Gating Syn-to-Real Knowledge for Pedestrian Crossing Prediction in Safe Driving

Jie Bai, Jianwu Fang, *Member, IEEE,* Yisheng Lv, *Senior Member, IEEE,* Chen Lv, *Senior Member, IEEE,* Jianru Xue, *Member, IEEE*, and Zhengguo Li, *Fellow, IEEE*

***Abstract*—Pedestrian crossing prediction (PCP) in driving scenes plays a critical role in ensuring the safe decision of intelligent vehicles. Due to the limited observations and annotations of pedestrian crossing behaviors in real situations, recent studies have begun to leverage synthetic data with flexible variation to boost prediction performance, employing domain adaptation frameworks. However, different domain knowledge has distinct cross-domain distribution gaps, which necessitates suitable domain knowledge adaption ways for PCP tasks. In this work, we propose a gated syn-to-real knowledge transfer approach for PCP (Gated-S2R-PCP), which has two aims: 1) designing the suitable domain adaptation ways for different kinds of crossing-domain knowledge, and 2) transferring suitable knowledge for specific situations with gated knowledge fusion. Specifically, we design a framework that contains three domain adaption methods including style transfer, distribution approximation, and knowledge distillation for various information, such as visual, semantic, depth, bounding boxes, *etc*. A learnable gated unit (LGU) is employed to fuse suitable cross-domain knowledge to boost pedestrian crossing prediction. We construct a new synthetic benchmark S2R-PCP-3181 with 3181 sequences (489,740 frames) which contains the pedestrian bounding boxes, RGB frames, semantic segmentation maps, and depth maps. With the synthetic S2R-PCP-3181, we transfer the knowledge to two real challenging datasets of PIE and JAAD, and superior PCP performance is obtained to the state-of-the-art methods.**

***Index Terms*—Pedestrian crossing prediction, safe driving, domain adaptation, gated network, Timesformer**

I. Introduction

THE new road safety era from 2025 is fully automated safety features [1], prioritized for the most vulnerable road users [2]. To enhance safe driving, this work targets the pedestrian crossing prediction (PCP) problem because pedestrian crossing is one of the most typical behaviors that compete for road rights with vehicles [3], [4], [5], [6].

PCP attracts increasing attention with the help of large-scale benchmarks [7], [8], [9], facilitating various data-driven deep PCP models. Pedestrian poses, bounding boxes (*abbrev.*, ***bboxes*** in this work), vehicle velocity, semantic segmentation maps, *etc.*, have all been explored. However, it is difficult to annotate these clues in real situations and the severe weather and light conditions also pose challenges to PCP. Therefore, sample diversity is vital to boost the generalization of PCP, and borrowing synthetic data is encouraged via domain adaptation models [10], [11], [12], [13], [14]. This involves the fundamental parallel verification of synthetic and real data for PCP [15], whereas the synthetic-real data shift is boring with varying modalities. Naturally, the synthetic (*more information*)-to-real (*limited information*) knowledge transfer is straightforward for PCP. The most related ones to our work are [10] and [11] with a syn-to-real pedestrian bbox fine-tuning strategy and syn-to-real knowledge distillation framework, respectively. However, the domain gaps are not explicitly investigated.

To illustrate the distinct domain gaps for different informa-

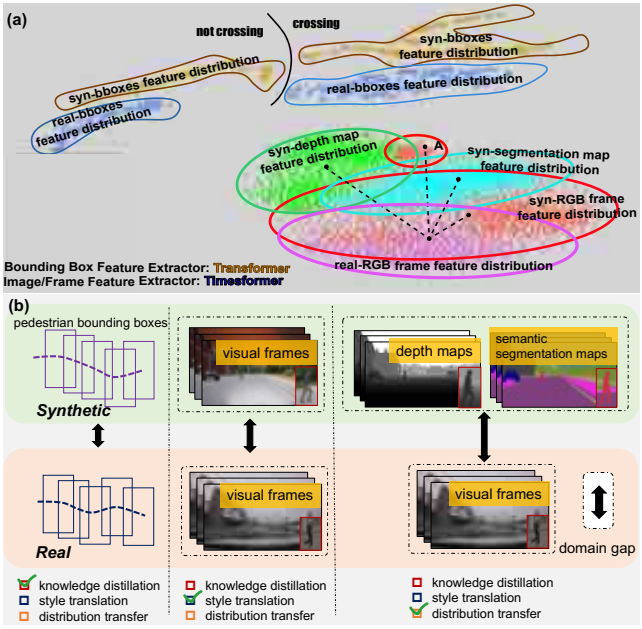

Fig. 1. Illustration for the suitable syn-to-real knowledge transfer for different kinds of information in driving scenes. (a) plots the feature distributions of real and synthetic datasets, for the pedestrian bounding boxes, RGB frames, semantic segmentation maps, and depth maps. (b) illustrates the differentiated syn-real knowledge transfer with different domain gaps. The feature distribution in (a) is obtained by t-SNE [16] on the feature vectors of 1000 randomly selected samples (each sample has 16 frames) in the synthetic PCP dataset (Syn-PCP-3181, to be described in Sec. IV-A) and a real PCP dataset (PIE [8]). Notably, because of the different input shapes of pedestrian bboxes and images, the feature vectors of pedestrian bboxes and image-like inputs are extracted by Transformer [17] and Timesformer [18], respectively.

This work is supported by the National Natural Science Foundation of China under Grants 62036008 and 62273057, and also supported by Shaanxi Qinchuangyuan "Scientist and Engineer" Project (2022KXJ-022).

J. Fang and J. Xue are with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China ({fangjianwu,jrxue}@mail.xjtu.edu.cn).

J. Bai is with the College of Transportation Engineering, Chang'an University, Xi'an, China.

Y. Lv is with the Institute of Automation, Chinese Academy of Sciences, Beijing, China (yisheng.lv@ia.ac.cn).

C. Lv is with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore (lyuchen@ntu.edu.sg).

Z. Li is with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore (ezgli@i2r.a-star.edu.sg).

tion in the PCP task, Fig. 1a plots the feature distributions of pedestrian bboxes, RGB frames, semantic segmentation and depth maps. We can observe that the pedestrian bbox feature distributions in synthetic and real datasets own a similar manifold and close intra-distance for crossing or not crossing behaviors. The feature distribution of syn-RGB frames covers the largest range and correlates with the ones of syn-depth maps, syn-segmentation maps, and real RGB frames. Notably, the subset **A** may be generated by the similar contour or object shapes of the scenes in these syn-images. The domain gaps for the pairs of <syn-bbox, real-bbox>, <syn-RGB, real-RGB>, and <syn-seg/depth, real-RGB> increase gradually.

As motivated, we *advocate a differentiated knowledge transfer* for different information within syn-real pairs and formulate a gated syn-to-real knowledge transfer model for adaptive PCP (Gated-S2R-PCP). As shown in Fig. 1b, Gated-S2R-PCP absorbs three different information from the pedestrian bboxes, visual context, and semantic-depth clues in the synthetic data to the real data only with the pedestrian bboxes and raw RGB frames. Specifically, a knowledge distiller, a style shifter (StyS), and a distribution approximator (DistA) are incorporated for transferring the syn-to-real knowledge in pedestrian bboxes, RGB frames, semantic segmentation and depth maps, respectively. The **contributions** are threefold.

- We propose Gated-S2R-PCP, that incorporates a style shifter, a distribution approximator, and a knowledge distiller together for cross-domain learning of the visual frames, semantic-depth maps, and pedestrian bboxes.
- We propose a learnable gated unit (LGU) that adaptively learns the weights of different transfer methods to achieve suitable knowledge transfer learning for PCP, to obtain a general knowledge transfer model for different information with distinct synthetic-to-real domain gaps.
- We construct S2R-PCP-3181[1], which contains 3181 synthetic pedestrian crossing sequences with 498,740 frames. S2R-PCP-3181 contains pedestrian bboxes, RGB frames, depth maps, and semantic segmentation maps. Gated-S2R-PCP is verified on JAAD [7] and PIE [8] datasets and shows superiority to other SOTA methods. Additionally, we evaluate Gated-S2R-PCP on all pedestrian crossing sequences (50 ones) in the DADA-2000 dataset [19] to verify the PCP performance in near-collision scenes.

The remainder of the paper is organized as follows. Sec. II briefly reviews the related works. Gated-S2R-PCP is described in Sec. III, and evaluated by extensive experiments in Sec. IV. The conclusion is given in Sec. V.

II. Related Work

A. Pedestrian Crossing Prediction (PCP)

Pedestrian crossing prediction in driving scenes can be formulated by classifying the future crossing intention of pedestrians with a short time observation. Commonly, there is a time-to-crossing (TTC) interval (> 0 seconds) between the end of the observation and the future crossing state.

Probabilistic Model-driven Methods: Because of the high motility of pedestrians, traditional probabilistic model-driven methods model the PCP in the Markov process [20] or Bayesian models [21]. In 2013, the Daimler team of Germany and the University of Amsterdam designed a pedestrian path prediction method based on a recursive Bayesian filter [20]. Because of the common moving direction of the pedestrian group, the prediction task of the movement intention of the pedestrian crowd is described as a Markov Decision Process (MDP) problem [20]. Most traditional probabilistic models adopt the classical probability theory, which is highly interpretable but has limited generalization for diverse real scenes.

Sequential Encoding-Decoding Methods: With the promotion of deep learning models, most recent PCP works are formulated as sequential encoding-decoding frameworks [22]. Within this field, the pedestrian appearance, posture, trajectory, surrounding scenes, speed, *etc.*, are all investigated by appropriate temporal feature coding and fusion methods [23], [7], [8], [24], [25]. Among them, the most representative works are the joint attention in autonomous driving (JAAD) [7] and pedestrian intention estimation (PIE) [8]. These two datasets drive the PCP task significantly. Besides LSTM and GRU, 3D convolution is also popular in temporal encoding in the PCP task [26]. For example, Saleh *et al.* [27] adopt a series of 3D DenseNet modules to encode the spatiotemporal information of pedestrians, to judge the real-time crossing intention. However, due to the expensive encoding of 3D convolution, temporal 3D convNet (T3D) is introduced by Chaabane *et al.* [28] for PCP and improves the performance.

Interaction and Attention Driven Models: Some popular deep learning models, such as graph convolution network (GCN) [29] and Transformer [17], [30], are frequently used in PCP task, with the consideration of the pose, optical flow, ego-car velocity, pedestrian bboxes, and so on [31], [32]. For instance, Cadena *et al.* [33] extend their work, Pedestrian-Graph [34], and propose Pedestrian-Graph+ [33]. The GCN structure in Pedestrian-Graph+ explores multimodal data information on vehicle speed and image features. PedAST-GCN [32] achieves powerful PCP performance only with pedestrian pose, vehicle velocity encoded by GCN. Besides, specific occasions, *e.g.*, the non-signalized intersection [28] and red-light zone [35], are also focused. Since the intervention of the Transformer, it becomes tantalizing to attentively encode the information of pedestrians [36], where PCP with attention (PCPA) [25] is a milestone work. Pedformer [30] involves the pedestrian trajectory, ego-car velocity, RGB frames, and semantic segmentation maps together by a multi-head attention (MHA) and achieves the highest accuracy (0.93) on the PIE dataset.

The aforementioned PCP methods demonstrate significant progress while the variability insufficiency of real datasets makes the existing works involve much more information on pedestrians, ego-cars, and pedestrian-scene interactions to boost the PCP performance, which poses challenges for practical use when some information is not available, while domain knowledge transfer method is leveraged in this work.

