EfficientPIE: Real-Time Prediction on Pedestrian Crossing Intention with Sole Observation

Fang Qu[1], **Pengzhan Zhou**[1*], **Yuepeng He**[1], **Kaixin Gao**[1], **Youyu Luo**[1], **Xin Feng**[2], **Yu Liu**[3] and **Songtao Guo**[1]
[1]College of Computer Science, Chongqing University
[2]School of Computer Science and Engineering, Chongqing University of Technology
[3]Department of Computing, Hong Kong Polytechnic University
{fq, hyp, kaixingaocs, yyluo}@stu.cqu.edu.cn, {pzzhou, guosongtao}@cqu.edu.cn, xfeng@cqut.edu.cn, yu-y.liu@polyu.edu.hk

Abstract

Present Advanced Driving Assistance System (ADAS) responds to the dangerous crossing of pedestrians after the occurrence of the incident, occasionally causing severe accidents due to the stringent response window. Inference of pedestrian crossing intention may help vehicles operate in advance and enhance the safety of the vehicle by predicting the crossing probability. Recent studies usually ignore the demand of real-time forecast that required in the realistic driving scenario, and mainly focus on improving the model representation capacity on public datasets by increasing modality and observation time. Consequently, a new framework named EfficientPIE is proposed to predict the pedestrian crossing intention in real time with sole observation of the incident. To achieve reliable predictions, we propose incremental learning based on intention domain to relieve forgetting and promote performance with a progressive perturbation method. Our EfficientPIE outperforms all the SOTA models on two datasets PIE and JAAD, running nearly 7.4x faster than the previously fastest model. Our code is available at https://github.com/heinideyibadiaole/EfficientPIE.

1 Introduction

Traditional vehicle offers quite limited assistance for drivers to avoid accidents and requires driver's experiences and attention to prevent potential risks, which promotes the development of Advanced Driver Assistance System (ADAS). The enhanced perception ability of ADAS benefits from the recent evolution of AI technology, which are widely deployed in almost all the aspects of autonomous driving [Fabbri *et al.*, 2021; Phong *et al.*, 2023]. In order to avoid severe accidents involving pedestrians, many of recent works focus on the prediction of pedestrian crossing intention[Rasouli *et al.*, 2020; Kotseruba *et al.*, 2020].

Many traffic accidents happen when the crossing intention is neglected, such as pedestrians crossing from inconspicuous corners, revealing the importance of predicting the crossing behavior in advance. Even though the Autonomous Emergency Braking (AEB) may help ease this problem, it has drawbacks like that actions are taken only if the danger has already happened, which occasionally fails to prevent the accidents due to the quite stringent response window. Actually, crossing behavior can be successfully foreseen if the crossing intention of pedestrian and the risk of the crossing are estimated ahead, sparing more time to respond timely for the vehicles to avoid accidents. In this research scope, the dataset JAAD [Rasouli *et al.*, 2017] was first proposed to provide a benchmark for pedestrian crossing prediction. Afterwards, PIE [Rasouli *et al.*, 2019] was proposed to supplement the intention label for pedestrian crossing intention estimation.

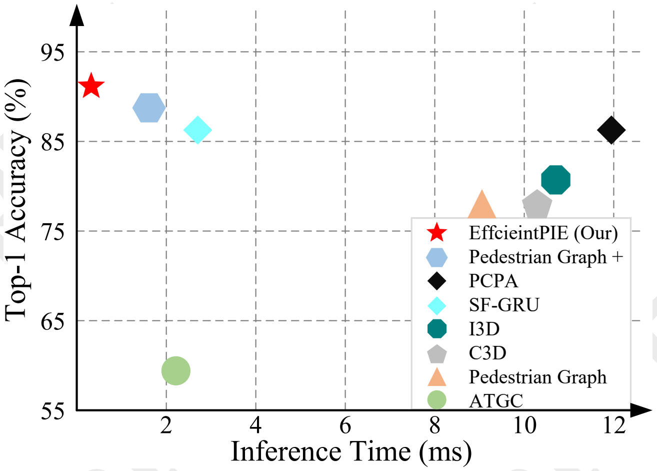


Figure 1: **PIE top-1 Accuracy vs. Inference Time.** Our proposed EfficientPIE achieves SOTA performance and runs 7.4x faster than the second best Pedestrian Graph +.

For common methods of PIE studies, crossing intention prediction consists of receiving related information, observing one specific pedestrian before the crossing event and predicting whether pedestrian has the intention to cross the road or not. For the purpose of improving the accuracy of crossing intention prediction, more input information is usually fed into model by providing more images and modalities, which

[*]Corresponding author.

also brings many drawbacks and prevents it from real deployment. For example, the base model proposed in PIE needs 15 consecutive frames as an input, which corresponds to cold-start time lasting for 0.5 sec. A recent framework obtained the superior performance at the expense of adding modalities including semantic segmentation map[Rasouli *et al.*, 2021], which could be limited in the realistic driving scenario, especially in face of emergencies[Li *et al.*, 2023]. Moreover, more complicated input means heavier computational cost, preventing the algorithms from being implemented in vehicles due to the limited resources and considerable latency.

