t}\}$.

Global context: Global context $C_g = \{c_g^{t-m}, c_g^{t-m+1}, \ldots, c_g^{t}\}$ offers the visual features that account for multi-interactions between the road and road users, or among road users. In our work, we use pixel-level semantic masks to represent the global

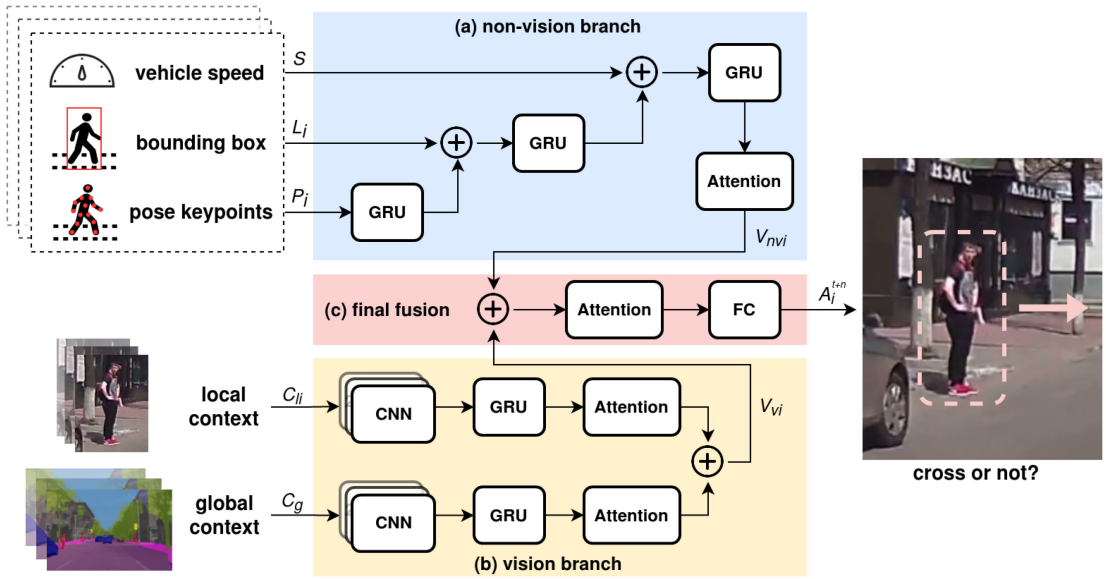


Fig. 2. Overview of the proposed pedestrian crossing intention prediction model. The yellow part denotes the fusion of visual features. 2D convolutional features of local context and global context are encoded by GRUs and fed to the attention blocks respectively. The two outputs are concatenated as final visual features. The blue part denotes the fusion of local features (non-visual). These non-visual features are encoded by another GRU and fused hierarchically, and then fed to an attention block to obtain the final non-visual features. The red part denotes the final fusion. Final visual features and final non-visual features are concatenated and fed to an attention block. A fully-connected (FC) layer is then applied to make the final prediction.

context. The semantic masks classify and localize different objects in the image by labeling all the pixels associated with the objects. Since the JAAD dataset does not have annotated ground truth of semantic masks, we use the DeepLabV3 model [30] pre-trained on the Cityscapes Dataset [31] to extract the semantic masks and select important objects (e.g. road, street, pedestrians and vehicles) as the global context. For the model to learn the interactions between the target pedestrian i and these objects, the target pedestrian is masked by an unique label. The mask area uses the target pedestrian i's bounding box (obtained from L_i). The semantic segmentation of all input frames are scaled to be of size [224,224] pixels, which is the same as the local context.

C. Model Architecture

The overall architecture is shown in Fig. 2. It consists of CNN modules, RNN modules, attention modules, and a novel way of fusing different features.

CNN module: We use the VGG19 [13] model pre-trained on the ImageNet dataset [32] to build the CNN module. Sequential RGB images are collected as a 4D array input with the dimensions of [number of observed frames, row, cols, channels] ($[16, 224, 224, 3]$ in this work), and then loaded by the CNN module. First, the feature map of every image from the fourth maxpooling layer of VGG19 is extracted with size [512,14,14]. Second, every feature map is averaged by a pooling layer with a 14×14 kernel, and then flattened and concatenated, to obtain a final feature tensor with size [16, 512], as sequential visual features.

RNN module: We use a gated recurrent unit (GRU) [16] to build the RNN module. The reason for choosing a GRU is that the GRU is more computationally efficient than its counterpart LSTM [15], which is older, and its architecture is relatively simple. The applied GRUs have 256 hidden units, which result in a feature tensor of size [16, 256].