B. Cross-Domain Adaptation in Traffic Scenes

Because of the limited variability of dynamic traffic scenes in one dataset, cross-domain adaptation becomes popular

[1]https://github.com/JWFanggit/Gated-S2R-PCP

between different datasets (with source datasets and target datasets) [37]. In the cross-domain adaption in the driving scene, the core problem is to approximate the distribution between the source domain and target domain. Therefore, distribution alignment [38], [39], cross-domain feature distance minimization [40], adversarial distribution approximation [41], [42], *etc.*, are commonly investigated in this field. For instance, Jiao *et al.* [38] propose a selective alignment network (SAN) for cross-domain pedestrian detection. Wang *et al.* [39] model an augmented feature alignment network for cross-domain object detection, where the domain image generation and domain-adversarial training are integrated into a unified framework. Recently, cross-domain object detection has been observed in the autonomous driving community. Li *et al.* [40] model a stepwise domain adaptation for the popular object detectors and excellent detection performance is obtained.

In addition to the visual domain adaptation, multi-spectral domain adaptation is another mainstream prototype in the traffic scene [43] to restrict the low visibility conditions issue in RGB frames. For example, Brehar *et al.* [22] explore the role of the far-infrared information of pedestrians in promoting the PCP in low-light conditions, and a CROSSIR dataset is constructed. In this approach, only the far-infrared images are used which does not explore the domain adaptation framework. Marnissi *et al.* [44] propose a thermal-to-visible domain adaptation for pedestrian detection, where the feature distribution alignments are adopted in the Faster R-CNN.

With the overview of cross-domain adaption works in traffic scenes, semantic segmentation, and detection are commonly concerned, while the PCP task needs further investigation.

C. Syn-to-Real Knowledge Transfer in Driving Scenes

Unlike traditional cross-domain adaptation in traffic scenes for real data, syn-to-real knowledge transfer approaches leverage the virtual simulator to generate vast samples based on the traffic layout structure and rules [45]. With the larger diverse weather and light conditions, occasion, and road types, syn-to-real knowledge transfer aims to cover the scene variability and fulfill parallel traffic system (PTS) [46], [12]. Based on this, researchers in this field begin to construct large-scale datasets with various realistic graph engines (*e.g.*, CARLA [47], UE4, *etc.*) for simulating the virtual traffic scene, and adopt the synthetic data in detection, semantic segmentation, and tracking [48], [49]. For instance, Wang *et al.* [49] contribute a parallel teacher for syn-to-real domain adaptation in traffic object detection, which consists of the data level and knowledge level domain distribution shift and takes the knowledge in the real world and virtual world together to achieve the knowledge distillation. As for the PCP task, the syn-to-real knowledge transfer emerged recently [10], [11], such as the syn-to-real fine-tuning for bounding box information learning [10] and syn-to-real knowledge distillation [11] involved by ourselves.

Building upon the state-of-the-art, similar to previous PCP works, we advocate involving multiple information of pedestrians and road scenes. Differently, we encode various information in the synthetic data and prefer a PCP in the real datasets only with pedestrian bboxes and RGB frames.

III. Methodology

Problem Formulation. Assume we have the short observation with T video frames, the PCP aims to predict the crossing or not-crossing label at time $T+\tau$, where $\tau > 0$ means the time-to-crossing (TTC) interval. We assume that there are four modalities (RGB frames, ped. boxes, depth maps, and semantic segmentation maps) and two modalities (only with RGB frames and ped. boxes) in synthetic $\mathcal{V}_{syn} = \{\boldsymbol{I}_{syn}, \boldsymbol{B}_{syn}, \boldsymbol{D}_{syn}, \boldsymbol{S}_{syn}\}$ and real $\mathcal{V}_{real} = \{\boldsymbol{I}_{real}, \boldsymbol{B}_{real}\}$ datasets, respectively. As defined by Eq. 1, PCP can be formulated as the following inference problem

$$\hat{\mathbf{p}}_{T+\tau} = G(\phi(\mathcal{V}_{real}|_{1:T}, \mathbf{W})), \tag{1}$$

where G is the pedestrian crossing determiner with the input of *ped. boxes* and *RGB frames* of the T frames in real datasets, and $\phi(.,.)$ is the feature encoder for $\mathcal{V}_{real}$. $\hat{\mathbf{p}}_{T+\tau} = \{c, nc\}$ denotes the predicted label of crossing or not-crossing. $\mathbf{W}$ is the model parameters in $\phi(.,.)$, which is obtained by training the synthetic-to-real knowledge transfer models.

The framework of the Gated-S2R-PCP model is presented in Fig. 2. For the "*ped. boxes*", we formulate a **knowledge distiller (KnowD)** with a teacher model and a student model for $\boldsymbol{B}_{syn}$ and $\boldsymbol{B}_{real}$, respectively. As for the RGB frames, because of the same modality, the main difference is the image styles, and we model a **style shifter (StyS)** to transfer the diverse weather and light conditions in $\boldsymbol{I}_{syn}$ to $\boldsymbol{I}_{real}$. For the depth and semantic segmentation maps, that have the largest distribution gap to RGB frames, we design a **distribution approximator (DistA)** to project the feature embedding of different modalities into a *shared* feature space for the PCP task. To achieve an adaptive domain knowledge selection, we fuse different knowledge transfer functions by LGU to fulfill an ensemble knowledge transfer learning, as defined by Eq. 2.

$$\begin{aligned} f_{gate} = \underset{\mathbf{W}_{LGU}}{LGU}\,(&\underset{\mathbf{W}_K}{KnowD}(\boldsymbol{B}_{syn} \rightarrow \boldsymbol{B}_{real}), \\ &\underset{\mathbf{W}_S}{StyS}(\boldsymbol{I}_{syn} \rightarrow \boldsymbol{I}_{real}), \\ &\underset{\mathbf{W}_D}{DistA}(\{\boldsymbol{D}_{syn}, \boldsymbol{S}_{syn}\} \longleftrightarrow \boldsymbol{I}_{real})), \end{aligned} \tag{2}$$

where $\mathbf{W}_K, \mathbf{W}_S$, $\mathbf{W}_D$, and $\mathbf{W}_{LGU}$ are the weights in $KnowD(\rightarrow)$, $StyS(\rightarrow)$, $DistA(\longleftrightarrow)$, and LGU, respectively. f_{gate} is the feature vector for PCP task.

A. Knowledge Distiller (KnowD)

KnowD contains a teacher model and a student model for pedestrian bbox learning of $\boldsymbol{B}_{syn}$ and $\boldsymbol{B}_{real}$, respectively. Suppose the pedestrian bboxes $\boldsymbol{B}$ be $\{\mathbf{b}_t\}_{t=1}^{T}$, where $\mathbf{b}_t = (x_t, y_t, w_t, h_t)$ represents the center point (x_t, y_t), width, and height of the bounding box, respectively.

In KnowD, the teacher model $\mathcal{T}$ uses synthetic bbox data to learn the pedestrian behavior pattern, and the pre-trained $\mathcal{T}$ is then adopted to guide the student model learning $\mathcal{S}$ in real bbox data. As formulated by Eq. 3, KnowD is defined as:

$$f_{\mathcal{S}}(\boldsymbol{B}_{real}) = KnowD(\mathcal{T}(\boldsymbol{B}_{syn}) \rightarrow \mathcal{S}(\boldsymbol{B}_{real}), \mathbf{W}_K), \tag{3}$$

where $f_{\mathcal{S}}(\boldsymbol{B}_{real})$ represents the feature representation after the distillation learning process, which is utilized for subsequent pedestrian crossing prediction task. Fig. 3 presents the

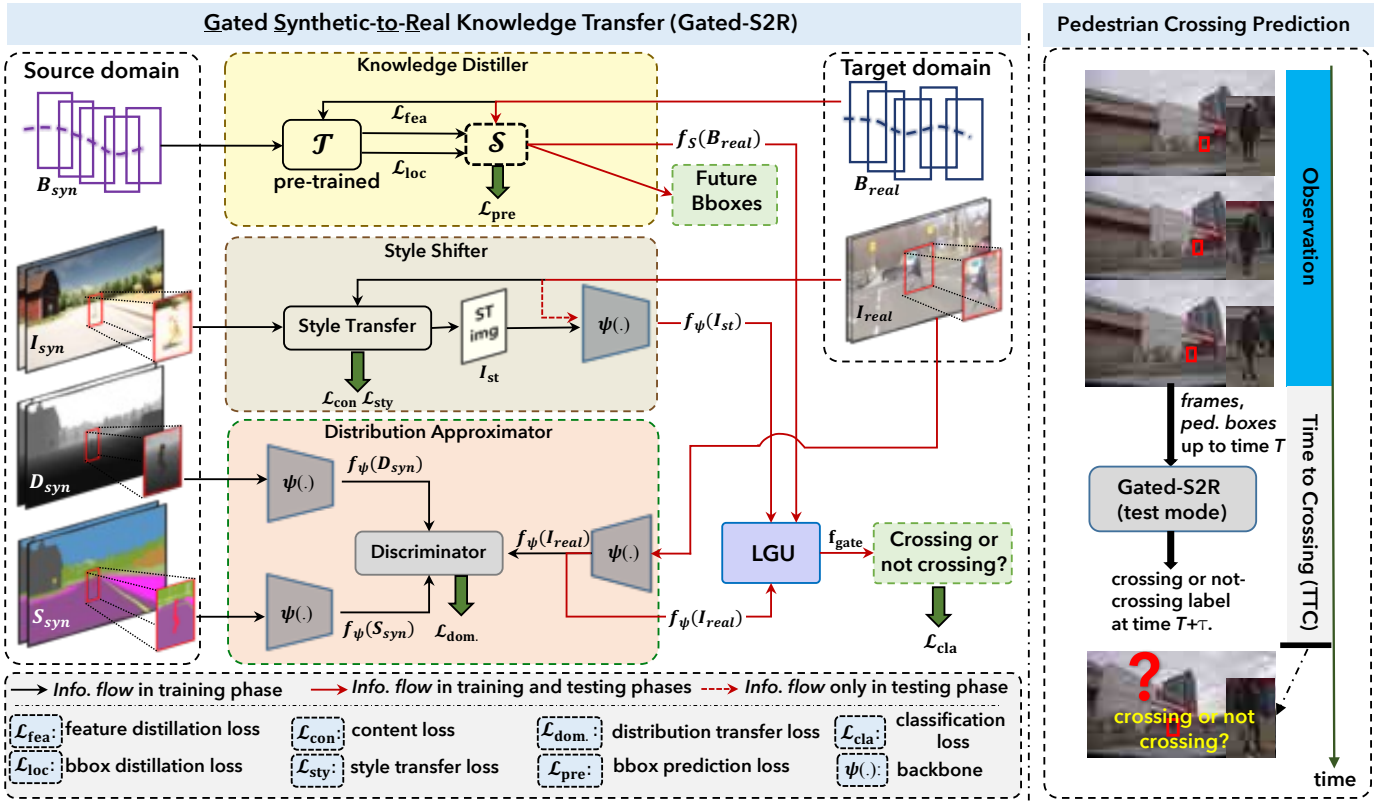

Fig. 2. The framework of **Gated-S2R-PCP**. **Synthetic domain:** pedestrian bounding boxes $\boldsymbol{B}_{syn}$, RGB frames $\boldsymbol{I}_{syn}$, depth maps $\boldsymbol{D}_{syn}$ and semantic segmentation maps $\boldsymbol{S}_{syn}$. **Real domain:** pedestrian bounding boxes $\boldsymbol{B}_{real}$ and RGB frames $\boldsymbol{I}_{real}$. During the training phase, the knowledge distiller, style shifer, and distribution approximator are trained to fulfill the knowledge transfer of $\boldsymbol{B}_{syn} \rightarrow B_{real}$ (Eq. 3), $\boldsymbol{I}_{syn} \rightarrow \boldsymbol{I}_{real}$ (Eq. 5), and $\{\boldsymbol{D}_{syn}, \boldsymbol{S}_{syn}\} \longleftrightarrow \boldsymbol{I}_{real}$ (Eq. 7), respectively. With these domain adaptation approaches, we obtain the feature embedding of $f_{\mathcal{S}}(\boldsymbol{B}_{real})$ of $\boldsymbol{B}_{real}$, $f_{\psi}(\boldsymbol{I}_{st})$ for style-transferred image set $\boldsymbol{I}_{st}$ and the $f_{\psi}(\boldsymbol{I}_{real})$ for $\boldsymbol{I}_{real}$ over T frames. Finally, LGU adaptively fuses $f_{\mathcal{S}}(\boldsymbol{B}_{real})$, $f_{\psi}(\boldsymbol{I}_{st})$ and $f_{\psi}(\boldsymbol{I}_{real})$ to form the multi-source feature f_{gate} for pedestrian crossing prediction. During the testing phase, only $\boldsymbol{I}_{real}$ and $\boldsymbol{B}_{real}$ are the input up to time T (1:T). The pedestrian crossing predictor determines the crossing or not crossing label of the pedestrians at time $T+\tau$, where τ means the time-to-crossing (TTC).