In order to realize PIE in realistic autonomous driving, it is necessary to simplify the input, reduce the computational cost and shorten the latency. Though fewer images contain less information, [Cadena *et al.*, 2022] explores the influence of the number of input frames on the accuracy of two datasets PIE and JAAD, concluding that the accuracy does not attain significant reduction if the observation window is narrowed down. Inspired by this, we utilize the effectiveness of sole observation and presume that the feature of one frame could be sufficient to infer the crossing intention of pedestrians. Especially, it needs to be verified that, the trained model can obtain adequate representation capacity by observing the pedestrian just once, and extra images are redundant in this case if we intend to achieve the balance between accuracy and efficiency.

Motivated by this, we propose a new architecture named EfficientPIE to predict crossing intention of pedestrians in real time in this paper. Specifically, EfficientPIE focuses on exploiting implicit feature of pedestrians and local context effectively, excluding extra modalities and images. In addition, even though the standard convolution operation module helps obtain acceptable performance on accuracy, it could be more efficient if it is replaced by the depthwise separable convolution. For the purpose of solving catastrophic forgetting and getting a stable prediction, we are the first to apply incremental learning to crossing intention estimation. Inspired by progressive learning, progressive perturbation is proposed to enhance performance and generalization ability. Attributed to the above, we validate in the following sections that EfficientPIE achieves state-of-the-art accuracy of **92%** with the shortest inference time **0.21ms** on PIE as shown in Figure 1. Our contributions are mainly three-fold:

- We propose a real-time neural network architecture called EfficientPIE. To the best of our knowledge, EfficientPIE is the first framework to predict pedestrian crossing intention effectively using just one image, rather than a series of continuous images. Combined with object detection, the crossing intentions of all pedestrians in an image can be inferred.
- We propose incremental learning in the intention domain. Combined with a progressive perturbation method, EfficientPIE can exploit the feature of pedestrians more effectively to achieve stable performance.
- We conduct sufficient experiments on PIE and JAAD. The results demonstrate that EfficientPIE outperforms state-of-the-art models and strikes a good trade-off between efficiency and accuracy, which is crucial in realistic autonomous driving system.

2 Related Work

Pedestrian intention prediction is similar to pedestrian behavior prediction, but the application scenarios are different. In [Rasouli *et al.*, 2017], the authors realize that the danger can be prevented if the crossing event can be predicted and propose a public dataset JAAD, focusing on the crossing prediction. But JAAD has unbalanced distribution of samples where negative samples are obviously more than the positive. Due to the drawback of JAAD, PIE is proposed to provide a more balanced benchmark [Rasouli *et al.*, 2019] and officially defines the intention, which is the potential goal of pedestrians.

A recent study has established a human reference of intention estimation, proving that mankind is capable of understanding and predicting intention [Kotseruba *et al.*, 2020]. Based on the intuition that the crossing intention comes from the previous behaviors, base model is proposed by using RNN and the variants [Rasouli *et al.*, 2019], which is similar to video prediction. Furthermore, the study investigates the influence of the context information, demonstrating the necessity of surroundings.

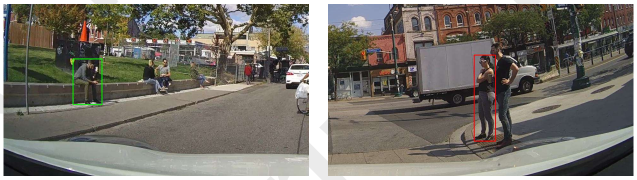
However, the base model from PIE inputs the feature of next time step into fully connected (FC) layer to obtain the classification, which exploits the temporal feature of image sequences ineffectively. Meanwhile, just simply connecting multiple modalities does not contribute to the prediction [Rasouli *et al.*, 2020], so that attention mechanism is added to capture the feature in connected input [Kotseruba *et al.*, 2021]. Due to the insufficient temporal feature extraction in base model, temporal attention mechanism and 3D convolution are also added to exploit the implicit temporal relationships. Moreover, BiPed [Rasouli *et al.*, 2021] investigates the representation ability of multi-modal feature and achieves state-of-the-art accuracy. Despite the accuracy has been improved, the model has higher computational cost. A recent study intends to design a fast framework based on GCN to predict the intention in a fast speed [Cadena *et al.*, 2022], but the input includes human key points which are generated in advance. Even if the generation of human key points is included, the whole inference speed does not gain significant promotion according to their experiments.

Overall, the previous works have not explored an efficient framework for the pedestrian intention prediction tasks since the feature extraction of the continuous images has heavy computational cost. Therefore, the computation can be accelerated from fewer input images and modalities, and less utilization of RNN, leading to the research of our EfficientPIE.

3 Methodology

3.1 Problem Formulation

We formulate the pedestrian crossing intention prediction as an image classification process in which the objective is to classify whether the concerned pedestrian is crossing or not. Given the video containing sequential image samples and the corresponding bounding boxes [Rasouli *et al.*, 2017; Rasouli *et al.*, 2019] which can be expressed as $X_t = \{x_{1t}, x_{2t}, \ldots, x_{it}\}$ and $B_t = \{b_{1t}, b_{2t}, \ldots, b_{it}\}$, the predictions and labels can be formulated as $\hat{Y}_t$ and Y_t. Note that,



(a) Negative

(b) Positive

Figure 2: An image from the PIE dataset (cropped for better visibility). Green bounding box represents the absence of crossing intention, while the red one represents the presence of crossing intention.