Attention module: An attention module [33], by selectively focusing on parts of features, is used for better memorizing sequential sources. Sequential features (e.g. the output of RNN-based encoder) are represented as hidden states $h = \{h_1, h_2, \ldots, h_e\}$. The attention weight is computed as:

$$\alpha = \frac{exp(score(h_e, \tilde{h_s}))}{\sum_{s'} exp(score(h_e, \tilde{h_{s'}}))},$$

where $score(h_e, \tilde{h_s}) = h_e^T W_s \tilde{h_s}$ and W_s is a weight matrix. Such attention weight trades off the end hidden state h_e with each previous source hidden state $\tilde{h_s}$. The output vector of the attention module is produced as

$$V_{attention} = tanh(W_c[h_c; h_e]),$$

where W_c is a weight matrix, and h_c is the sum of all attention weighted hidden states as $h_c = \sum_{s'} \alpha \tilde{h_{s'}}$. The output of the attention module in our work is a feature tensor with size [1,256].

Hybrid fusion: We applied a hybrid way of fusing the features from different sources. The strategy is shown in Fig. 2. The proposed architecture has two branches, one for non-visual features and one for visual features.

The non-vision branch fuses three non-visual features (bounding boxes, pose keypoints, and vehicle speed). They are hierarchically fused according to their complexity and level of abstraction. The later the fusion stage occurs, the more impact the fused features will have on the final prediction. This is illustrated in Fig. 2(a). First, sequential pedestrian pose keypoints P_i are fed to an RNN-based encoder. Second, the output of the first stage is concatenated with 2D location trajectory L_i and fed

to a new RNN-based encoder. Finally, the output of the second stage is concatenated with ego-vehicle speed S and fed to a final RNN-based encoder. The output of the final encoder is then fed to an attention block to obtain the final non-visual feature vectors V_{nvi}.

The vision branch fuses two visual features, consisting of local context (enlarged pedestrian appearance around the bounding box) and global context (semantic segmentation of important objects in the whole scene), as shown in Fig. 2(b). Local context C_{li} is encoded by first extracting spatial features from the CNN module (as explained in the previous section) and then extracting temporal features from the GRU module. Global context C_g is encoded in the same way. Both local and global features are then fed into their attention modules, and finally, concatenated together to generate the final visual feature vectors V_{vi}.

Lastly, as shown in Fig. 2(c), the final non-visual feature vectors V_{nvi} and the final visual feature vectors V_{vi} are concatenated and fed into another attention block, followed by a fully-connection (FC) layer to obtain the final predicted action:

$$A_i^{t+n} = f_{FC}(f_{attention}(V_{nvi}; V_{vi})).$$

IV. Experiments

A. Dataset and Benchmark

The proposed model was evaluated using both the JAAD [1] and PIE [4] datasets. The JAAD dataset contains two subsets, JAAD behavioral data (JAAD$_{beh}$) and JAAD all data (JAAD$_{all}$). JAAD$_{beh}$ contains pedestrians who are crossing (495 samples) or are about to cross (191 samples). JAAD$_{all}$ has additional pedestrians (2100 samples) with non-crossing actions. To create a fair benchmark, the dataset configuration is the same as used in [10]. It uses a data sample overlap of 0.8 and a local context scale of 1.5.

The PIE dataset is a more comprehensive dataset compared to the JAAD dataset. It contains 1322 non-crossing samples and 512 crossing samples. Besides, the PIE dataset covers pedestrians with more different appearances and scenes with more different surroundings than those in the JAAD dataset.

The evaluation metrics use accuracy, AUC, F1 score, precision, and recall. These are the most recognized metrics and are used by most related works. False alarm rate is another important metric when deploying this algorithm in autonomous driving systems. False alarms may cause unnecessary brakes for autonomous cars, resulting in unpleasant experiences for the passengers. The above metrics inherently include the false alarm rate. They are more balanced metrics for evaluating a prediction system. Most related works and benchmarks adopt this metric system to report their results. Using this metric system, we can fairly compare our works with others.

B. Implementation

In the experiments, the proposed model was compared with the following methods: SingleRNN [2], SF-GRU [7] and PCPA [10]. We adopted the benchmark implementation released with the PCPA model [10]. This benchmark collects the implementations of most pedestrian intention prediction methods. Our model was developed based on this benchmark. We use a dropout of 0.5 in the attention module, L2 regularization of 0.001 in the FC layer, binary cross-entropy loss, and the Adam optimizer [34]. For the JAAD dataset, we use learning rate $= 5 \times 10^{-7}$, epochs $= 40$, and batch size $= 2$. For the PIE dataset, we use learning rate $= 5 \times 10^{-5}$, epochs $= 60$, and batch size $= 2$. ll models were trained and tested on the same split of the dataset, as suggested by the benchmark [10]. Note that the JAAD dataset does not provide explicit vehicle speed. Instead, the driver's action is recorded as an abstract encoding of the vehicle speed. The action contains [stopped (0), moving slow (1), moving fast (2), decelerating (3), accelerating (4)].