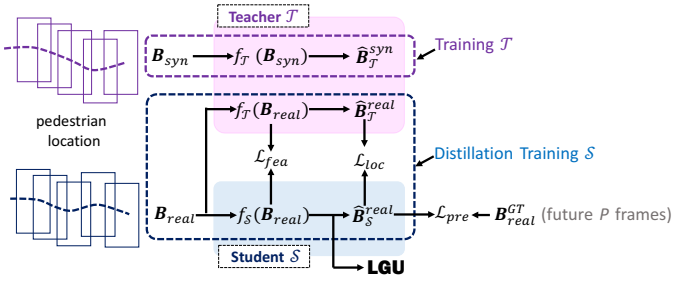

Fig. 3. The structure of Knowledge Distiller.

structure of KnowD. In addition to the feature representation learning, we incorporate a prediction path of the future object location (**FOL**) to improve the PCP performance.

Consequently, given T successive frames, the output is the feature representation of pedestrian bboxes of T successive frames for the PCP and pedestrian bboxes in future P frames of the FOL, *i.e.*, predicting the bboxes at the time window $[T+1:T+P]$, where P is evaluated in the experiments.

Training $\mathcal{T}$: We adopt the Transformer network [17] as the backbone model of $\mathcal{T}$. Firstly, the pedestrian bboxes $\boldsymbol{B}_{syn} = \{\mathbf{b}_t^{syn}\}_{t=1}^{T}$ in the synthetic dataset, is the input of the teacher model $\mathcal{T}$, and $\hat{\boldsymbol{B}}_{\mathcal{T}}^{syn} = \{\hat{\mathbf{b}}_t\}_{t=T+1}^{T+P}$ is the pedestrian bboxes to be predicted. $\boldsymbol{B}_{syn}$ is first encoded by $\mathcal{T}$ to a feature embedding by position embedding layer in Transformer. It is then encoded by Transformer (with seven interleaved multi-head attention (MHA) and linear normalization (LN)) as a 512-dimension feature vector $f_{\mathcal{T}}(\boldsymbol{B}_{syn})$. The feature vector can be used to classify pedestrian crossing labels at time $T+\tau$ and predict $\hat{\boldsymbol{B}}_{\mathcal{T}}^{syn}$ in the synthetic dataset. Once $\mathcal{T}$ is trained, the parameters in $\mathcal{T}$ are fixed in the following distillation process.

Distillation Training $\mathcal{S}$: For the student model $\mathcal{S}$, we prefer a lightweight but efficient architecture. In this work, we take the ResNet+LSTM to encode the pedestrian bboxes on real dataset, where $\boldsymbol{B}_{real} = \{\mathbf{b}_t^{real}\}_{t=1}^{T}$ and $\hat{\boldsymbol{B}}_{\mathcal{S}}^{real} = \{\hat{\mathbf{b}}_t^{real}\}_{t=T+1}^{T+P}$ are the input and output of $\mathcal{S}$ with the same format as $\mathcal{T}$. To be lightweight, ResNet18 [50] is chosen as the backbone model, and LSTM temporally associates the feature embedding of ResNet18 for the FOL of pedestrians. The feature vector $f_{\mathcal{S}}(\boldsymbol{B}_{real})$ with dimension 512 is obtained. To achieve effective distillation, we model three types of losses.

1) *Feature Distribution Consistency* ($\mathcal{L}_{fea}$) advocates the shared label for crossing or not-crossing determined by $\mathcal{T}$ and $\mathcal{S}$, and computed by the Kullback–Leibler Divergence (KLD) $\mathrm{KLD}(f_{\mathcal{T}}(\boldsymbol{B}_{real}), f_{\mathcal{S}}(\boldsymbol{B}_{real}))$.

2) *FOL Consistency* ($\mathcal{L}_{loc}$) expects the equivalent future

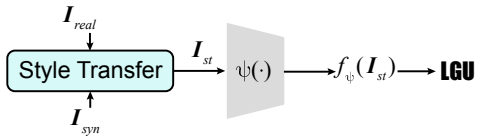

Fig. 4. The structure of Style Shifter. The real RGB frames $\boldsymbol{I}_{real}$ are the content images and the synthetic RGB frames $\boldsymbol{I}_{syn}$ are the style images.

pedestrian bboxes predicted by $\mathcal{T}$ and $\mathcal{S}$, specified by mean square error, *i.e.*, MSE($\hat{\boldsymbol{B}}_{\mathcal{S}}^{real}, \hat{\boldsymbol{B}}_{\mathcal{T}}^{real}$).

3) *FOL Accuracy* ($\mathcal{L}_{pre}$) prefers a minimum location distance between $\hat{\boldsymbol{B}}_{\mathcal{S}}^{real}$ and the ground truth $\boldsymbol{B}_{real}^{GT}$, and calculated by MSE($\hat{\boldsymbol{B}}_{\mathcal{S}}^{real}, \boldsymbol{B}_{real}^{GT}$), where $\boldsymbol{B}_{real}^{GT}$ represents the ground truth for the pedestrian bboxes in the time window $[T+1, T+P]$.

In total, the loss functions in KnowD are summarized as

$$\mathcal{L}_{kd} = \mathcal{L}_{fea} + \mathcal{L}_{loc} + \mathcal{L}_{pre}. \tag{4}$$

B. Style Shifter (StyS)

StyS is used to transfer the frame appearance of synthetic RGB frames $\boldsymbol{I}_{syn}$ under different conditions to the real RGB frames $\boldsymbol{I}_{real}$. This module can be viewed as a scene appearance augmentation for the real dataset. We use the synthetic RGB frames as the style image set and the real RGB frame as the content image set. Then, we formulate StyS using Eq. 5, *i.e.*,

$$f_\psi(\boldsymbol{I}_{st}) = StyS(\boldsymbol{I}_{syn} \rightarrow \boldsymbol{I}_{real}, \mathbf{W}_S), \tag{5}$$

where $f_\psi(\boldsymbol{I}_{st})$ is the feature embedding of the style-transferred image $\boldsymbol{I}_{st}$. The structure of StyS is shown in Fig. 4.

In this module, both synthetic and real RGB frames stand at the local regions of pedestrians. The global image contains much irrelevant background clutter, which may degrade the following PCP task and cause a slow convergence. We crop the rectangle region with β (1.5 in this work) times the size around the pedestrian bounding box, and the cropped regions are scaled with the same size. Consequently, the input region size for StyS is $(T, 3, w, h)$, where $[w, h]$ is the region size.

Feature-Level StyS: We advocate a feature-level style transfer to boost the style diversity in StyS. Here, we introduce the efficient AdaIN method [51] to generate the style-transferred image set $\boldsymbol{I}_{st}$ of T frames. Then, $\boldsymbol{I}_{st}$ is encoded by a spatial-temporal backbone model $\psi(.)$, and the encoded feature is then fed into the PCP classifier. In this work, different backbones of $\psi(.)$ are experimentally evaluated.

To measure the quality of generated $\boldsymbol{I}_{st}$, it is necessary to calculate the content loss $\mathcal{L}_{con}$ and style loss $\mathcal{L}_{sty}$. Similarly, the MSE loss between the feature embedding $f_\psi(\boldsymbol{I}_{real})$ and $f_\psi(\boldsymbol{I}_{st})$ is adopted to compute the $\mathcal{L}_{con}$. $\mathcal{L}_{sty}$ calculates the MSE of the vector mean and variance of $f_\psi(\boldsymbol{I}_{syn})$ and $f_\psi(\boldsymbol{I}_{st})$. Therefore, the total loss in StyS is

$$\mathcal{L}_{st} = \mathcal{L}_{con} + \alpha\mathcal{L}_{sty}, \tag{6}$$

where α is a hyper-parameter and set as 10 in this work.

C. Distribution Approximator (DistA)

As the largest domain gap between the synthetic depth maps ($\boldsymbol{D}_{syn}$), synthetic semantic segmentation maps ($\boldsymbol{S}_{syn}$), and real RGB frames ($\boldsymbol{I}_{real}$), we model the DistA by Eq. 7 to find the shared feature distribution, *i.e.,*

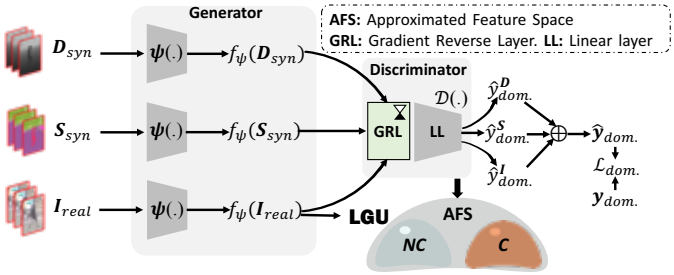

Fig. 5. The structure of Distribution Approximator. The generator encodes the input to the high-dimensional feature embedding, and the discriminator acts as the domain classifier, which aims to confuse the domain label and approximate the feature embedding of different domains to a shared space.

$$f_\psi(\boldsymbol{I}_{real}) = DisA(\{\boldsymbol{D}_{syn}, \boldsymbol{S}_{syn}\} \longleftrightarrow \boldsymbol{I}_{real}, \mathbf{W}_D). \tag{7}$$

Different from KnowD and StyS, DisA enforces a bi-directional approximation.

DisA is inspired by the adversarial insights of generative adversarial networks (GAN) [52] with the generator-discriminator structure. Differently, DisA focuses on feature-level adversarial training. The generator encodes the $\boldsymbol{D}_{syn}$, $\boldsymbol{S}_{syn}$, and $\boldsymbol{I}_{real}$ by the same backbone model $\psi(.)$ of Style Shifter, which results in the feature embedding of $f_\psi(\boldsymbol{D}_{syn})$, $f_\psi(\boldsymbol{S}_{syn})$, and $f_\psi(\boldsymbol{I}_{real})$, respectively. The discriminator determines whether the feature embedding comes from the synthetic or real domain. The discriminator expects to be confused by the generator for different domains as much as possible while avoiding the GAN-type training. A gradient reverse layer (GRLayer) in [53] is used to reverse the feature space gradient and promote the optimization of the discriminator during training, as shown in Fig. 5.

In DisA, the data form is the same as the input of StyS, *i.e.*, rectangle regions around the pedestrian boxes. All input data is processed with the same dimensions (T, c, w, h), where channel c equals 1 for $\boldsymbol{D}_{syn}$ and $\boldsymbol{S}_{syn}$, while c is 3 for $\boldsymbol{I}_{real}$. The spatial regions over T frames are fed into the generator $\psi(.)$ to obtain the feature embedding $f_\psi(\mathcal{Z})$, where $\mathcal{Z}$ denotes the $\boldsymbol{D}_{syn}$, $\boldsymbol{S}_{syn}$, or $\boldsymbol{I}_{real}$.