	PIE		JAAD	
	Accuracy	AUC	Accuracy	AUC
Single	0.873	0.869	0.844	0.818
Multiple	0.871	0.872	0.845	0.822

Table 1: The difference of single image and multiple images.

each input focuses on one pedestrian, and our input only needs one image. So we choose a key image x_{it} before the happening of the crossing event to predict, where the value of i means the i-th crossing event and the value of t represents the time step which varies from 0 to 14.

3.2 Effectiveness of Methods

In previous work, history images are fed into a model composed of RNN and FC layers to achieve higher performance [Rasouli *et al.*, 2020; Rasouli *et al.*, 2021]. The intuition of the methods is that only if enough observed images can be used to infer the intention.

However, we found that the crossing intention may be predicted through analyzing just one image. Intuitively, the upcoming behavior could be speculated from local context in conjunction with pedestrian motion. Furthermore, when a snapshot is taken in the realistic driving scenario, the previous states are related to the current frame, thus can be inferred by the actions of the pedestrians. Therefore, the information of one frame is sufficient to derive the prediction. For example, as shown in Figure 2(a), the pedestrian labelled by a green bounding box is highly likely to keep sitting, which results in the lack of crossing intention. Vice versa, While owing to the standing pose of the pedestrian labelled by red bounding box and the presence of crossroad, the crossing intention is more obvious to predict in Figure 2(b).

As shown in Table 1, the accuracy of different input types is nearly consistent, indicating the redundancy of multiple images and the reliability of the proposed method. Due to the similarity of sequential images in realistic scenarios, the predictions of different input frames (i.e. indexed with $0, 2, 4, \ldots, 14$) show little difference, which is shown in Table 2. A possible explanation of this observation may be that the movement of most pedestrians is modest and continuous, thus any frame making no huge difference in indicating the crossing intention within the time window of 15 consecutive frames lasting for just 0.5 seconds.

	Accuracy							
	0	2	4	6	8	10	12	14
PIE	0.869	0.870	0.868	0.869	0.872	0.873	0.871	0.873
JAAD	0.835	0.836	0.841	0.838	0.837	0.843	0.844	0.844

Table 2: The difference of key image.

3.3 Architecture

Owing to the effectiveness demonstrated previously, EfficientPIE does not include any RNN, which is the largest difference compared with other models used for PIE tasks. In contrast, to exploit the image feature, EfficientPIE is only composed of convolution operation modules. The Mobile inverted bottleneck MBConv, which is optimized with squeeze-and-excitation mechanism [Hu *et al.*, 2018], shows excellent performance in image classification [Tan and Le, 2019]. Furthermore, the utilization of Fused-MBConv improves the efficiency and accuracy in [Tan and Le, 2021]. For rapider computation, the squeeze and excitation mechanism is removed from Fused-MBConv. Built upon the two types of convolution operation modules, EfficientPIE consists of six blocks. As shown in Figure 3, the input of EfficientPIE is a 300×300 pixels cropped figures, and the output is the predicted crossing intention of pedestrians within the figure. In between, it contains two Common-Conv, two Fused-MBConv, and two MBConv blocks.

Depthwise Separable Convolutions

In order to compute the prediction in a faster inference speed, it is important to use more efficient operation modules. The standard convolution, which is often used in CNNs, filters and combines inputs into a new set of outputs in one step. However, compared to depthwise separable convolution [Zhang *et al.*, 2018; Howard *et al.*, 2017; Chollet, 2017], it has been proved being inefficient owing to the redundant computation.

Formally, suppose $D_K \times D_K$ is the size of the convolution kernel, M is the number of input channels, N is the number of output channels and $D_F \times D_F$ is the spatial dimension of the output feature map. Thus, for the output feature map, number of parameters and computational cost of standard convolution layer are as follows respectively:

$$D_K \cdot D_K \cdot M \cdot N, \tag{1}$$

$$D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F. \tag{2}$$

The operation can change both channel dimension and spatial dimension. However, depthwise separable convolution splits the filtering and combination steps into two steps, consisting of depthwise convolution and pointwise convolution. Depthwise convolution applies a single filter to each input channel in order to change spatial dimension and maintain channel dimension. It has the number of parameters and computational cost of:

$$D_K \cdot D_K \cdot M, \tag{3}$$

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F. \tag{4}$$

Then, pointwise convolution, which is a simple 1×1 convolution, is used to combine the output of the depthwise convolution to derive the new feature map. The parameters and the