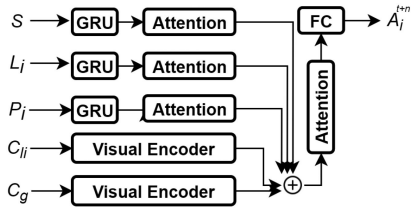


Fig. 3. Illustration of Later Fusion.

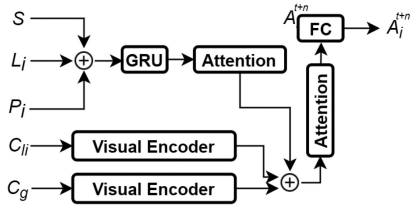


Fig. 4. Illustration of Early Fusion.

C. Ablation Study

An ablation study was also conducted to compare different strategies of fusing different features. In addition to the baseline methods (SingleRNN [2], SF-GRU [7] and PCPA [10]) and the proposed model (Ours), a total of 7 variants of the proposed model (Ours1, Ours2,... Ours7, as indicated in Tables V and VI) were trained and compared with the proposed one. First, for the visual encoder, we tried (1) a 2D CNN combined with RNN (VGG and GRU in our experiments) and (2) a 3D CNN as proposed in the PCPA model. Second, we tried the models with and without the global feature (semantic segmentation). Finally, we tried different fusion strategies that include later fusion, early fusion, and hierarchical fusion so that they can be compared with the proposed hybrid fusion strategy. Later fusion (Fig. 3) is the same as that proposed in PCPA [10]. Early fusion (Fig. 4) concatenates non-visual features and visual features directly and then sends them into one RNN module followed by an attention module. Hierarchical fusion (Fig. 5) gradually fuses both visual features and non-visual features by RNN modules in the same manner as in Fig. 2(a), followed by an attention module.

V. Results

A. Quantitative Results

Table I shows the qualitative results on the JAAD$_{beh}$ dataset. It compares the proposed model with baseline models of

TABLE I
QUANTITATIVE RESULTS ON THE JAAD BEHAVIOR SUBSET

Models	Model Variants			JAAD_{beh}				
	Visual Encoder	Global Context	Fusion Approach	Accuracy	AUC	F1 Score	Precision	Recall
SingleRNN [2]	VGG + GRU	✗	✗	0.60	0.54	0.70	0.65	0.76
SF-GRU [7]	VGG + GRU	✗	hierarchical-fusion	0.58	**0.56**	0.65	**0.68**	0.62
PCPA [10]	3D CNN	✗	later-fusion	0.56	0.54	0.63	0.66	0.60
Ours	VGG + GRU	✓	hybrid-fusion	**0.62**	0.54	**0.74**	0.65	**0.85**

TABLE II
QUANTITATIVE RESULTS ON THE JAAD ALL DATASET

Models	Model Variants			JAAD_{all}				
	Visual Encoder	Global Context	Fusion Approach	Accuracy	AUC	F1 Score	Precision	Recall
SingleRNN [2]	VGG + GRU	✗	✗	0.78	0.77	0.54	0.42	0.75
SF-GRU [7]	VGG + GRU	✗	hierarchical-fusion	0.76	0.77	0.53	0.40	0.79
PCPA [10]	3D CNN	✗	later-fusion	0.77	0.79	0.56	0.42	**0.83**
Ours	VGG + GRU	✓	hybrid-fusion	**0.83**	**0.82**	**0.63**	**0.51**	0.81

TABLE III
QUANTITATIVE RESULTS ON THE PIE DATASET

Models	Model Variants			PIE				
	Visual Encoder	Global Context	Fusion Approach	Accuracy	AUC	F1 Score	Precision	Recall
SingleRNN [2]	VGG + GRU	✗	✗	0.83	0.78	0.69	0.72	0.67
SF-GRU [7]	VGG + GRU	✗	hierarchical-fusion	0.84	0.80	0.71	0.72	0.71
PCPA [10]	3D CNN	✗	later-fusion	0.87	0.85	0.78	0.76	0.81
Ours	VGG + GRU	✓	hybrid-fusion	**0.89**	**0.86**	**0.80**	**0.79**	**0.81**

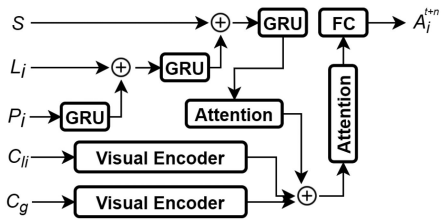


Fig. 5. Illustration of Hierarchical Fusion.