The discriminator $\mathcal{D}(.)$ consists of the GRLayer and a LN layer. The feature embedding is classified after gradient inversion operation in synthetic and real data domain determination. Therefore, the discriminator can be treated as a domain classifier to determine $\hat{y}^{\boldsymbol{D}}_{dom.}$, $\hat{y}^{\boldsymbol{S}}_{dom.}$, and $\hat{y}^{\boldsymbol{I}}_{dom.}$, and the loss is defined by following binary cross-entropy (BCE)

$$\mathcal{L}_{dom.} = \sum_{\mathcal{Z} \in \{\boldsymbol{D}_{syn}, \boldsymbol{S}_{syn}, \boldsymbol{I}_{real}\}} \text{BCE}(\mathcal{D}(f_\psi(\mathcal{Z})), \mathbf{y}_{dom.}), \tag{8}$$

where $\mathbf{y}_{dom.} = \{0, 1\}$ is the synthetic and real domain labels.

D. Learnable Gated Unit

Since the knowledge importance may be different for different scenarios, it is desired to properly set weights for different

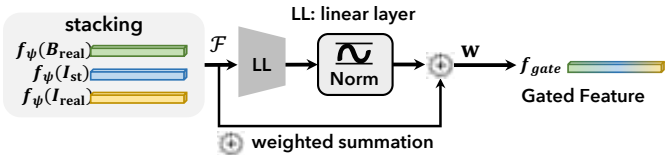

Fig. 6. The structure of LGU.

knowledge transfer functions accordingly. To fulfill an end-to-end ensemble learning, LGU adaptively fuses $f_{\mathcal{S}}(\boldsymbol{B}_{real})$ (Eq. 3), $f_\psi(\boldsymbol{I}_{st})$ (Eq. 5), and $f_\psi(\boldsymbol{I}_{real})$ (Eq. 7) over T frames. Hence, Eq. 2 is calculated by Eq. 9.

$$\begin{aligned} \mathcal{F} &= [f_{\mathcal{S}}(\boldsymbol{B}_{real}); f_\psi(\boldsymbol{I}_{st}); f_\psi(\boldsymbol{I}_{real})], \\ \mathbf{w} &= \text{Norm}(\text{LL}(\mathcal{F}, \mathbf{W}_{LGU})), \\ f_{gate} &= \mathbf{w} \oplus \mathcal{F}, \end{aligned} \tag{9}$$

where [;] denotes a stacking operation for three kinds of features, $\mathbf{W}_{LGU}$ denotes the parameters in LGU, and $\mathbf{w} \in R^{1\times 3}$ is the gated weight for feature fusion. LL(.,.) consists of 3 linear layers and a normalization operation. The gate operation is fulfilled by a weighted summation $\oplus$ of $\mathcal{F}$ by $\mathbf{w}$.

E. Pedestrian Crossing Prediction

Thus, the PCP problem in Eq. 1 is re-formulated as

$$\hat{\mathbf{p}}_{T+\tau} = G(f_{gate}), \tag{10}$$

where $G(.)$ contains 3 linear layers $l(.)$ and a Gumbel-softmax function [54] which is defined as

$$\sigma_t = \frac{e^{\frac{1}{\eta}(l(f^i_{gate}) - \log(-\log(\epsilon(i))))}}{\sum_j e^{\frac{1}{\eta}(l(f^j_{gate}) - \log(-\log(\epsilon(j))))}}, \tag{11}$$

where σ_t is the output of Gumbel-softmax operation on $l(f_{gate})$, and the logit score of the i^{th} feature item of $l(f_{gate})$ is denoted as $l(f^i_{gate}) - \log(-\log(\epsilon(i)))$, and $\epsilon(i)$ is a random variable that obeys the uniform distribution $U(0,1)$. $l(f_{gate})$ is the output of linear layers $l(.)$. η is the temperature parameter and is set as 1. $G(.)$ is optimized with the BCE loss as

$$\begin{aligned} \mathcal{L}_{cla} = \textstyle\sum_k (&-\hat{\mathbf{p}}_{T+\tau}(k) \log(\mathbf{p}_{T+\tau}(k)) + \\ &(1 - \hat{\mathbf{p}}_{T+\tau}(k)) \log(1 - \mathbf{p}_{T+\tau}(k))), \end{aligned} \tag{12}$$

where $\hat{\mathbf{p}}_{T+\tau}(k)$ and $\mathbf{p}_{T+\tau}(k)$ are the predicted and ground-truth intention label k of crossing or not-crossing.

The overall loss function of the Gated-S2R-PCP is

$$\min_{\{\mathbf{W}_K, \mathbf{W}_S, \mathbf{W}_D, \mathbf{W}_{LGU}\}} : \mathcal{L} = \mathcal{L}_{st} + \mathcal{L}_{dom.} + \mathcal{L}_{kd} + \mathcal{L}_{cla}. \tag{13}$$

The trained model weights are then used to generate the gated feature f_{gate} for PCP in the testing phase.

IV. Experiments and Discussions

A. Datasets

Synthetic PCP dataset: In this work, we construct a new synthetic PCP dataset, named Syn-PCP-3181. Syn-PCP-3181 contains 3,181 video sequences with more modalities including RGB frames, depth maps, semantic segmentation maps, and pedestrian bboxes, where RGB frames, depth maps, and semantic segmentation maps are with pixel-to-pixel registration automatically in CARLA [55] simulator. Fig. 7 displays several examples in Syn-PCP-3181. Syn-PCP-3181 is generated by CARLA under various weather and light conditions, where 1,226 sequences contain crossing behavior and 1955 non-crossing ones. When the dataset is generated, the camera selects the vehicle camera to record the behavior of the pedestrian in front of the vehicle, and the resolution is set to 1280×720 (pixels). To obtain the pedestrian crossing behavior, instead of random setting, we manually set different initial positions and various destination positions in varying urban scenes. To ensure the diversity of the dataset, we also assign different ages and genders to the pedestrians, such as youngest, old age, and middle age. In addition, Syn-PCP-3181 includes a variety of weather (sunny, cloudy, and rainy), light (sunset, daytime, and nighttime), and road occasions (urban, rural). Each kind of weather and light condition takes 1/3 in balance. The occasions, gender, and age statistics are shown in Fig. 8.

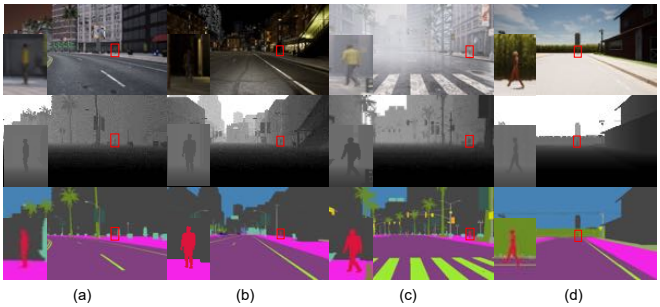

Fig. 7. Several scenarios in Syn-PCP-3181 dataset. (a), (b), (c), and (d) contain the RGB frames, depth maps, and semantic segmentation maps collected in the urban (daytime), urban (nighttime), rainy, and rural scenes, respectively.

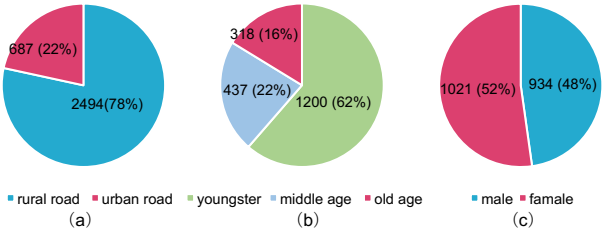

Fig. 8. The statistics of Syn-PCP-3181 for occasions, gender, and ages.

To avoid behavior confusion, each video sequence only contains one pedestrian. Therefore, the number of pedestrians is 3,181. Each crossing sequence contains 240 frames, while one not-crossing sequence contains 100 frames, so the total number of frames in the data set can reach 489,740. The most similar dataset to Syn-PCP-3181 is the Virtual-PedCross-4667 [11] collected by ourselves which only contains the RGB frames and pedestrian bboxes.

Real PCP datasets: JAAD [7] captures in different countries, which consists of a total of 346 video sequences with a total of 75K frames and 2,786 pedestrians. Among them, 495+191 samples are crossing. The remaining 2,100 pedestrians are not-crossing samples. JAAD has a subset, JAAD_beh,

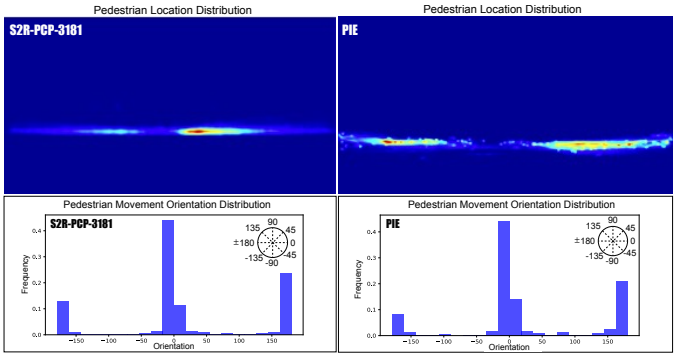

Fig. 9. The pedestrian location (center point here) and movement orientation distribution in Syn-PCP-3181 and PIE datasets, where the upper part (hotspot maps) represents the location distribution in which each point means a pedestrian passes through it. To be clear, we reshape the frame resolution of PIE as the same as Syn-PCP-3181 to 1280×720 (pixels) for comparable review. The bottom part denotes the pedestrian orientation histogram, which is computed by accumulating the moving direction ($\in [-180, 180]$) of pedestrians, *i.e.*, pedestrian location vector at two adjacent frames.

TABLE I
COMPARISON OF S2R-PCP-3181 WITH OTHER EXISTING PEDESTRIAN CROSSING DATASETS. INFO.: INFORMATION TYPE OF DATASET (I: RGB FRAMES; D: DEPTH MAPS; S: SEMANTIC SEGMENTATION MAPS)

dataset	#seqs.	#frames	#peds	Syn./Real.	Info.
JAAD [7]	346	75K	686	Real	I
PIE [8]	55	293K	1800	Real	I
Virtual-PedCross-4667 [11]	4667	745K	5835	Syn	I
S2R-PCP-3181	3181	490K	3181	Syn	I, D, S

that includes only 686 behavioral pedestrians. **PIE** [8] collects six consecutive hours of daytime data in the city of Toronto, which consists of 55 sequences with 293K frames and 1834 pedestrians. There are 512 crossing and 1322 not-crossing samples. We also evaluate the performance on the **PSI** [56] dataset, concentrating on the socialized interaction of ego-car and pedestrians, consisting of 110 representative vehicle-pedestrian encountering clips (15 seconds of each) with 25,881 frames collected by 110 drivers. Similar to [56], we adopt the first 75 sequences in training and the 76^{th}-110^{th} sequences for the testing.

It is known that there is no pedestrian behavior model in CARLA. However, the crossing behavior captured by the ego-view camera often shows relatively horizontal movement. As shown in Fig. 9, the pedestrian location (center point here) and the movement orientation in Syn-PCP-3181 show similar distributions to PIE. In addition, from the location distribution, the pedestrians in Syn-PCP-3181 own larger distances to the ego-vehicle than the PIE dataset, which indicates that the hidden pedestrian behavior model in Syn-PCP-3181 is complementary and positive for real dataset testing. Table. I shows the attribute comparison between different PCP datasets.

