# ESIA: Energy-Based Spatiotemporal Interaction-Aware Framework for Pedestrian Intention Prediction

Yanping Wu, Meiting Dang, Lin Wu, Edmond S. L. Ho, Zhenghua Chen, Chongfeng Wei*

***Abstract*— Pedestrian intention prediction aims to infer future crossing decisions by modeling temporal dynamics, social interactions, and environmental context. However, existing methods often rely on simplified interaction modeling and opaque reasoning processes, resulting in limited interpretability and inconsistent predictions. To address these issues, we propose an Energy-based Spatiotemporal Interaction-Aware framework (ESIA), formulated within a Conditional Random Field (CRF) paradigm. The scene is represented as a graph where pedestrians and the environment are treated as nodes. Unary potentials capture individual intentions, while pairwise potentials model pedestrian–pedestrian and pedestrian–environment interactions. These components are integrated into a unified global energy function to ensure coherent behavioral predictions across the scene. To enforce global consistency during inference, we introduce structural consistency constraints and solve the optimization using a Unary-Seeded Simulated Annealing (U-SSA) strategy that leverages reliable unary priors. Experiments on the JAAD and PIE datasets demonstrate that ESIA achieves superior performance and improved interpretability compared with state-of-the-art methods.**

***Index Terms*— Pedestrian Intention Prediction, Conditional Random Field, Social Interaction, Environmental Context**

## I. INTRODUCTION

Pedestrian intention prediction is a fundamental problem for autonomous driving and intelligent transportation systems. Despite recent progress, existing approaches remain limited by opaque interaction representations and fragmented modeling strategies. First, many methods oversimplify the complex interplay between individual intentions, social interactions, and environmental constraints, resulting in incomplete interaction modeling. Second, most approaches rely on black-box architectures that provide little interpretability, making it difficult to understand the rationale behind predictions. Third, the lack of a unified representation space often leads to predictions that are locally plausible but globally inconsistent with physical or social constraints. These limitations hinder the development of reliable and interpretable intention prediction systems in dynamic traffic scenarios. An ideal framework should therefore enable comprehensive interaction modeling, transparent reasoning, and globally consistent predictions within a unified formulation. To address this challenge, we propose an Energy-based Spatiotemporal Interaction-Aware (ESIA) framework, which explicitly models multi-level interactions under a unified energy-based formulation. Specifically, we represent individual intentions as unary potentials and model pedestrian–pedestrian (P–P) and pedestrian–environment (P–E) interactions as pairwise potentials, enabling explicit reasoning over factors. By integrating these components into a global energy framework with structural consistency constraints, the proposed approach produces interpretable and physically coherent predictions.

*Chongfeng Wei is the corresponding author.

Yanping Wu, Meiting Dang, Lin Wu, Zhenghua Chen, and Chongfeng Wei are with the James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, United Kingdom.

Edmond S. L. Ho is with the School of Computing Science, University of Glasgow, Glasgow, G12 8QQ, United Kingdom.

## II. PROPOSED METHOD

Unlike existing black-box approaches that implicitly fuse heterogeneous cues, ESIA explicitly models pedestrian intentions and their interactions within a unified Conditional Random Field formulation Fig. 1. **(1) Graph Construction:** Given an observation sequence, the scene is represented as an undirected graph where nodes correspond to pedestrians and a shared environment node. P–P edges capture social interactions, while P–E edges represent contextual constraints from the surrounding traffic environment. **(2) Interaction Modeling:** We define an energy function composed of unary and pairwise potentials on the graph. The unary potential models the intrinsic crossing intention of each pedestrian using visual appearance and motion cues. Two types of pairwise potentials are introduced to capture interactions: P–P interactions, which encode social behaviors such as group coherence and conflict avoidance. P–E interactions, which measure whether the environment supports crossing behavior. These interaction potentials transform complex spatiotemporal dependencies into measurable compatibility scores within the framework. **(3) Energy-based Inference:** The final intention configuration for all pedestrians is obtained by minimizing the total energy. This global inference jointly considers individual intentions, social interactions, and environmental constraints, enabling the model to correct locally inconsistent predictions and produce globally coherent decisions.