TABLE IV
COMPARISON OF COMPUTATIONAL COST

Model	Number of Params.
SingleRNN [2]	1,016,321
SF-GRU [7]	2,595,329
PCPA [10]	31,165,953
Ours	2,988,545

SingleRNN [2], SF-GRU [7] and PCPA [10]. The proposed model achieved the best scores in accuracy, F1, and recall. F1 score is an balanced metric considering both recall and precision. For binary classification, it is the most important indicator of quality of the model. Our model achieved about 4% improvement in F1. In addition to F1, accuracy is another important metric, and our model also achieved the best score.

Table II shows the qualitative results on the JAAD_{all} dataset. JAAD_{all} has additional samples of non-crossing behaviors. It is larger than JAAD_{beh}. The data distribution is more similar to real world scenarios. As illustrated by Table II, the proposed method achieved the best in accuracy, AUC, F1, and precision. Similar to the results in JAAD_{beh}, our model achieved the best score in terms of the two important metrics, F1 and accuracy.

Table III shows the qualitative results on the PIE dataset. On such a comprehensive dataset, our proposed method outperforms other methods with a considerable gap, which shows the importance and advantages of designing a hybrid fusion strategy and utilizing global context.

Note that the results of PCPA were generated based on the official implementation released by the PCPA author. We cannot reproduce the same results as reported in the PCPA paper. After communicating with PCPA's authors, they confirm that our reproduced result is normal.

We also analyzed the computational cost of the above models. The total number of model parameters is used as an indicator of computational cost. Table IV shows the comparison. PCPA [10] has the highest number of model parameters as it utilizes 3D convolution. Our model has only one-tenth of the parameters in PCPA, but still achieved better performance. This provides advantages in real-time deployment.

B. Qualitative Results

Fig. 6 provides qualitative results for the proposed model of pedestrian crossing intention prediction. We mainly compared the proposed method with the PCPA model. In the provided examples, our method correctly predicted the crossing intention but the PCPA failed. Taking a closer look at the examples, the following argument is raised. Without utilizing the global context, the task of crossing intention prediction may face the problems of (1) unknown direction of the pedestrian (Case a in Fig. 6), (2) occlusion (Case b in Fig. 6), and (3) poor vision (Case c in Fig. 6). Global context can provide additional information

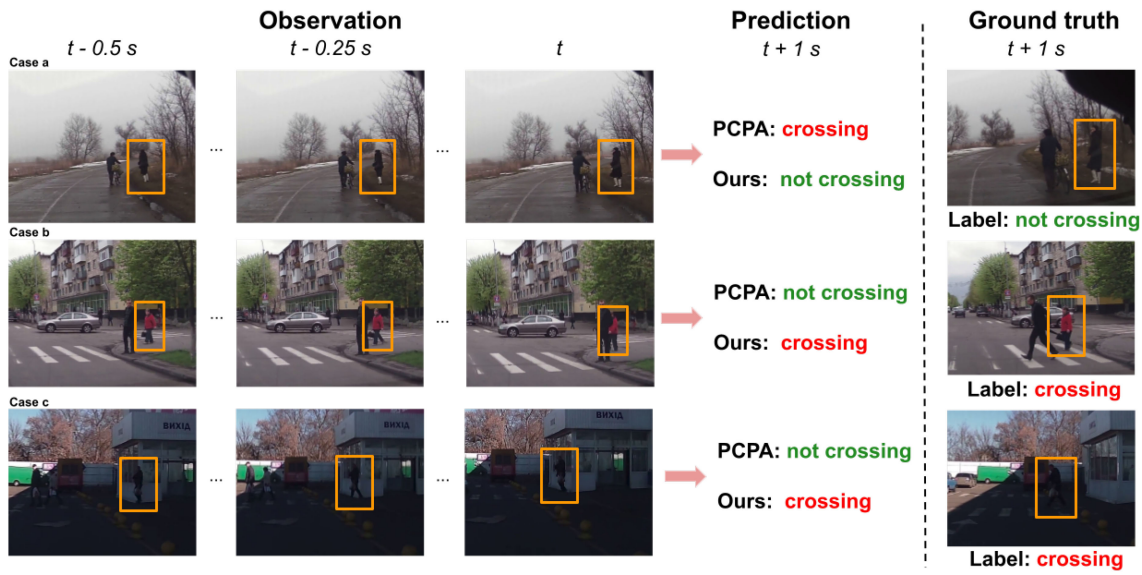


Fig. 6. Qualitative results on the JAAD dataset produced by PCPA [10] and our proposed model (Ours). The target pedestrians in images are enclosed by orange bounding boxes. The prediction results as well as ground truth labels are represented as red crossing or green not crossing.