B. Implementation Details

For JAAD and PIE datasets, following the setting of other PCP works, we adopt 16 consecutive video frames (0.5 seconds) as the observation to predict whether pedestrians would cross with the τ of 30 or 60 frames (1 to 2 seconds).

In the Knowledge Distiller (Sec. III-A), the teacher model $\mathcal{T}$ uses 3 default Transformer layers [17] with an 8-head self-attention module for each Transformer layer, and generates a 64-dimensional feature vector. The student model $\mathcal{S}$ only has a ResNet18 module and two LSTM layers [57], and each LSTM layer has 100-dimensional hidden states. For pedestrian bbox information, the input dimension is ($bsize$, 32, 4) of the Knowledge Distiller. For the FOL path in the Knowledge Distiller, the future $P = 16$ frames (0.5 seconds) are used.

As stated in the Style Shifter (Sec. III-B), we crop the local image regions around the pedestrian boxes for the input of the Style Shifter. The local regions are scaled to 112×112. Therefore, the shape of all RGB data is ($bsize$, 16, 3, 112, 112), where $bsize = 2$ denotes the batch size in all experiments. In the Distribution Approximator (Sec. III-C), the input local image region is divided into patches with the size of 16×16, which forms the visual tokens in the backbone model $\psi(.)$. The backbone model $\psi(.)$ outputs the feature vector with a dimension of 64. Consequently, 16 frames of observation consist of 784 ($112/16 \times 112/16$) tokens.

Training Details: The training of teacher model $\mathcal{T}$ and the student model $\mathcal{S}$ takes the Adam optimizer with the learning rate is 10^{-5} and a decay of 0.8. The decay step is set to 10, which means that the learning rate is reduced to 80% in every 10 training epochs. $\mathcal{T}$ is trained with 10 epochs. For the JAAD and PSI datasets, the training epoch is set to 40 and the epoch is set to 20 for the PIE dataset because the sample size of PIE is larger than JAAD and converged at 20 training epochs. The training epochs for JAAD are the same as the ones in PCPA [58] for a fair comparison. To prevent overfitting, a dropout ratio of 0.5 is introduced. This work is implemented with Python 3.7 and PyTorch 1.7 on a platform with one NVIDIA GeForce RTX 2080Ti GPU and 64 RAM.

C. Metrics

Following other PCP works [17], [35], [36], the accuracy (**Acc**), the area under ROC curve (**Auc**), precision (**Pre**), **Recall** and **F1**-measure are used to evaluate the models. These metrics all pursue higher values for better performance. These metrics evaluate the classification performance for the pedestrian crossing or not crossing in the future $T + \tau$, which determines an early prediction of certain behavior and saves time for safe decisions. Specifically, Acc, Pre, and Rec are computed by

$$\begin{aligned} \text{Acc} &= \frac{TP + TN}{TP + FP + FN + TN}, \\ \text{Pre} = \frac{TP}{FP + FP}&, \text{Rec} = \frac{TP}{TP + FN} \end{aligned} \tag{14}$$

where TP represents True Positive, TN denotes True Negative, FP and FN are the False Positive and False Negative, respectively. The **Auc** value is computed by the area under the receiver operating characteristic (ROC) curve represented by the curve between true positive rate (TPR) and false positive rate (FPR). F1$= 2 \times \frac{Pre \times Rec}{Pre + Rec}$ provides a comprehensive and balanced evaluation for the precision and recall rates.

TABLE II
PERFORMANCE COMPARISON ON THE PIE DATASET FOR DIFFERENT KNOWLEDGE TRANSFER WAYS, *i.e.*, **STYS**, **DISTA**, AND **KNOWD**.

StyS	DistA	KnowD	Acc	Auc	F1	Pre	Rec
			0.81	0.78	0.74	0.70	0.78
✓			0.82	0.74	0.62	0.67	0.58
	✓		0.79	0.71	0.57	0.63	0.53
		✓	0.88	0.86	0.80	0.76	**0.84**
✓	✓		0.82	0.75	0.63	0.63	0.63
✓		✓	0.87	0.87	0.81	0.79	0.82
	✓	✓	0.87	0.86	0.79	0.78	0.80
✓	✓	✓	**0.92**	**0.90**	**0.82**	**0.83**	0.80

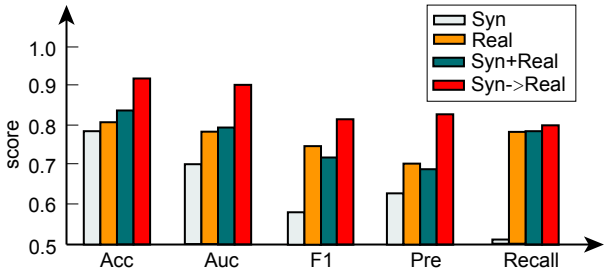

Fig. 10. Performance comparison on the PIE dataset for "Syn" training, "Real" training, "Syn+Real" training, and our "Syn→ Real" training modes.

D. Ablation Studies

In the ablation studies, we exhaustively evaluate the proposed method's components.

1) Performance check for different knowledge transfer ways: To check the role of **KnowD**, **StyS**, and **DistA**, we extensively evaluate different combinations of KnowD, StyS, and DistA. If we do not contain one knowledge transfer module, we can remove the synthetic data path to KnowD, StyS, or DistA, and maintain the real data path, as shown by the blue lines in Fig. 2. If we want to evaluate the performance without all the knowledge transfer modules, we can remove all the synthetic data paths, which means that we only have real data for training.

Table. II presents the performance on different combinations of KnowD, StyS, and DistA on the PIE dataset [8]. We can see that KnowD plays an important role in PCP, and each combination with KnowD shows better results than the ones without KnowD. Especially, only KnowD generates the best **Recall** value in this comparison. The main reason is that the pedestrian bbox information in synthetic and real data has the smallest domain gap. Therefore, the promotion role of bbox clues in synthetic can be directly borrowed. Contrarily, DisA exhibits the worst performance compared with StyS and KnowD, mainly due to the largest distribution gap between semantic segmentation or depth maps with RGB frames. When fusing all the knowledge transfer ways, the best performance (w/o Recall) is obtained, which means that the LGU can adaptively select the knowledge transfer ways.

2) Performance ablations by using different domain data in training: We evaluate the "**Syn**" training, "**Real**" training, "**Syn+Real**" training, and our "**Syn→ Real**" training modes here. As for the "**Syn**" training, "**Real**" training, and "**Syn+Real**" training modes, we only maintain the flowchart marked by red lines in Fig. 2, *i.e.*, the same flowchart setting as the testing phase without knowledge transfer modules. The testing set of these settings is the same, while the training set is different. "**Syn**" training mode inputs the RGB frames and pedestrian bboxes in 3181 synthetic sequences. "**Real**" is trained on the training data in the PIE dataset, and "**Syn+Real**" directly is trained by the combined set of 3181 synthetic sequences and the training set of PIE.

Fig. 10 displays the performance difference for these modes. We can see "**Syn**" training shows the worst performance because of the domain gap even though it owns the largest training set. "**Real**" training provides the comparative Recall value. If we directly combine the synthetic data and real data ("**Syn+Real**") in the same training way as "**Real**", The F1 and Pre values degrade mainly because of the domain confusion. On the contrary, our formulation, *i.e.*, "**Syn→ Real**", obtains the best performance through the suitable knowledge transfer ways for different information.

TABLE III
PERFORMANCE INFLUENCE OF THE *Backbone* FEATURE MODEL IN STYS OR DISA, AND THE *future object location (FOL)* PREDICTION IN KNOWD ON THE PIE DATASET, WHERE P. (MB) MEANS THE MODEL SIZE AND FPS DENOTES THE INFERENCE FRAMES PER SECOND.

Backbone/FOL	Acc	Auc	F1	Pre	Rec	P. (MB)	FPS
StyS/DistA+**C3D** [59]	0.87	0.87	0.81	0.80	**0.81**	186	52
+**TSM** [60]	0.88	0.86	0.79	0.80	0.79	159	57
+**Timesformer** [18]	**0.92**	**0.90**	**0.82**	**0.83**	0.80	307	30
KnowD	0.87	0.79	0.79	0.71	**0.88**	64	127
+**FOL**	0.88	0.86	0.80	0.76	0.84	65	126
Gated-S2R-PCP	0.90	0.87	0.80	0.76	0.84	306	30
+**FOL**	**0.92**	**0.90**	**0.82**	**0.83**	0.80	307	30

3) Role of various $\psi(.)$ for StyS or DistA: In the StyS and DistA, we introduce a backbone model for feature extraction, which may influent the final performance. We evaluate the role of different backbones here. We replace the Timesformer [18] with 3D Convolution (C3D) [59] and Temporal Shift Model (TSM) [60], respectively, both of which are commonly used video coding networks. C3D is pre-trained on the UCF101 dataset [70]. TSM is extended from the pre-trained ResNet 50 [50]. Timesformer [18] here is not pre-trained and directly used for training. The top half of Table. III displays the results with different backbones in StyS and DistA. From the results, we can see that Timesformer shows the best performance. However, C3D and TSM are also comparative. Therefore, the self-attention mechanism in Timesformer plays a positive role. Additionally, we also list the parameter size and inference efficiency here. It shows that the backbone model Timesformer has a more expensive computation cost than TSM and C3D. However, Timesformer boosts the PCP performance and meets the real-time demand.

4) Influence of FOL prediction: We also evaluate FOL influence in KnowD which forms multi-task learning for pedestrian bbox and crossing intention prediction. Here, we provide two evaluation groups: KnowD with or without FOL and Gated-S2R-PCP with or without FOL. The results are listed in the bottom half of the Table. III, and manifestly FOL improves the performance except for the Recall value, and

TABLE IV
PERFORMANCE COMPARISON OF OUR METHOD AND THE STATE-OF-THE-ART ON $JAAD_{all}$, $JAAD_{beh}$, AND PIE DATASETS. THE BEST AND SECOND VALUES FOR EACH METRIC ARE MARKED BY RED AND BLUE COLOR, RESPECTIVELY.