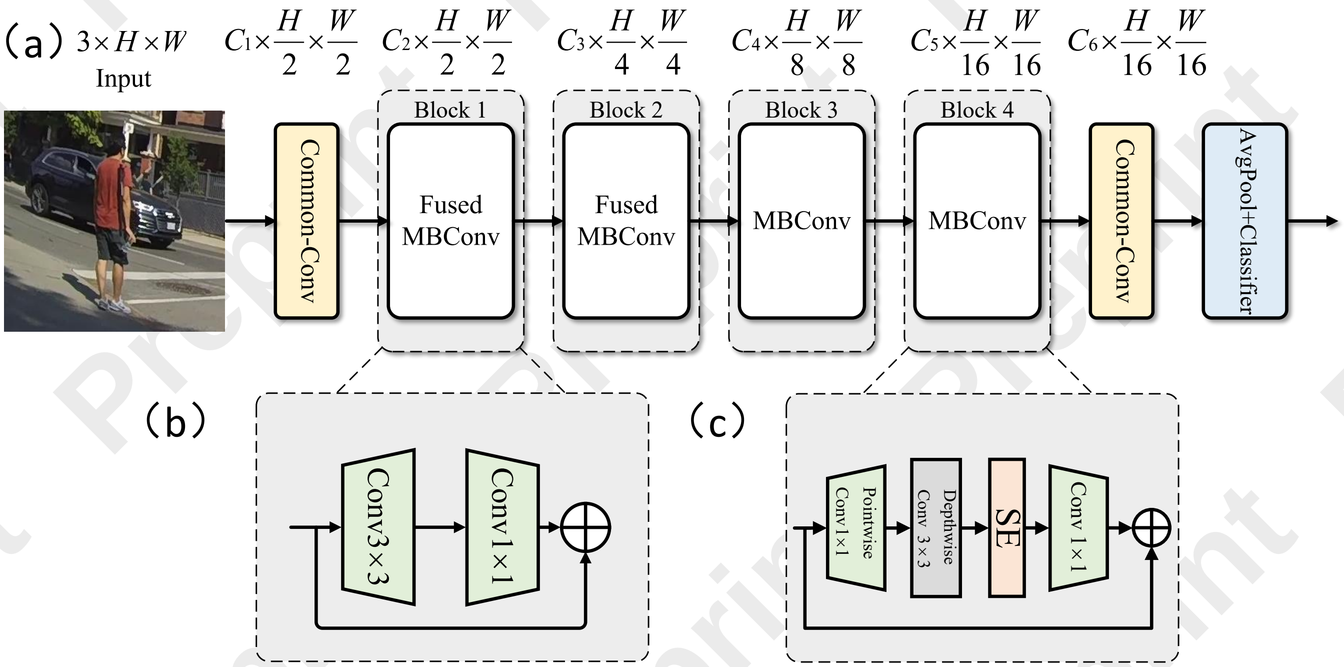


Figure 3: Overview of architecture.(a) Architecture of EfficientPIE; (b) Fused MBConv module; (c) MBConv module.

computational cost of it can be expressed respectively as:

$$M \cdot N, \tag{5}$$

$$M \cdot N \cdot D_F \cdot D_F. \tag{6}$$

Thus, the depthwise separable convolution, which is made up of depthwise convolution and pointwise convolution, has the number of parameters and the computational cost of:

$$D_K \cdot D_K \cdot M + M \cdot N \tag{7}$$

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F \tag{8}$$

Therefore, if standard convolution is replaced with depthwise separable convolution, the number of parameters and the computational cost could be reduced as follows:

$$\frac{D_K \cdot D_K \cdot M + M \cdot N}{D_K \cdot D_K \cdot M \cdot N} = \frac{1}{N} + \frac{1}{D_K \cdot D_K} \tag{9}$$

$$\begin{aligned} &\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F} \\ =&\frac{1}{N} + \frac{1}{D_K \cdot D_K} \end{aligned} \tag{10}$$

Usually, D_K is set to 3 and N is much larger than D_K. Therefore, the cost of depthwise separable convolution is nearly $1/9$ of the cost of standard convolution theoretically. Due to its excellent efficiency, the depthwise separable convolution serves as an important computational component of EffcientPIE.

3.4 Intention Domain Incremental Learning

EfficientPIE only needs 1/15 of all samples in our case to finish each training procedure, giving the possibility to enhance the performance by learning with the remaining 14/15 of whole datasets. However, although simply fine-tuning by training all samples is an acceptable method to obtain higher accuracy, the predictions of old tasks come with catastrophic forgetting, which contributes to unstable and unreliable outputs. For the sake of outputting reliable intentions, we propose an intention domain incremental learning approach and this is the first time that incremental learning is applied in crossing intention prediction.

Inspired by [Mishkin and Matas, 2015], the model is pretrained on Imagenet [Deng *et al.*, 2009] and trained on the first 1/15 of the two datasets as the starting step. Then the process of the incremental learning can be denoted as:

$$\hat{Y}_{t-1} = \text{EfficientPIE}(X_t, \theta_s, \theta_{t-1}) \tag{11}$$

$$\hat{Y}_t = \text{EfficientPIE}(X_t, \theta_s, \theta_t) \tag{12}$$

$$\theta_s^*, \theta_{t-1}^*, \theta_t^* = \text{argmin}(\mathcal{L}_a) \tag{13}$$

$$\mathcal{L}_a = \begin{cases} \lambda_{t-1}\mathcal{L}_{t-1}(Y_{t-1}, \hat{Y}_{t-1}) & \\ +\mathcal{L}_t(Y_t, \hat{Y}_t), & \text{if } \mathcal{L}_t > \mathcal{L}_{t-1}, \\ \mathcal{L}_t(Y_t, \hat{Y}_t), & \text{otherwise,} \end{cases} \tag{14}$$

where θ_s is the parameter of the feature extraction layers of EfficientPIE, which is shared by the two models. θ_{t-1} and θ_t are the parameters of the classifier layer of the old model and the new model respectively. But unlike [Li and Hoiem, 2017], the parameters of all layers are optimized by an adaptive loss function, which varies during the training procedure to prevent the loss from getting stuck in the local optimal solution of the task of the previous time step.