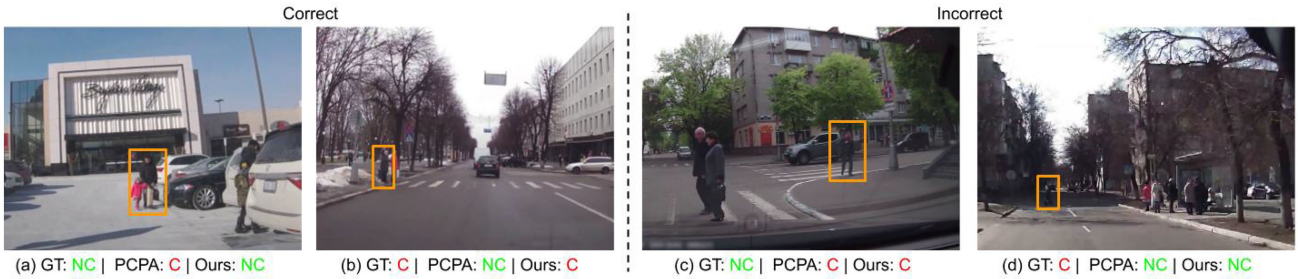


Fig. 7. More qualitative results. (a) and (b) show cases of correct prediction by the proposed model for which the PCPA failed. (c) and (d) show results when both the proposed and the PCPA model failed.

to account for the interaction between the whole scene and the target pedestrian.

Fig. 7 provides more qualitative results to analyze the advantages of the proposed model over the PCPA model as well as a few failure cases. Fig. 7(a) and (b) show cases when the proposed model generated correct predictions but the PCPA failed. The main reason is that our model considers the global visual context that contains the semantic segmentation of the drivable area. The model can learn from this whether the pedestrian is moving toward or on the drivable area, which is an important indicator of pedestrian crossing intention.

Fig. 7(c) and (d) show cases when both the proposed model and the PCPA failed. Fig. 7(c) shows an intersection scenario. The pedestrian (yellow bounding box) has already crossed the ego road but is near the edge of the road on the other side. This may mislead the model to generate a prediction of crossing. The failure in Fig. 7(d) was mainly due to the poor illumination such that the model cannot obtain enough detailed features.

C. Results of Ablation Study

Tables V and VI show the ablation study on the JAAD_{beh} and JAAD_{all} datasets, respectively. Table VII shows the ablation study on the PIE dataset. Different model variants are denoted by Ours1, Ours2,..., Ours7. By comparing Ours5 with Ours4 and Ours1 with the PCPA model, it shows that introducing global context can improve the model performance. In terms of fusion strategies, the proposed hybrid fusion strategy achieved the best performance, as seen by comparing Ours with Ours5, Ours6, and Ours7. If we further compare Ours4 with the PCPA model, it shows that using a 2D CNN plus RNN instead of a 3D CNN has a minimal impact on performance. This evidence also demonstrates that the improvement of our proposed method is mainly due to the new hybrid fusion strategy and global context.

D. Effect of Longer Prediction Horizon

It is claimed in some traffic studies [35] that the most suitable prediction horizon, i.e., time-to-event (TTE), is 1-2 seconds, because longer prediction horizon is impractical due to unpredictable nature of most urban scenarios and human dynamics [35]. However, to show the generalization ability, we still evaluated the proposed model with a longer TTE prediction horizon of 2-3 seconds. This was done by recreating the samples with a larger number of future frames. Table VIII shows the effect of different prediction horizons for the proposed model. It can

TABLE V
ABLATION STUDY ON THE JAAD BEHAVIOR SUBSET

Models	Model Variants			JAAD_{beh}				
	Visual Encoder	Global Context	Fusion Approach	Accuracy	AUC	F1 Score	Precision	Recall
Ours	VGG + GRU	✓	hybrid-fusion	0.62	0.54	**0.74**	0.65	**0.85**
Ablations								
Ours1	3D CNN	✓	later-fusion	0.59	0.53	0.69	0.65	0.75
Ours2	3D CNN	✓	early-fusion	0.59	0.54	0.69	0.65	0.74
Ours3	3D CNN	✓	hierarchical-fusion	0.57	0.48	0.70	0.62	0.81
Ours4	VGG + GRU	✗	later-fusion	0.59	0.51	0.72	0.63	0.83
Ours5	VGG + GRU	✓	later-fusion	**0.64**	**0.59**	0.73	**0.68**	0.78
Ours6	VGG + GRU	✓	early-fusion	0.60	0.56	0.70	0.67	0.73
Ours7	VGG + GRU	✓	hierarchical-fusion	0.54	0.50	0.64	0.63	0.65