Model	Input Clues	$JAAD_{all}$					$JAAD_{beh}$					PIE				
		Acc	Auc	F1	Pre	Rec	Acc	Auc	F1	Pre	Rec	Acc	Auc	F1	Pre	Rec
ATGC [7] ICCVW2017	I, 2DB	0.67	0.62	0.76	0.72	0.80	0.48	0.41	0.62	0.58	0.66	0.59	0.55	0.39	0.33	0.47
MultiRNN [61]CVPR2018	2DB	0.79	0.79	0.58	0.45	0.79	0.61	0.50	0.74	0.64	0.86	0.83	0.80	0.71	0.69	0.73
SFGRU [23]BMVC2019	I, 2DB, PO, ECV	0.83	0.77	0.58	0.51	0.67	0.58	0.56	0.65	0.68	0.62	0.86	0.83	0.75	0.73	0.77
SingleRNN-LSTM [24]IV2020	I, 2DB, ECV	0.78	0.75	0.54	0.44	0.70	0.51	0.48	0.61	0.63	0.59	0.81	0.75	0.64	0.67	0.61
PCPA [58]WACV2021	I, PO, ECV, 2DB	0.83	0.83	0.64	-	-	0.56	0.56	0.68	-	-	0.87	0.86	0.77	-	-
BiPed [62]ICCV2021	T, I, 2DB, ECV, SS	0.83	0.79	0.60	0.52	-	-	-	-	-	-	0.91	0.90	0.85	0.82	-
IntFormer [63]2021	I, 2DB, PO, ECV	0.86	0.78	0.62	-	-	0.59	0.54	0.69	-	-	0.89	0.92	0.81	-	-
ST-CrossingPose [64]TITS2022	PO	-	-	-	-	-	0.63	0.56	0.74	0.66	0.83	-	-	-	-	-
Yang *et al.* [65]TIV2022	I, 2DB, PO, ECV	0.83	0.82	0.63	0.51	0.81	0.62	0.54	0.74	0.65	0.85	0.89	0.86	0.80	0.79	0.80
Pedestrian Graph +(32) [66]TITS2022	GCN, CNN	0.86	0.88	0.65	0.58	0.75	0.70	0.70	0.76	0.77	0.75	0.89	0.90	0.81	0.83	0.79
VR-PCP [11]ITSC2022	I, 2DB	0.86	0.81	0.77	0.74	0.81	0.64	0.66	0.76	0.70	0.89	0.89	0.88	0.67	0.74	0.84
Dong [67]ICLRW2023	I, 2DB, SS	0.88	0.81	0.69	-	-	0.64	0.56	0.75	-	-	0.89	0.88	0.79	-	-
PIT* [68]IEEE-TITS2023	I, 2DB, PO, ECV	0.86	0.87	0.66	0.55	0.82	0.69	0.67	0.77	0.74	0.83	0.90	0.90	0.80	0.81	0.80
Pedformer [30]ICRA2023	I, 2DB, T, SS, ECV	0.93	0.76	0.54	0.65	-	-	-	-	-	-	-	-	-	-	-
PIP-Net [26]Arxiv2024	I, 2DB, PO, ECV, SS, D	-	-	-	-	-	-	-	-	-	-	0.91	0.90	0.84	0.85	0.84
DPCIAN [69]IEEE-TITS2024	I, 2DB, PO	0.88	0.77	0.59	0.61	0.58	0.71	0.66	0.77	0.73	0.85	0.91	0.88	0.83	0.83	0.84
PedAST-GCN [32]IEEE-TITS2024	2DB, PO, ECV	0.88	0.81	0.68	0.66	0.70	0.69	0.59	0.80	0.68	0.96	0.90	0.86	0.81	0.83	0.79
PedCMT [36]IEEE-TITS2024	2DB, ECV	0.88	0.77	0.65	-	-	0.71	0.64	0.80	-	-	0.92	0.92	0.87	-	-
Gated-S2R-PCP (Ours)	I, 2DB	0.88	0.84	0.79	0.77	0.82	0.66	0.68	0.77	0.71	0.85	0.92	0.90	0.82	0.83	0.80

*PIT has four method versions and the best one for certain metrics in JAAD and PIE are different. Therefore, we take the average metric value of method versions here.
Clues: image (I); 2D bboxes (2DB); ego-car velocity (ECV); trajectory (T); pose (PO); depth map (D); body orientation (BO); semantic segmentation maps (SS).

the pure **KnowD-w/o-FOL** obtains 0.88 of Recall value. The reason is that pedestrian bbox has the closest relation to the crossing behavior with large displacement. Because FOL may involve small displacement, PCP+FOL may enforce a motion-consistent constraint on the PCP and show weak adaptation to the large motion in crossing. In Sec. IV-H2, we will evaluate the FOL influence with different prediction windows.

E. Comparison with SOTA Methods

To verify the proposed method, we take $JAAD_{all}$, $JAAD_{beh}$, and PIE in the evaluation, and compare it with eighteen state-of-the-art methods. They are ATGC [7] , MultiRNN [61], SFGRU [23], SingleRNN-LSTM [24], PCPA [58], BiPed [62], IntFormer [63], ST-CrossingPose [64], Pedestrian Graph+(32)[66], VR-PCP [11], PIT [68], Pedformer [30], PIP-Net [26], DPCIAN [69], PedAST-GCN [32], PedCMT [36], the methods proposed by Yang *et al.* [65] and Dong [67], respectively. The results are listed in Table. IV.

From the results, we can observe that our Gated-S2R-PCP obtains the comparative performance for all the metrics in three datasets. It seems that Pedestrian Graph+(32) [66], our Gated-S2R-PCP, PIT [68], VR-PCP [11], and PedAST-GCN [32] are the top five methods for JAAD dataset with a consideration of all metrics. Pedestrian Graph+(32) adopts 32 frames as the observation and the graph relation between pedestrians is considered. VR-PCP also leverages the knowledge of synthetic data for the PCP task, where only the RGB frames and pedestrian boxes are used. PIT fulfills the Progressive Transformer for PCP, where it designs 15 transformer layers with the inputs of RGB frames, pedestrian poses, and ego-vehicle motion state. PedAST-GCN only uses the pedestrian pose, bounding boxes, and vehicle velocity to avoid illumination and occlusion issues in video observation in the JAAD dataset, and leverages the spatial-temporal GCN to fulfill a fast PCP. Notably, Pedformer generates 0.93 accuracy in the $JAAD_{all}$ dataset, while the other metrics are not promising because of the promotion gap of different feature modalities for PCP performance.

Fig. 11. Visualization comparison of some examples between PCPA [58], Pedformer [30], and our method, where GT is ground truth, C is Crossing, and NC is not crossing. TTC denotes the video frames of time-to-crossing.

For the PIE dataset, the top five ones are Gated-S2R-PCP, BiPed [62], Pedestrian Graph+(32) [66], DPCIAN [69], and PedCMT [36]. Compared with JAAD, the video quality of PIE is better. Therefore, visual information can provide a promotion role for PCP. PIP-Net [26] and BiPed involving more information perform well for the PIE dataset. PedCMT formulates an attentive fusion for pedestrian pose and bounding boxes, and the importance of the time-varying interaction can be extracted. Like our Gated-S2R-PCP, BiPed takes the

future trajectory prediction to guide PCP, and promising performance is obtained. These observations indicate that more input frames and attributes will boost the performance, and future trajectory guidance can reduce the determination uncertainty of pedestrian crossing.

We also compare Gated-S2R-PCP, Pedformer [30] (with the best Acc value in the JAAD$_{all}$ test set) and PCPA [58] with several snapshots for the frames at time $T+\tau$ (*i.e.*, crossing or not crossing intention frames) in Fig. 11. Notably, we adopt the pre-trained weights of Pedformer [30] (trained by the training set of JAAD$_{all}$) to infer the same test set as our model. We can see that the small-scale issue of pedestrians has a more negative impact on PCPA than our method. Pedformer can almost generate consistent and accurate results with our Gated-S2R-PCP, while showing a failure in the frame #05118.

F. Further Comparison on the PSI Dataset

In this comparison, we take VR-GCN [71], the methods in PIE [8] (*abbrev.*, PIE-PCP) and PSI [56] (*abbrev.*, PSI-PCP), and our Gated-S2R-PCP for evaluation, and the results are listed in Table V. From the results, we can see that the performance of each method shows a decrease compared with the values in the PIE dataset [8]. Because there are only 25 sequences in the test set. The F1 values are persuasive for the not crossing and crossing intention prediction, and our Gated-S2R-PCP performs the best.

TABLE V
PCP PERFORMANCE COMPARISON (NOT CROSSING AND CROSSING CLASSIFICATION) ON THE PSI DATASET.

Methods	Acc	F1
VR-GCN [71]	0.74	0.64
PIE-PCP [8]	0.69	0.79
PSI-PCP [56]	0.76	0.66
Gated-S2R-PCP	**0.78**	**0.80**

G. Further PCP Analysis on Near-Collision Scenes

To verify Gated-S2R-PCP for extra pedestrian crossing scenarios, this work takes all pedestrian crossing sequences (50 ones) in the challenging DADA-2000 [19] dataset (*abbrev.,* **PC50-DADA**), where each crossing pedestrian is hit by the ego-vehicle. Certainly, for further PCP analysis, we can also implement the proposed method to the practical perception system of self-driving vehicles by Tensor RT or some other platforms. However, it needs to pre-detect and track the pedestrians, which involves perception bias for the PCP evaluation. Therefore, we instead take the near-collision scenes here for an in-depth analysis of our Gated-S2R-PCP, to be with a closer relation to the safe-driving function in this work.

Unlike JAAD or PIE datasets that assume the observed pedestrian state of input is not crossing and predict the pedestrian state with a long TTC (τ) of 1-2 seconds, pedestrians in the sequences 86% of PC50-DADA appear suddenly and begin to cross the road without a long τ. Hence, for the 86% sudden crossing sequences, we fix the first 16 frames of each sequence as the observation and narrow τ (1 to 16 frames) to adapt to the input observation setting of Gated-S2R-PCP.

TABLE VI
PCP PERFORMANCE OF THE PROPOSED METHOD IN PC50-DADA.

TTC τ	1.0s	0.8s	0.6s	0.4s	0.2s	0.03s	all τ
Acc	0.63	0.60	0.57	0.58	0.57	0.53	0.66
Auc	0.65	0.62	0.61	0.59	0.63	0.58	0.68

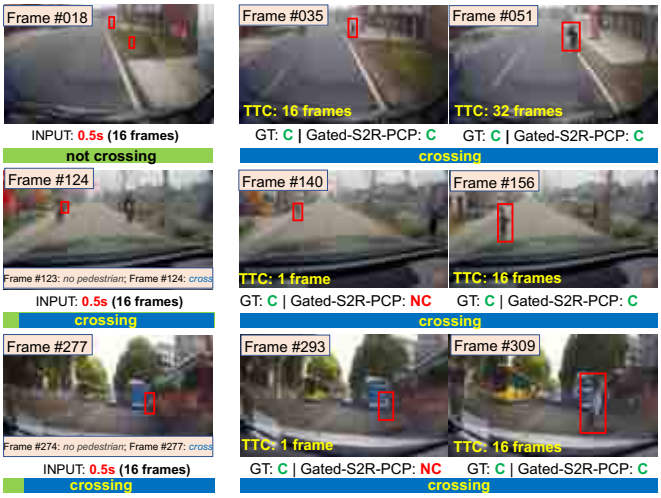

Fig. 12. The visualization of PCP results in PC50-DADA, where the red boxes mark the pedestrians whose crossing intention is to be predicted.

This setting may cause the last frame in observation to have a crossing state, as shown in the second and third rows of Fig. 12. In addition, 100 pedestrians not crossing on PC50-DADA are selected to serve the negative samples.

Table VI lists the results with varying τ. The performance of Gated-S2R-PCP increases when the pedestrian is close to the ego vehicle (larger τ), which meets our expectation because a larger τ means that the pedestrian at time $T+\tau$ has a higher risk of collision. Fig. 12 also presents the PCP results in three samples of PC50-DADA. In particular, the second and third rows of PCP results generate failure predictions at time $T+1$ and re-obtain accurate results for $T+16$. This observation indicates that PCP models prefer the large pedestrian appearance and location distances at time T and time $T+\tau$, and the near-collision involved PCP task has a large space to be improved mainly from two insights: 1) synthesizing more sudden pedestrian crossing samples with occlusion handling; 2) leveraging the crossing object grounding for long-tailed sample learning or event camera for sudden motion augmentation.