3.5 Progressive Perturbation

The crossing intention of a pedestrian may also change during the crossing event, resulting in dynamic probability of crossing, which can be recognized as the uncertainty of intention.

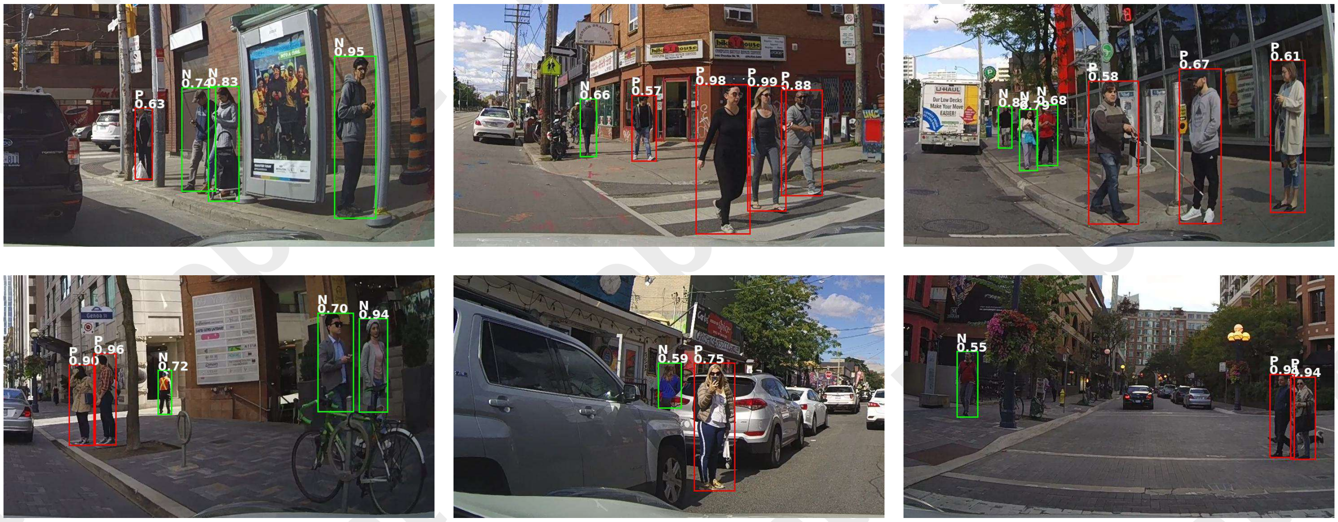


Figure 4: Detection results on several typical samples. “P” (i.e. “Positive”) represents crossing, and “N” (i.e. “Negative”) represents not crossing. The numbers below the characters P or N represent the probability of the corresponding class.

Since the crossing intention is used as the label, the performance could improve if the uncertainty of intention is utilized effectively. To validate the observation, a robust perturbation method is proposed to capture the implicit uncertainty before computing the loss. Inspired by [Tan and Le, 2021], a noise which is generated progressively is utilized to modify the prediction module before back propagation as follows:

$$\hat{Z} = \hat{Y} + \vec{\delta}, \tag{15}$$

where $\hat{Y}$ and $\vec{\delta}$ denote the predictions and the noises respectively. $\hat{Z}$ is the modified prediction after perturbation. It is composed of two elements generated randomly:

$$\vec{\delta} = [p_j \quad -p_j], \tag{16}$$

$$p_j = \text{rand}(-\frac{mj}{E}, \frac{mj}{E}), j \in [0, E], j \in N, \tag{17}$$

where p_j represents the seed of noise. m is the adjustable perturbation level. j and E denote the current epoch and total number of epochs respectively. As summarized in algorithm 1, the neural network is trained with weak perturbation to learn primary representations in the early training epochs and the perturbation will be enhanced gradually to promote the performance. Meanwhile, since the accuracy is possible to decrease if the image size is changed dynamically [Hoffer *et al.*, 2019], our algorithm maintain the image size to apply stronger performance to the network.

Algorithm 1 Progressive Perturbation

```
Input: Initialize the perturbation level $m$, the image $X$ and label $Y$, number of total training epochs E
Output: Parameters of trained model
for $j = 0$ to $E - 1$ do
    Train the model and compute prediction $\hat{Y}$
    Compute the perturbation $\vec{\delta}$ using Eqns. 16 and 17
    Alter the prediction to obtain $\hat{Z}$ using Eq. 15
    Compute the loss $\mathcal{L}$ based on the altered $\hat{Z}$ and $Y$
    Back propagation and update the network parameters
end for
```

4 Experiments

4.1 Datasets

To validate our method, we use the **Pedestrian Intention Estimation (PIE)** and **Joint Attention in Autonomous Driving (JAAD)** as the main dataset for experiment.