TABLE VI
ABLATION STUDY ON THE JAAD ALL DATASET

Models	Model Variants			JAAD_{all}				
	Visual Encoder	Global Context	Fusion Approach	Accuracy	AUC	F1 Score	Precision	Recall
Ours	VGG + GRU	✓	hybrid-fusion	**0.83**	**0.82**	**0.63**	**0.51**	0.81
Ablations								
Ours1	3D CNN	✓	later-fusion	0.77	0.77	0.54	0.42	0.76
Ours2	3D CNN	✓	early-fusion	0.77	0.74	0.51	0.41	0.69
Ours3	3D CNN	✓	hierarchical-fusion	0.78	0.77	0.55	0.43	0.75
Ours4	VGG + GRU	✗	later-fusion	0.75	0.79	0.54	0.40	**0.85**
Ours5	VGG + GRU	✓	later-fusion	0.77	0.80	0.56	0.43	0.84
Ours6	VGG + GRU	✓	early-fusion	0.79	0.74	0.52	0.43	0.66
Ours7	VGG + GRU	✓	hierarchical-fusion	0.80	0.81	0.59	0.46	0.84

TABLE VII
ABLATION STUDY ON THE PIE DATASET

Models	Model Variants			PIE				
	Visual Encoder	Global Context	Fusion Approach	Accuracy	AUC	F1 Score	Precision	Recall
Ours	VGG + GRU	✓	hybrid-fusion	**0.89**	**0.86**	**0.80**	**0.79**	0.81
Ablations								
Ours1	3D CNN	✓	later-fusion	0.84	0.85	0.76	0.67	0.86
Ours2	3D CNN	✓	early-fusion	0.85	0.85	0.76	0.68	**0.87**
Ours3	3D CNN	✓	hierarchical-fusion	0.86	0.84	0.76	0.75	0.77
Ours4	VGG + GRU	✗	later-fusion	0.85	0.85	0.76	0.69	0.84
Ours5	VGG + GRU	✓	later-fusion	0.86	0.84	0.76	0.74	0.78
Ours6	VGG + GRU	✓	early-fusion	0.74	0.64	0.47	0.55	0.41
Ours7	VGG + GRU	✓	hierarchical-fusion	0.86	0.84	0.77	0.74	0.80

TABLE VIII
EFFECT OF LONGER PREDICTION HORIZON

Dataset	TTE	Acc.	AUC	F1	Precision	Recall
JAAD_{beh}	1-2s	**0.62**	**0.54**	**0.74**	**0.65**	**0.85**
	2-3s	0.53	0.47	0.65	0.62	0.68
JAAD_{all}	1-2s	**0.83**	**0.82**	**0.63**	**0.51**	**0.81**
	2-3s	0.79	0.78	0.57	0.46	0.76
PIE	1-2s	**0.89**	**0.86**	**0.80**	**0.79**	**0.81**
	2-3s	0.78	0.77	0.65	0.59	0.73

• The **bold** result indicates he best result among the models.

be seen from the table that the model performance drops on both the JAAD and PIE datasets. This supports the claims that a TTE of 1-2 seconds is more suitable than a TTE of 2-3 seconds.

E. Comparison of Different Prediction Task Configurations

There are some works that formulate the pedestrian intention prediction task in a different setting. Although using the same datasets, JAAD and PIE, they prepare the training, evaluation, and testing samples in a different way. The quantitative results cannot be directly compared with the proposed method. We analytically compared their results with ours with the conditions described. The comparison can be found in Table IX. The following works were compared:

- Liu's work [14] formulated the prediction task in two different perspectives of pedestrian-centric and location-centric settings. In addition to the JAAD dataset, it also introduces a specifically designed new dataset. The spatio-temporal information is encoded by GCN and RNN. They reported an accuracy of 0.77 on the JAAD dataset for predicting exactly 1 s into the future. We use the more commonly recognized benchmark proposed in [10] in our work. We achieved an accuracy of 0.83 on the JAAD dataset for future actions of 1-2 seconds. With a longer prediction horizon and higher accuracy scores, the effectiveness of our proposed method is validated.
- Zhang's work [36] focuses on pedestrian's crossing intention at red-light scenario. It analyzed 4 different machine learning models, SVM, RF, GBM, and XGBoost,

TABLE IX
COMPARISON OF DIFFERENT PEDESTRIAN INTENTION PREDICTION TASK CONFIGURATIONS

Work	Prediction Task Configuration	Results	Comparison to Ours
Liu et al. [14]	Both pedestrian-centric and location-centric configuration	Achieved 0.77 accuracy on JAAD dataset for action prediction at 1s	Achieved 0.83 accuracy on JAAD dataset for 1-2s future action prediction
Zhang et al. [36]	Pedestrian crossing intention at red-light scenario	Achieved 0.91 accuracy on self-created red-light scenario dataset	Unable to directly compare as the self-collected dataset is not accessible
Chen et al. [37]	Utlized a balanced sampling strategy, observing 15 frames (0.5s), predicting the action for 45 frames (1.5s)	Achieved 0.79 accuracy and 0.78 F1 score on randomly sampled testing set (balanced) using PIE dataset	Achieved 0.89 accuracy and 0.80 F1 score on PIE dataset with the commonly adopted configuration in [10]
Rasouli et al. [9]	Jointly predicting pedestrian action (intention), trajectory, and grid position	Achieved 0.91 accuracy with the auxiliary labels on PIE dataset	Achieved 0.89 accuracy with only action (intention) labels on PIE dataset

that are fed with pedestrian pose features. It achieved an accuracy of 0.91. However, the results were obtained on a self-collected and self-labeled red-light scenario dataset. Our proposed model cannot be directly compared with this method due to the inability of accessing the dataset and the different task configurations. Nevertheless, our method achieved an accuracy of 0.89 on the PIE dataset, which is very similar to Zhang's results.