## III. PRELIMINARY RESULTS

*1) **Overall Performance Comparison**:* We evaluate ESIA on *pedestrian intention prediction task* against SOTA baseline methods on three real-world datasets, using Acc, P, R, F1, and AUC, as summarized in Table I. From the average results in the table, we observe that ESIA consistently outperforms most of the baseline methods. Specifically, ESIA demonstrates state-of-the-art performance on JAAD-all (Acc, P, R), PIE (P, F1), and JAAD-beh (Acc, AUC), while ranking among the top performers on other metrics. Notably, ESIA achieves absolute gains of **4**% in P on JAAD-all, **2**% in

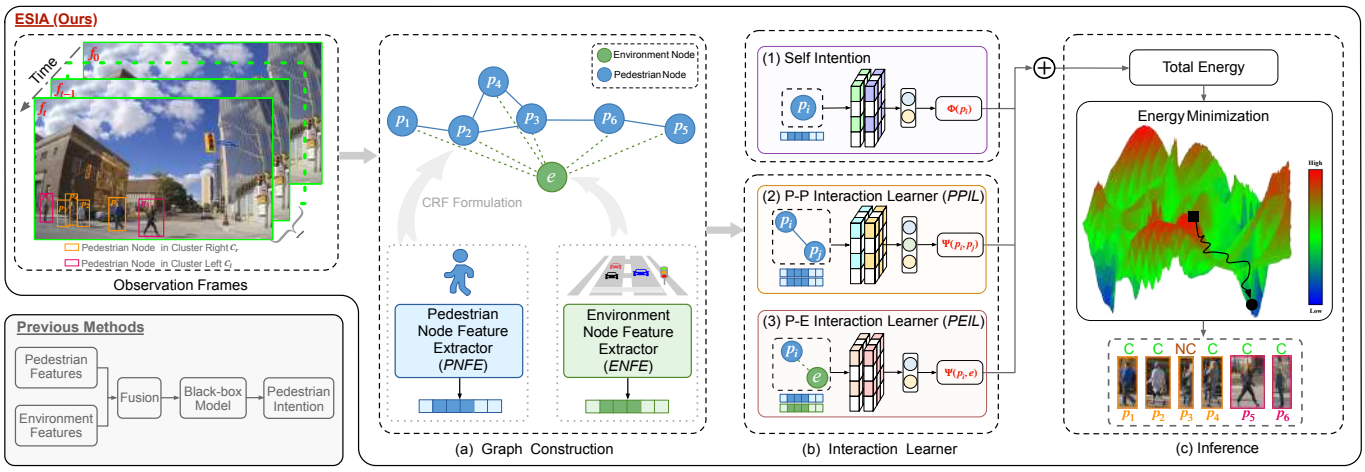


Fig. 1. Overview of ESIA. Unlike previous methods (*bottom left*), ESIA framework consists of three components: a) Graph Construction,; b) Interaction Learner; and c) Inference. Here, *NC* denotes Not Crossing, and *C* denotes Crossing.

TABLE I

OVERALL PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON JAAD AND PIE DATASETS. **BOLD**:BEST, UNDERLINED:SECOND BEST. INPUT: B = BOUNDING BOX, S = SPEED, P = POSE, I = IMAGE. BLOCKS: ATT = ATTENTION. MISSING VALUES ARE INDICATED BY –.