JAAD is a large dataset for pedestrian crossing prediction, which is composed of recorded video clips. It has 3955 training sequences while the insufficient positive samples prevents models from learning representation of crossing intention. Attributed by the weakness, PIE is proposed to provide more positive samples, having 3980 training sequences and 995 of them are crossing events, which is made up of 6 hours video footage of pedestrians in Toronto, Canada. Compared to JAAD, PIE are generated from longer and continuous videos and focus more on the pedestrian samples that are likely to cross the road. Specifically, the positive samples of PIE are more sufficient than JAAD, which is beneficial to the model to capture the semantic pattern of crossing event. Moreover, PIE provides the pedestrian intention label while JAAD uses the crossing action label as the substitute. Both datasets provide bounding box annotations for each concerned pedestrian.

The pedestrian tracks generated in the same way as [Rasouli *et al.*, 2019] and the tracks are clipped with an overlap ratio of 0.5. After the clipping, JAAD has 40046 samples and PIE has 19086 samples, which are all taken before the happening of crossing event. To compute the intention more efficiently, we choose the last frame of samples and crop the image to 300 × 300 around the labelled pedestrian as input.

		PIE				JAAD			
Method	**Infer-time(ms)**	Accuracy	AUC	F1	Precision	Accuracy	AUC	F1	Precision
ATGC	2.19	0.59	0.55	0.36	0.35	0.64	0.60	0.53	0.50
Pedestrian Graph	9.07	0.76	0.69	0.48	0.62	0.80	0.84	0.55	0.46
C3D	10.32	0.77	0.67	0.52	0.63	0.84	0.81	0.65	0.57
I3D	10.64	0.79	0.75	0.64	0.61	0.82	0.75	0.55	0.49
SingleRNN	-	0.81	0.75	0.64	0.67	0.78	0.75	0.54	0.44
StackedRNN	-	0.82	0.78	0.67	0.67	0.79	0.79	0.58	0.46
MultiRNN	-	0.83	0.80	0.71	0.69	0.79	0.79	0.58	0.45
MM-LSTM	-	0.84	0.84	0.75	0.68	0.80	0.77	0.58	0.51
SF-GRU	2.65	0.86	0.83	0.75	0.73	0.83	0.77	0.58	0.51
PCPA	11.89	0.86	0.84	0.76	0.73	0.83	0.77	0.57	0.50
MMHA	-	0.89	0.88	0.81	0.77	0.84	0.80	0.62	0.54
Pedestrian Graph +	1.56	0.89	0.90	0.81	0.83	0.86	**0.88**	0.65	0.58
BiPed	-	0.90	0.90	0.84	0.80	0.83	0.79	0.60	0.52
DPCIAN	5.86	0.91	0.88	0.83	0.83	**0.89**	0.77	0.59	0.61
PIT	4.80	0.91	0.90	0.82	0.85	0.87	0.87	**0.66**	0.54
EfficientPIE	**0.21**	**0.92**	**0.92**	**0.95**	**0.96**	**0.89**	0.86	0.62	**0.63**

Table 3: Performance of EffcientPIE and the previous models tested on PIE and JAAD datasets.

		PIE	**JAAD**
IDIL	**Perturbation**	Accuracy	Accuracy
×	×	0.87	0.84
×	✓	0.88	0.85
✓	×	0.91	0.88
✓	✓	**0.92**	**0.89**

Table 4: The impact of different variations on EfficientPIE.

4.2 Setup

EfficientPIE is trained and evaluated with an NVIDIA RTX 3090 GPU. Before the input is computed, the image is applied a random horizontal flip and a color transformation to augment multiplicity of samples. RMSProp optimizer is used with weight decay $1e-4$, set learning rate to $1e-5$ and apply cosine annealing algorithm to decrease the learning rate, which are possible to contribute to the performance [He *et al.*, 2019]. The model is trained for 50 epochs and batch size is set to be 32. The training setting for two datasets is absolutely identical.

To demonstrate the efficiency of EfficientPIE, except evaluating the common metrics including accuracy, precision, F1 and Area Under Curve (AUC), the inference time is also a necessary metric and should be tested strictly. Before measuring inference time, the GPU is warmed up to utilize the computational resource by loading several random samples. Then, the total inference time is recorded in the procedure of computing the prediction of 128 random selected samples. This procedure is repeated 100 times and the final inference time is calculated by the average of the total inference time during the whole procedure.

4.3 Results

Table 3 summarizes the quantitative results of experiments on PIE and JAAD. We conduct sufficient experiments by testing ATGC [Rasouli *et al.*, 2017], Pedestrian Graph [Cadena *et al.*, 2019], C3D [Tran *et al.*, 2015], I3D [Carreira and Zisserman, 2017], SingleRNN [Kotseruba *et al.*, 2020], StackedRNN [Yue-Hei Ng *et al.*, 2015], MultiRNN [Bhattacharyya *et al.*, 2018], MM-LSTM [Aliakbarian *et al.*, 2018], SF-GRU [Rasouli *et al.*, 2020], PCPA [Kotseruba *et al.*, 2021], MMHA [Rasouli *et al.*, 2022], Pedestrian Graph + [Cadena *et al.*, 2022], BiPed [Rasouli *et al.*, 2021], DPCIAN [Yang *et al.*, 2023], PIT [Zhou *et al.*, 2023] and EfficientPIE.