- Chen's work [37] utilized a balanced sampling strategy to extract the samples for pedestrian crossing prediction. They use 15 frames (0.5 s) as observation to predict the pedestrian crossing action for 45 future frames (1.5 s). A graph convolutional autoencoder is used to embed spatio-temporal information. It achieved 0.79 accuracy and 0.78 F1 scores on a randomly sampled testing set (balanced) using the PIE dataset. Our model uses the commonly adopted configuration in [10]. We achieved 0.89 accuracy and 0.80 F1 score on the PIE dataset.
- Rasouli's work [9] formulates the pedestrian crossing intention as a sub-task of a multitask prediction framework, i.e., jointly predicting action (intention), trajectory, and grid position. They use a combined independent and joint encoding strategy with a categorical interaction module to fuse all the input channels. With the auxiliary labels, their work achieved 0.91 accuracy on the PIE dataset. As a comparison, our model achieved 0.89 accuracy on PIE but with only the action (intention) labels. Without auxiliary labels, our model still achieves comparable results. This validates the effectiveness of our model.

VI. CONCLUSION

In this work, we proposed a novel method for vision-based pedestrian crossing intention prediction. Our method explicitly considers the global context as a channel representing the interaction between the target pedestrian and the whole scene. We also proposed a hybrid fusion strategy for different features using 2D CNNs, RNNs, and attention mechanisms. Experiments on the JAAD and PIE datasets show that the proposed method achieves the state-of-the-art against baseline methods in the pedestrian action prediction benchmark.

Future work can focus on improving our model's robustness in unexpected situations, e.g., poor vision and occlusion. Additionally, feature fusion with more information sources can be explored. Finally, fine-tuning the model for particular pedestrian subsets, such as children and disabled people, can increase overall safety and performance.

REFERENCES

[1] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 206–213.

[2] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Do they want to cross? Understanding pedestrian intention for behavior prediction," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2020, pp. 1688–1693.

[3] J. Lorenzo, I. Parra, F. Wirth, C. Stiller, D. F. Llorca, and M. A. Sotelo, "RNN-based pedestrian crossing prediction using activity and pose-related features," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2020, pp. 1801–1806.

[4] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6262–6271.

[5] Z. Fang and A. M. López, "Is the pedestrian going to cross? Answering by 2D pose estimation," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2018, pp. 1271–1276.

[6] F. Piccoli *et al.*, "Fussi-Net: Fusion of spatio-temporal skeletons for intention prediction network," in *Proc. 54th Asilomar Conf. Sig., Syst., Comput.*, 2020, pp. 68–72.

[7] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian action anticipation using contextual feature fusion in stacked RNNs," in *Proc. 30th Brit. Mach. Vis. Conf.*, 2019.

[8] A. Rasouli, T. Yau, M. Rohani, and J. Luo, "Multi-modal hybrid architecture for pedestrian action prediction," 2020, *arXiv:2012.00514*.

[9] A. Rasouli, M. Rohani, and J. Luo, "Bifold and semantic reasoning for pedestrian behavior prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15 600–15 610.

[10] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1258–1268.

[11] J. Hariyono and K.-H. Jo, "Detection of pedestrian crossing road," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 4585–4588.

[12] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 304–311.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 7–9, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[14] B. Liu *et al.*, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 3485–3492, Apr. 2020.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *Syntax, Semantics Struct. Stat. Transl.*, pp. 103–111, 2014.

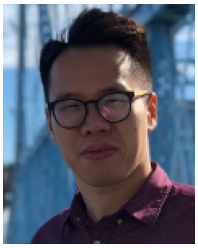
[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[18] K. Saleh, M. Hossny, and S. Nahavandi, "Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal DenseNet," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 9704–9710.

[19] K. Saleh, M. Hossny, and S. Nahavandi, "Spatio-temporal DenseNet for real-time intent prediction of pedestrians in urban traffic environments," *Neurocomputing*, vol. 386, pp. 317–324, 2020.

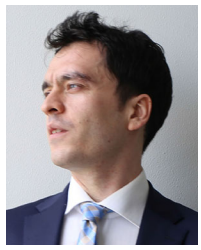
[20] M. Chaabane, A. Trabelsi, N. Blanchard, and R. Beveridge, "Looking ahead: Anticipating pedestrians crossing with future frames prediction," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2297–2306.