H. Discussions

1) Shared feature space analysis after DisA: In this work, DistA aims to narrow the feature space of synthetic semantic segmentation and depth maps and one of RGB frames in real data. We verify the effect of DistA by visualizing the feature point distribution using t-SNE [16]. Specifically, we randomly select 1000 samples (crossing or not-crossing clips) in the synthetic S2R-PCP-3181 dataset and the real JAAD or PIE dataset. Here, we adopt the Timesformer [18] as the feature backbone model and generate a 512-dimensional vector for

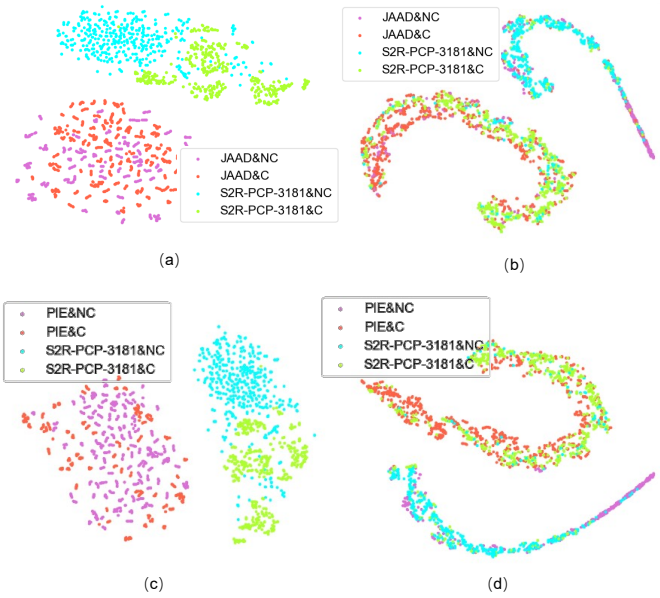

Fig. 13. Feature distributions of *Before-DistA* and *After-DistA* for pedestrian crossing and not-crossing using t-SNE [16] with 1000 samples from S2R-PCP-3181 or real datasets (taking 500 crossing and 500 not-crossing samples in each dataset). (a) and (c) denote the feature distributions of *Before-DistA*. (b) and (d) specify the feature distributions of *After-DistA*.

the observation (16 frames of 0.5 seconds) of each sample. We use t-SNE to reduce the feature dimension, *i.e.*, $512 \rightarrow 2$. Then the feature vectors are visualized in Fig. 13, where (a) and (c) are the feature distribution before DistA. (b) and (d) are the feature distribution after DistA.

The visualization shows that the confused crossing and not-crossing feature points are separated from the structured manifolds. Notably, from the feature distribution before DistA, we can see that the synthetic samples (semantic segmentation and depth maps) have more clearer margin for crossing and not-crossing behavior than the real data (RGB frames). Therefore, synthetic data's behavior knowledge can help guide the distinction of behavior patterns in real data via the distribution approximating process. In addition, there is a long-tailed distribution for the not-crossing samples, which may be caused by the more confused motion pattern compared with the crossing behaviors. Through this evaluation, we find the shared feature space between crossing or not-crossing in synthetic and real data obtained by the DistA.

2) The interactive role between FOL and PCP: As evaluated in the ablation studies, the FOL prediction has a positive role for PCP except for the Recall value. To further check the reason, we evaluate the interactive influence between FOL and PCP here. As for FOL, we adopt the average intersection of unit (**AIoU**,%) and final intersection of unit (**FIoU**,%) between the predicted bounding boxes and the ground truth. FIoU is computed by the predicted bbox in the last frame while AIoU focuses on all the prediction windows. Similarly, the trajectory point errors (Euclidean distance between the predicted trajectory points with the ground-truth), *i.e.*, average distance error (**ADE**, *pixels*) and final distance error (**FDE**, *pixels*), are also utilized. The results are displayed in Table. VII. We also present some prediction results of FOL with

TABLE VII
PERFORMANCE INFLUENCE PCP AND FOL (WITH 0.25S (8 FRAMES), 0.5S (16 FRAMES), AND 1.0S (32 FRAMES) PREDICTION) ON THE PIE DATASET, WHERE ↑ AND ↓ PREFER SMALL AND LARGE VALUE, RESPECTIVELY.

Pred. Set	Acc	Auc	F1	Pre	Rec	AIoU↑	FIoU↑	ADE↓	FDE↓
0.25s (8-f.)	0.89	0.89	**0.83**	0.81	**0.84**	44.17	24.06	**20.78**	**23.44**
0.5s (16-f.)	**0.92**	**0.90**	0.82	**0.83**	0.80	**47.72**	**42.13**	22.09	27.72
1.0s (30-f.)	0.88	0.87	0.80	0.81	0.78	39.14	24.04	44.05	68.50

Fig. 14. Some prediction results of future object location (FOL) with time window (0.25s, 0.5s, and 1.0s) on the PIE dataset.

different time windows in Fig. 14.

We can observe that 0.5s of FOL prediction can obtain the best PCP performance for Acc, Auc, and Pre values. The Recall value may be weakened by long-term FOL mainly because of the inconsistent behavior type in observation (not-crossing or wandering) and crossing in prediction. In addition, we can see FOL can provide a clear separation between crossing and not-crossing behaviors, as shown in the top-right image in Fig. 14. The woman may cross but the future bbox presents a clear not-crossing behavior along the roadside. Longer-term FOL generates weaker trajectory prediction performance (with larger ADE and FDE). In common sense, AIoU and FIoU are larger for the shorter-term FOL prediction, and vice versa. However, there is one interesting phenomenon the AIoU and FIoU of 0.5s are larger than the ones of 0.25s. The reason for this is that when predicting the FOL in a short time, the small pedestrian displacement has little influence on the FOL loss computation in training. When increasing the prediction window, the consistency of the movement direction has an important impact on the FOL loss computation in prediction. Besides, longer-term FOL prediction may involve a large loss value. Therefore, AIoU and FIoU increase from 0.25s to 0.5s predictions and decrease from 0.5s to 1.0s predictions.

3) Parameter Count and Inference Speed: To show the parameter count and inference speed, we list the parameter count (Millions) and inference time (ms) in Table VIII. Actually, it is difficult to conduct an absolute fair comparison of the inference speed because of the rapid updating of computation

TABLE VIII
PARAMETERS (MILLIONS) AND EFFICIENCY (MS) COMPARISON ON THE $JAAD_{all}$ TEST SET, WHERE TIME (MS) MEANS THE INFERENCE TIME FOR EACH FRAME.

Methods	Acc	Params.	time (ms)	Platform (GPU)
PCPA [58]	0.83	31.2M	194.6	1 GTX 1080ti
Yang *et al.** [65]	0.83	155.2M	416.8	1 GTX 1080ti
Pedestrian-Graph+* (32) [66]	0.86	128.1M	351.5	1 GTX 1080ti
PIT [68]	0.86	173M	4.9	1 RTX 3060
PedAST-GCN* [32]	0.88	108.3M	166.0	1 GTX 1080ti
Gated-S2R-PCP	0.88	176.7M	33.0	1 RTX 2080ti

Yang *et al.** pre-detects the human pose and semantic segmentation images by Openpose method [72] (52.3M) and and deeplab-v3 [73] (42M) approach, respectively. Pedestrian-Graph+* needs to pre-obtain the pedestrian pose and scene depth by Alphapose [74] (59.M) and a depth estimation approach [75] (68.5M). PedAST-GCN* also needs to pre-obtain the skeleton pose by HRNet [76] (63.6M) and semantic segmentation maps by deeplab-v3 [73] (42M).

platforms for different methods. In this work, for a rational comparison, we count all the parameters in the inference phase consisting of the pre-steps of each method, such as the pedestrian pose estimation and depth map estimation in Pedestrian-Graph+ [66]. The parameter counts and inference speed of PCPA [58], Yang *et al.* [65], Pedestrian-Graph+ [66], and PedAST-GCN* [32] take the reported values in PedAST-GCN [32] run by a GTX 1080ti GPU. Because all models take the ground-truth object bounding boxes in evaluation, we do not involve the time cost of object detection in this comparison. From the results, although our Gated-S2R-PCP have 176.7M parameters, mainly by two Timesformer backbones [60] (as shown in Fig. 6) for the Style Shifter and the Distribution Approximator of RGB frame encoding in inference phase, we can see that the inference efficiency of our Gated-S2R-PCP is competitive with real-time capability by a RTX 2080ti GPU.

V. CONCLUSIONS

This paper proposes the gated syn-to-real knowledge transfer model for pedestrian crossing prediction (Gated-S2R-PCP). Three knowledge transfer ways, *i.e.*, a Style Shifter, a Distribution Approximator, and a Knowledge Distiller, are designed to transfer the visual context, semantic and depth information, and the bboxes of pedestrians in synthetic data (our S2R-PCP-3181) to two real datasets (JAAD and PIE). A learnable gated unit (LGU) is modeled to adaptively fuse knowledge features for PCP. Through extensive experiments, the superiority of Gated-S2R-PCP is verified. Besides, we also evaluate the interactive role of FOL prediction with PCP, and 0.5s of FOL prediction is the best choice for the PCP task.

From the investigation of this work, some highlights can be concluded. (1) A differentiated knowledge transfer from synthetic to real data is promising when meeting different information domain gaps. (2) The extra FOL prediction task benefits the precision and accuracy of PCP. (3) The PCP task in the near-collision scenes is challenging and with large space to be improved from sudden crossing pedestrian grounding or event-camera-based sudden motion augmentation.

REFERENCES

[1] W. H. Organization, "Five eras of safety," https://www.nhtsa.gov/vehicle-safety/automated-vehicles-safety.

[2] E. Marti, M. A. De Miguel, F. Garcia, and J. Perez, "A review of sensor technologies for perception in automated driving," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 4, pp. 94–108, 2019.

[3] L. Li, C. Zhao, X. Wang, Z. Li, L. Chen, Y. Lv, N.-N. Zheng, and F.-Y. Wang, "Three principles to determine the right-of-way for avs: Safe interaction with humans," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7759–7774, 2021.

[4] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Understanding pedestrian behavior in complex traffic scenes," *IEEE Trans. Intell. Veh.*, vol. 3, no. 1, pp. 61–70, 2018.

[5] M. Prédhumeau, A. Spalanzani, and J. Dugdale, "Pedestrian behavior in shared spaces with autonomous vehicles: An integrated framework and review," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 438–457, 2023.

[6] A. Vial, W. Daamen, A. Y. Ding, B. van Arem, and S. Hoogendoorn, "Amsense: how mobile sensing platforms capture pedestrian/cyclist spatiotemporal properties in cities," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 1, pp. 29–43, 2020.

[7] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," in *ICCVW*, 2017, pp. 206–213.

[8] 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 *ICCV*, 2019, pp. 6261–6270.

[9] J. Fang, F. Wang, J. Xue, and T. Chua, "Behavioral intention prediction in driving scenes: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 8334–8355, 2024.

[10] L. Achaji, J. Moreau, T. Fouqueray, F. Aioun, and F. Charpillet, "Is attention to bounding boxes all you need for pedestrian action prediction?" in *IV*, 2022, pp. 895–902.

[11] J. Bai *et al.*, "Deep virtual-to-real distillation for pedestrian crossing prediction," in *ITSC*, 2022, pp. 1586–1592.

[12] Q. Miao *et al.*, "Parallel learning: Overview and perspective for computational learning across syn2real and sim2real," *IEEE CAA J. Autom. Sinica*, vol. 10, no. 3, pp. 603–631, 2023.

[13] K. Liu, L. Li, Y. Lv, D. Cao, Z. Liu, and L. Chen, "Parallel intelligence for smart mobility in cyberphysical social system-defined metaverses: A report on the international parallel driving alliance," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 6, pp. 18–25, 2022.

[14] M. N. Riaz, M. Wielgosz, A. G. Romera, and A. M. López, "Synthetic data generation framework, dataset, and efficient deep model for pedestrian intention prediction," in *ITSC*, 2023, pp. 2742–2749.

[15] F. Wang, J. Ibañez-Guzmán, and Y. Lv, "Call for v&viv: Verification and validation of intelligent vehicles," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 401–406, 2022.

[16] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[18] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, 2021, pp. 813–824.

[19] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: driver attention prediction in driving accident scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4959–4971, 2022.

[20] Y. Hashimoto, Y. Gu, L. Hsu, and S. Kamijo, "A probabilistic model for the estimation of pedestrian crossing behavior at signalized intersections," in *ITSC*, 2015, pp. 1520–1526.

[21] N. Schneider and D. M. Gavrila, "Pedestrian path prediction with recursive bayesian filters: A comparative study," in *GCPR*, J. Weickert, M. Hein, and B. Schiele, Eds., vol. 8142, 2013, pp. 174–183.