ATGC is the first model to predict pedestrian crossing intentions, but it does not solve the problem effectively and only achieves an accuracy of 0.59 on PIE. I3D and C3D are 3D convolution neural networks for action recognition but show evidently higher latency than most models, since their input are videos. SingleRNN, StackedRNN, MultiRNN, MM-LSTM and SF-GRU are all based on common RNN operation modules and express insignificantly different performance due to the similar feature extraction procedures. Being aware of the positive contribution of different input modalities, PCPA, MMHA, BiPed, DPCIAN promote their accuracy by using multi-modality, and DPCIAN achieves the accuracy of 0.91. However, the increase in accuracy actually comes at the expense of inference speed, contributing the longest inference time of 38 ms for PCPA. Therefore, Pedestrian Graph + tries to predict the intention more quickly and achieves the inference time of 1.56 ms. However, the speed does not include the time it takes to generate the input data such as gesture key points.

In summary, our model outperforms all the existing models, achieving state-of-the-art performance on both PIE and JAAD. Due to the architecture, EfficientPIE only needs **0.21ms** to compute the prediction, running nearly **7.4x faster** than the previous fastest model. Moreover, compared with prior models, the performance of EfficientPIE demonstrates that only sole observation is enough to apply superior performance to models by convolution operation modules.

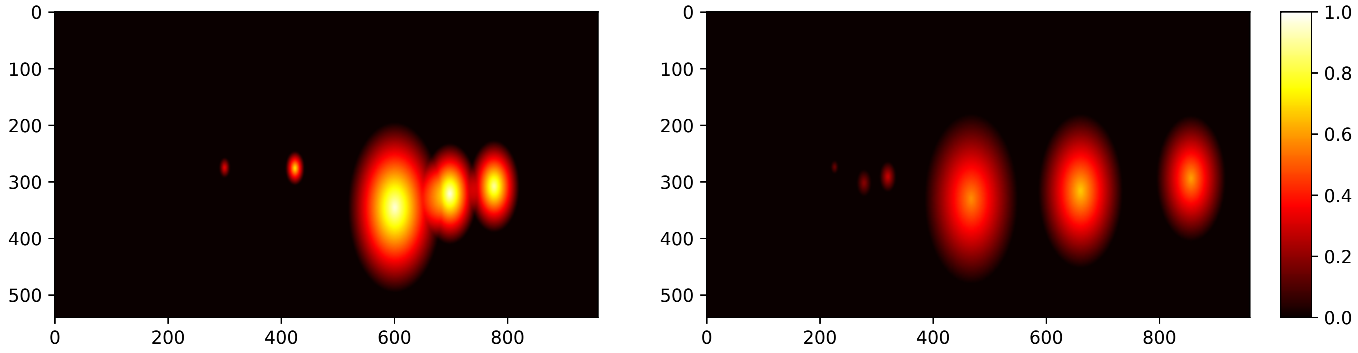


Figure 5: Risk analyses on cases, where the value represents the risk degree. The heat maps are generated from the second and the third figures on the first row of Figure 4 respectively.

4.4 Ablation Study

We conduct experiments to test the proposed methods and follow the same training setup. As shown in Table 4, according to the accuracy of 0.87 on PIE and 0.84 on JAAD, our base model already has an acceptable performance to predict the intention. Meanwhile, optimizing the parameters through the adaptive loss function, EfficientPIE is more capable of capturing higher semantic patterns and derives the improvement of 4.6% on PIE. The proposed perturbation method contributes less than the incremental learning in the intention domain, since perturbation denotes the variance of intention while continual learning captures more natural features. By applying both transfer learning and perturbation, EfficientPIE improves the accuracy by 5.7% and 6.0% on PIE and JAAD respectively.

4.5 Detection

The crossing intention prediction need the bounding box to locate the concerned pedestrian to forecast the crossing intention. Although the previous model can use object detection to generate the boxes, the stringent observation windows prevents those models from implementing the prediction successfully in autonomous driving. However, on account of the advantage of sole observation, EfficientPIE is qualified to implement real-time inference of all pedestrians' crossing intention in an image with the help of YOLO. Since each training sample only focuses on one pedestrian, most of the pedestrians in images are not considered in the training procedure. But as shown in Figure 4, the intention of most pedestrians from the cases is predicted accurately, showing sufficient generalization ability of EfficientPIE. According to the results, we draw two conclusions as follows.

First, intention is influenced by the combination of the pedestrian and the position. In the first figure of Figure 4 , the pedestrians near the bus stop are predicted "not crossing" due to the pose and the position. Especially, the male standing with his back to the driver has the highest probability of "not crossing", which is consistent with human driver's intuition. In the second figure, the persons who appear at crosswalk tend to have the crossing intention, meaning that the classification of them is mainly based on the zebra crossing.

Second, the intention has the infectivity. The pedestrians tend to have the same intention as the nearby pedestrians if they express the apparent intention, which can be demonstrated in the third and the fourth figure. There is a blind man in the third figure whose intention is predicted to be positive. Despite it is easy to understand that he does not plan to cross the road, the intention prediction is affected by the infectivity of intention of his nearby pedestrians.

4.6 Risk Analyses

In the realistic driving scenario, the attention of drivers tends to be paid to the implicit risk rather than a certain pedestrian among the crowd, accounting for their defensive driving style. This means the risk degree of an area may be more important than the crossing intention of a specific pedestrian from their perspective.