[21] P. Gujjar and R. Vaughan, "Classifying pedestrian actions in advance using predicted video of urban driving scenes," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 2097–2103.
[22] Z. Fang and A. M. López, "Intention recognition of pedestrians and cyclists by 2D pose estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4773–4783, Nov. 2020.
[23] Z. Wang and N. Papanikolopoulos, "Estimating pedestrian crossing states based on single 2D body pose," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, vol. 2, 2020, pp. 2205–2210.
[24] P. R. G. Cadena, M. Yang, Y. Qian, and C. Wang, "Pedestrian graph: Pedestrian crossing prediction based on 2D pose estimation and graph convolutional networks," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2019, pp. 2000–2005.
[25] A. Rasouli, M. Rohani, and J. Luo, "Bifold and semantic reasoning for pedestrian behavior prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15 600–15 610.
[26] A. Rasouli, T. Yau, P. Lakner, S. Malekmohammadi, M. Rohani, and J. Luo, "PePScenes: A novel dataset and baseline for pedestrian action prediction in 3D," 2020, *arXiv:2012.07773*.
[27] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[28] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3645–3649.
[29] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[30] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[31] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[33] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[35] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 900–918, Mar. 2020.
[36] S. Zhang, M. Abdel-Aty, Y. Wu, and O. Zheng, "Pedestrian crossing intention prediction at red-light using pose estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2331–2339, Mar. 2022.
[37] T. Chen, R. Tian, and Z. Ding, "Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3103–3109.



Dongfang Yang (Member, IEEE) received the B.E. degree in microelectronics from Sun Yat-sen University, Guangzhou, China, in 2014, and the M.S. and Ph.D. degrees in electrical and computer engineering from The Ohio State University, Columbus, OH, USA, in 2019 and 2020, respectively. He is currently a Senior Algorithm Engineer with Chongqnig Chang'an Automobile Company, Ltd. and a Postdoc Researcher with Chongqing University, Chongqing, China. His research interests include data analysis, machine learning, deep learning, and control, with applications in behavior prediction, decision-making, and motion planning of autonomous systems.

Haolin Zhang received the B.E. degree from the Shanghai University of Electric Power, Shanghai, China, and the M.S. degree from The Ohio State University, Columbus, OH, USA. His research interests include artificial intelligence, computer vision, and perception in the applications of intelligent vehicles.



Ekim Yurtsever (Member, IEEE) received the B.S. and M.S. degrees from Istanbul Technical University, Istanbul, Turkey, in 2012 and 2014, respectively, and the Ph.D. degree in information science from Nagoya University, Nagoya, Japan, in 2019. Since 2019, he has been a Research Associate with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA.

His research focuses on artificial intelligence, machine learning, computer vision, reinforcement learning, intelligent transportation systems, and automated driving systems.

Keith A. Redmill (Senior Member) received the B.S.E.E. and B.A. degrees in mathematics from Duke University, Durham, NC, USA, in 1989, and the M.S. and Ph.D. degrees from The Ohio State University, Columbus, OH, USA, in 1991 and 1998, respectively. Since 1999, he has been with the Department of Electrical and Computer Engineering, The Ohio State University, initially as a Research Scientist. He is currently a Research Associate Professor.

He has significant experience and expertise in intelligent transportation systems, intelligent and automated vehicle control and safety systems, sensors including computer vision, LiDAR, GNSS, IMU, and other sensing modalities, sensor fusion, wireless vehicle to vehicle communication, multi-agent systems including autonomous ground and aerial vehicles and robots, systems and control theory, virtual environment and dynamical systems modeling and simulator development, scientific computing, traffic monitoring and data collection, and real-time embedded and electromechanical systems.

Ümit Özgüner (Life Fellow, IEEE) is currently the TRC Inc. Chair on ITS with The Ohio State University (OSU), Columbus, OH, USA. He has authored or coauthored extensively on control design and vehicle autonomy and has coauthored one book on Ground Vehicle Autonomy. His current projects are on machine learning for driving, pedestrian modeling with OSU, and participates externally on V&V and risk mitigation, and self-driving operation of specialized vehicles. He has developed and taught a course on Ground Vehicle Autonomy for many years and has advised more than 35 students during their studies towards the Ph.D. He is a well known expert on Intelligent Vehicles. He holds the title of Fellow in IEEE for his contributions to the theory and practice of autonomous ground vehicles and is the Editor in Chief of the IEEE ITS Society, IEEE TRANSACTIONS ON INTELLIGENT VEHICLES. He has led and participated in many autonomous ground vehicle related programs, such as DoT FHWA Demo'97, DARPA Grand Challenges, and DARPA Urban Challenge. His research was/is supported by many industries, including Ford, GM, Honda, and Renault.