[22] R. D. Brehar, M. P. Muresan, T. Marita, C. Vancea, M. Negru, and S. Nedevschi, "Pedestrian street-cross action recognition in monocular far infrared sequences," *IEEE Access*, vol. 9, pp. 74 302–74 324, 2021.

[23] A. Rasouli *et al.*, "Pedestrian action anticipation using contextual feature fusion in stacked rnns," in *BMVC*, 2019, pp. 171–184.

[24] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Do they want to cross? understanding pedestrian intention for behavior prediction," in *IV*, 2020, pp. 1688–1693.

[25] ——, "Benchmark for evaluating pedestrian action prediction," in *WACV*, 2021, pp. 1258–1268.

[26] M. Azarmi, M. Rezaei, H. Wang, and S. Glaser, "Pip-net: Pedestrian intention prediction in the wild," *CoRR*, vol. abs/2402.12810, 2024.

[27] K. Saleh, M. Hossny, and S. Nahavandi, "Intent prediction of vulnerable road users from motion trajectories using stacked lstm network," in *ITSC*, 2017, pp. 327–332.

[28] M. Chaabane, A. Trabelsi, N. Blanchard, and R. Beveridge, "Looking ahead: Anticipating pedestrians crossing with future frames prediction," in *WACV*, 2020, pp. 2297–2306.

[29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[30] A. Rasouli and I. Kotseruba, "Pedformer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning," in *ICRA*, 2023, pp. 9844–9851.

[31] K. Kitchat *et al.*, "Pedcross: Pedestrian crossing prediction for auto-driving bus," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 8730–8740, 2024.

[32] Y. Ling, Z. Ma, Q. Zhang, B. Xie, and X. Weng, "Pedast-gcn: Fast pedestrian crossing intention prediction using spatial-temporal attention graph convolution networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13 277 – 13 290, 2024.

[33] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, "Pedestrian graph +: A fast pedestrian crossing prediction model based on graph convolutional networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21 050–21 061, 2022.

[34] 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 *ITSC*, 2019, pp. 2000–2005.

[35] S. Zhang, M. A. 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, 2022.

[36] X. Chen, S. Zhang, J. Li, and J. Yang, "Pedestrian crossing intention prediction based on cross-modal transformer and uncertainty-aware multi-task learning for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 12 538–12 549, 2024.

[37] Y. Kuznietsov, M. Proesmans, and L. V. Gool, "Towards unsupervised online domain adaptation for semantic segmentation," in *WACV*, 2022, pp. 261–271.

[38] Y. Jiao, H. Yao, and C. Xu, "SAN: selective alignment network for cross-domain pedestrian detection," *IEEE Trans. Image Process.*, vol. 30, pp. 2155–2167, 2021.

[39] H. Wang, S. Liao, and L. Shao, "AFAN: augmented feature alignment network for cross-domain object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 4046–4056, 2021.

[40] G. Li, Z. Ji, and X. Qu, "Stepwise domain adaptation (SDA) for object detection in autonomous vehicles using an adaptive centernet," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17 729–17 743, 2022.

[41] Y. Lv, Y. Chen, L. Li, and F.-Y. Wang, "Generative adversarial networks for parallel transportation systems," *IEEE Intell. Transp. Syst. Mag.*, vol. 10, no. 3, pp. 4–10, 2018.

[42] D. Rempe, J. Philion, L. J. Guibas, S. Fidler, and O. Litany, "Generating useful accident-prone driving scenarios via a learned traffic prior," in *CVPR*, 2022, pp. 17 305–17 315.

[43] J. Zhang, C. Liu, B. Wang, C. Chen, J. He, Y. Zhou, and J. Li, "An infrared pedestrian detection method based on segmentation and domain adaptation learning," *Comput. Electr. Eng.*, vol. 99, p. 107781, 2022.

[44] M. A. Marnissi, H. Fradi, A. Sahbani, and N. E. B. Amara, "Unsupervised thermal-to-visible domain adaptation method for pedestrian detection," *Pattern Recognit. Lett.*, vol. 153, pp. 222–231, 2022.

[45] C. Zhao, Y. Lv, J. Jin, Y. Tian, J. Wang, and F.-Y. Wang, "Decast in transverse for parallel intelligent transportation systems and smart cities: Three decades and beyond," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 6, pp. 6–17, 2022.

[46] F. Zhu, Y. Lv, Y. Chen, X. Wang, G. Xiong, and F. Wang, "Parallel transportation systems: Toward iot-enabled smart urban traffic control and management," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4063–4071, 2020.

[47] A. Dosovitskiy, G. Ros, F. Codevilla, A. M. López, and V. Koltun, "CARLA: an open urban driving simulator," in *CoRL*, vol. 78, 2017, pp. 1–16.

[48] T. Sun, M. Segù, J. Postels, Y. Wang, L. V. Gool, B. Schiele, F. Tombari, and F. Yu, "SHIFT: A synthetic driving dataset for continuous multi-task domain adaptation," in *CVPR*, 2022, pp. 21 339–21 350.

[49] J. Wang *et al.*, "A parallel teacher for synthetic-to-real domain adaptation of traffic object detection," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 441–455, 2022.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[51] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017, pp. 1510–1519.

[52] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680.

[53] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, vol. 37, 2015, pp. 1180–1189.

[54] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *ICLR*, 2017.

[55] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *CoRL*, 2017, pp. 1–16.

[56] T. Chen, R. Tian, Y. Chen, J. Domeyer, H. Toyoda, R. Sherony, T. Jing, and Z. Ding, "Psi: A pedestrian behavior dataset for socially intelligent autonomous car," *arXiv preprint arXiv:2112.02604*, 2021.

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

[58] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for Evaluating Pedestrian Action Prediction," in *WACV*, 2021, pp. 1257–1267.

[59] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *ICCV*, 2015, pp. 4489–4497.

[60] J. Lin, C. Gan, K. Wang, and S. Han, "TSM: Temporal Shift Module for Efficient and Scalable Video Understanding on Edge Devices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2760–2774, 2022.

[61] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *CVPR*, 2018, pp. 4194–4202.

[62] A. Rasouli, M. Rohani, and J. Luo, "Bifold and semantic reasoning for pedestrian behavior prediction," in *ICCV*, 2021, pp. 15 580–15 590.

[63] J. Lorenzo, I. Parra, and M. Á. Sotelo, "Intformer: Predicting pedestrian intention with the aid of the transformer architecture," *CoRR*, vol. abs/2105.08647, 2021.

[64] X. Zhang, P. Angeloudis, and Y. Demiris, "ST CrossingPose: A Spatial-Temporal Graph Convolutional Network for Skeleton-Based Pedestrian Crossing Intention Prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20 773–20 782, 2022.

[65] D. Yang *et al.*, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Trans. Intell. Veh.*, vol. 7, no. 2, pp. 221–230, 2022.

[66] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, "Pedestrian graph +: A fast pedestrian crossing prediction model based on graph convolutional networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21 050–21 061, 2022.

[67] M. Dong, "Pedestrian cross forecasting with hybrid feature fusion," in *ICLR Workshop on Scene Representations for Autonomous Driving*, 2023.

[68] Y. Zhou, G. Tan, R. Zhong, Y. Li, and C. Gou, "PIT: progressive interaction transformer for pedestrian crossing intention prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14 213–14 225, 2023.

[69] B. Yang, Z. Wei, H. Hu, R. Wang, Y. C. Chun, and R. Ni, "DPCIAN: A novel dual-channel pedestrian crossing intention anticipation network," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 6, pp. 6023–6034, 2024.

[70] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

[71] T. Chen, R. Tian, and Z. Ding, "Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention," in *ICCV Workshops*, 2021, pp. 3096–3102.

[72] Z. Cao, G. Hidalgo, T. Simon, S. 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, 2021.

[73] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.

[74] J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *CVPR*, 2019, pp. 10 863–10 872.

[75] J. H. Lee, M. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *CoRR*, vol. abs/1907.10326, 2019.

[76] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019, pp. 5693–5703.

Jie Bai received the Bachelor's and Master's degrees in Automation and Traffic Information Engineering and Control from Chang'an University in 2020 and 2023. He is currently an Engineering in BYD Co., Ltd.. His research interests include intelligent transportation and pedestrian intent prediction.

Jianwu Fang received a Ph.D. degree in SIP (signal and information processing) from the University of Chinese Academy of Sciences, China in 2015. He is currently an Associate Professor at the Institute of Artificial Intelligence and Robotics at Xian Jiaotong University, Xian, China. He was a visiting scholar at the NExT++ Research Centre in the School of Computing, National University of Singapore (NUS), Singapore. He is the chair of many special sessions of ITSC, IV, and ACM MM. He received the Best Paper Nominees Award in ITSC 2023. He is the Associate Editor of IEEE Transactions on Intelligent Transportation Systems and IEEE Open Journal of Intelligent Transportation Systems. His research interests include computer vision, multimedia analysis, and their applications in intelligent transportation.

Yisheng Lv (Senior Member, IEEE) received the B.E. and M.E. degrees in transportation engineering from Harbin Institute of Technology, in 2005 and 2007, respectively, and the Ph.D. degree in control theory and control engineering from the Chinese Academy of Sciences, in 2010. He is a Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. He is the EiC of IEEE Intelligent Transportation Systems and AE of IEEE Transactions on Intelligent Transportation Systems. His research interests include traffic data analysis, dynamic traffic modeling, and parallel traffic management and control systems.

Chen Lv (Senior Member, IEEE) received the Ph.D. degree from the Department of Automotive Engineering, Tsinghua University, China, in January 2016. He is currently an Associate Professor at the School of Mechanical and Aerospace Engineering and the Cluster Director of Future Mobility Solutions, Nanyang Technological University, Singapore. His research interests include intelligent vehicles, automated driving, and human-machine systems, where he has contributed two books, over 100 papers, and obtained 12 granted patents.

Jianru Xue (Member, IEEE) received his B.S. degree from Xi'an University of Technology in 1994, and both MS and Ph.D. degrees from Xi'an Jiaotong University in 1999 and 2003, respectively. He joined the Institute of Artificial Intelligence and Robotics at Xi'an Jiaotong University, Xi'an, China, in 1999, where he currently is a full professor. He worked in FujiXerox, Tokyo, Japan, from 2002 to 2003, and visited the University of California, Los Angeles, from 2008 to 2009. His research interests include computer vision, visual localization and navigation, and video coding based on analysis. He and his team were winners of the IEEE ITS Institute Lead Award in 2014. He and his students won the Best Application Paper award at the Asian Conference on Computer Vision 2012.

Zhengguo Li (Fellow, IEEE) received a B.Sc. degree in applied mathematics and an M.Eng. degree in automatic control from Northeastern University, Shenyang, China, in 1992 and 1995, respectively, and a Ph.D. degree in automatic control from Nanyang Technological University, Singapore, in 2001.

He is currently with the Agency for Science, Technology and Research, Singapore. Dr. Li has been actively involved in the development of H.264/AVC and HEVC since 2002. He served as a TPC Chair for IEEE IECON in 2023 and 2020, a special session Chair for IEEE ICASSP 2022, a Technical Brief Co-Founder for SIGGRAPH Asia, and a leading Workshop Chair for IEEE ICME in 2013. He is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology from 2020 to 2024; a Senior Area Editor of IEEE Transactions on Image Processing from 2020 to 2024; and the Editor of MDPI Sensors from 2022 to 2024 and APSIPA Transactions on Signal and Information Processing from 2022 to 2025. His current research interests include computational photography, mobile imaging, video processing and delivery, and switched and impulsive control. He is awarded IEEE Fellow and Asia-Pacific Artificial Intelligence Association Fellow.