Consequently, we visualize the detection results by using the heat map in Figure 5 and further show the risk generated from the crossing intention of the second and the third case of Figure 4. Apparently, comparing the two heat maps, the left scene is more dangerous than the right owing to the intenser crossing intention, with brighter color and denser circles on the map. Note that, given that the pedestrians having no crossing intention still contribute to the implicit risk, the probabilities of negative are converted to positive by replace them with the absolute value of the difference from 1. In other words, the streets which exist pedestrians are more dangerous than those without pedestrians, explaining the existence of the tiny red dots in the heat map.

5 Conclusion

In this paper, we introduce a new efficient framework named EfficientPIE for pedestrian intention prediction. According to the experiments on public benchmarks PIE and JAAD, our model achieves state-of-the-art performance and runs nearly 7.4x faster than prior fastest model owing to the sole observation and the efficient architecture design. Combined with pedestrian detection, experiments illustrate that all the crossing intention of pedestrians in an image can be inferred thoroughly, which also reveal that intention has infectivity and depends on the pedestrians' position. In future work, we are interested in incorporating scene classification to further enhance the model generalization ability.

Acknowledgments

The work in this paper was supported by the National Natural Science Foundation of China under Grant 62202072, 62272069, and supported by the Project No. 2024CDJGF-036 of the Fundamental Research Funds for the Central Universities. This work was also supported by the Natural Science Foundation of Chongqing, China, under Grant No. CSTB2022NSCQMSX0493, and the key project of Humanities and Social Sciences under the Chongqing Ministry of Education with Grant No. 24sKD134.

References

[Aliakbarian *et al.*, 2018] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Viena: A driving anticipation dataset. In *Asian Conference on Computer Vision*, pages 449–466, 2018.

[Bhattacharyya *et al.*, 2018] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4194–4202, 2018.

[Cadena *et al.*, 2019] Pablo Rodrigo Gantier Cadena, Ming Yang, Yeqiang Qian, and Chunxiang Wang. Pedestrian graph: Pedestrian crossing prediction based on 2d pose estimation and graph convolutional networks. In *IEEE Intelligent Transportation Systems Conference*, pages 2000–2005, 2019.

[Cadena *et al.*, 2022] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. Pedestrian graph+: A fast pedestrian crossing prediction model based on graph convolutional networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):21050–21061, 2022.

[Carreira and Zisserman, 2017] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Fabbri *et al.*, 2021] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10849–10859, 2021.

[He *et al.*, 2019] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019.

[Hoffer *et al.*, 2019] Elad Hoffer, Berry Weinstein, Itay Hubara, Tal Ben-Nun, Torsten Hoefler, and Daniel Soudry. Mix & match: training convnets with mixed image sizes for improved accuracy, speed and scale resiliency. *arXiv preprint arXiv:1908.08986*, 2019.

[Howard *et al.*, 2017] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[Kotseruba *et al.*, 2020] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Do they want to cross? understanding pedestrian intention for behavior prediction. In *IEEE Intelligent Vehicles Symposium*, pages 1688–1693, 2020.

[Kotseruba *et al.*, 2021] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Benchmark for evaluating pedestrian action prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1258–1268, 2021.

[Li and Hoiem, 2017] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[Li *et al.*, 2023] Rongqing Li, Changsheng Li, Dongchun Ren, Guangyi Chen, Ye Yuan, and Guoren Wang. Bcdiff: Bidirectional consistent diffusion for instantaneous trajectory prediction. In *Advances in Neural Information Processing Systems*, pages 14400–14413, 2023.

[Mishkin and Matas, 2015] Dmytro Mishkin and Jiri Matas. All you need is a good init. *arXiv preprint arXiv:1511.06422*, 2015.

[Phong *et al.*, 2023] Tran Phong, Haoran Wu, Cunjun Yu, Panpan Cai, Sifa Zheng, and David Hsu. What truly matters in trajectory prediction for autonomous driving? In *Advances in Neural Information Processing Systems*, pages 71327–71339, 2023.

[Rasouli *et al.*, 2017] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017.

[Rasouli *et al.*, 2019] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6262–6271, 2019.

[Rasouli *et al.*, 2020] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked rnns. *arXiv preprint arXiv:2005.06582*, 2020.

[Rasouli *et al.*, 2021] Amir Rasouli, Mohsen Rohani, and Jun Luo. Bifold and semantic reasoning for pedestrian behavior prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15600–15610, 2021.

[Rasouli *et al.*, 2022] Amir Rasouli, Tiffany Yau, Mohsen Rohani, and Jun Luo. Multi-modal hybrid architecture for pedestrian action prediction. In *IEEE intelligent Vehicles symposium*, pages 91–97, 2022.

[Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[Tan and Le, 2021] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021.

[Tran *et al.*, 2015] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[Yang *et al.*, 2023] Biao Yang, Zhiwen Wei, Hongyu Hu, Rui Wang, Changchun Yang, and Rongrong Ni. Dpcian: A novel dual-channel pedestrian crossing intention anticipation network. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[Yue-Hei Ng *et al.*, 2015] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702, 2015.

[Zhang *et al.*, 2018] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.

[Zhou *et al.*, 2023] Yuchen Zhou, Guang Tan, Rui Zhong, Yaokun Li, and Chao Gou. Pit: Progressive interaction transformer for pedestrian crossing intention prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2023.