| Models | Year | Blocks | Inputs | JAAD-beh | | | | | JAAD-all | | | | | PIE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Acc | P | R | F1 | AUC | Acc | P | R | F1 | AUC | Acc | P | R | F1 | AUC |
| Single-GRU [1] | 2020 | GRU | B,S,I | 0.58 | 0.67 | 0.68 | 0.67 | 0.54 | 0.65 | 0.26 | 0.49 | 0.34 | 0.59 | 0.83 | 0.70 | 0.64 | 0.67 | 0.77 |
| Single-LSTM [1] | 2020 | LSTM | B,S,I | 0.51 | 0.63 | 0.59 | 0.61 | 0.48 | 0.78 | 0.44 | 0.70 | 0.54 | 0.75 | 0.81 | 0.67 | 0.61 | 0.64 | 0.75 |
| TrouSPI-Net [2] | 2021 | GRU+Att | B,S,P | 0.64 | 0.66 | 0.91 | 0.76 | 0.56 | 0.85 | 0.57 | 0.55 | 0.56 | 0.73 | 0.88 | 0.73 | **0.89** | 0.80 | 0.88 |
| PCPA [3] | 2021 | CNN+RNN+Att | B,S,P,I | 0.58 | – | – | 0.71 | 0.50 | 0.85 | – | – | 0.68 | 0.86 | 0.87 | – | – | 0.77 | 0.86 |
| PedGraph+[4] | 2022 | GCN+CNN | S,P,I | 0.70 | **0.77** | 0.75 | 0.76 | **0.70** | 0.86 | 0.58 | 0.75 | 0.65 | **0.88** | 0.89 | 0.83 | 0.79 | 0.81 | 0.90 |
| FFSTP [5] | 2022 | CNN+GRU+Att | B,S,P,I | 0.62 | 0.65 | 0.85 | 0.74 | 0.54 | 0.83 | 0.51 | 0.81 | 0.63 | 0.82 | 0.89 | 0.79 | 0.81 | 0.80 | 0.86 |
| PIT [6] | 2023 | Att | P,S,I | 0.70 | 0.71 | 0.93 | **0.81** | 0.65 | 0.87 | 0.54 | **0.85** | 0.66 | 0.87 | 0.91 | 0.85 | 0.79 | 0.82 | 0.90 |
| PedCMT [7] | 2024 | Att | B,S | 0.71 | – | – | 0.80 | 0.64 | **0.88** | – | – | 0.65 | 0.77 | **0.93** | – | – | **0.87** | **0.92** |
| PPCI_att [8] | 2024 | LSTM+Att | B,S,P | 0.67 | – | – | 0.77 | 0.60 | 0.81 | – | – | **0.75** | 0.78 | 0.91 | – | – | 0.84 | 0.89 |
| PedSA [9] | 2025 | Att+ViT | B,S,P,I | 0.67 | 0.68 | 0.90 | 0.77 | 0.60 | 0.83 | 0.47 | 0.82 | 0.62 | 0.80 | – | – | – | – | – |
| LSOP-Net [10] | 2025 | GRU+Att+CNN | B,S,P,I | 0.65 | 0.65 | **0.98** | 0.78 | 0.54 | 0.85 | 0.56 | 0.61 | 0.58 | 0.75 | 0.89 | 0.80 | 0.82 | 0.81 | 0.87 |
| **ESIA (Ours)** | 2026 | Att+ViT+CRF | B,S,I | **0.73** | 0.73 | 0.87 | 0.79 | **0.70** | **0.88** | **0.62** | **0.85** | 0.72 | 0.87 | 0.92 | **0.86** | 0.88 | **0.87** | 0.91 |

Acc on JAAD-beh, and **1**% in P on PIE. From the results across three datasets, we also observe some lower scores in certain metrics (e.g., P on JAAD-beh and F1 on JAAD-all) compared to certain baselines. The underlying reason lies in the optimization trade-off inherent in learning from imbalanced scenarios (e.g., JAAD and PIE). Prioritizing overall robustness and global consistency inevitably leads to fluctuations between P and R across different subsets.

## REFERENCES


[1] Kotseruba I, Rasouli A, Tsotsos J K. Do they want to cross? understanding pedestrian intention for behavior prediction[C]//2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2020: 1688-1693.

[2] Gesnouin J, Pechberti S, Stanciulcscu B, et al. TrouSPI-Net: Spatio-temporal attention on parallel atrous convolutions and U-GRUs for skeletal pedestrian crossing prediction[C]//2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). IEEE, 2021: 01-07.

[3] Kotseruba I, Rasouli A, Tsotsos J K. Benchmark for evaluating pedestrian action prediction[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 1258-1268.

[4] Cadena P R G, Qian Y, Wang C, et al. Pedestrian graph+: A fast pedestrian crossing prediction model based on graph convolutional networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 21050-21061.

[5] Yang D, Zhang H, Yurtsever E, et al. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention[J]. IEEE Transactions on Intelligent Vehicles, 2022, 7(2): 221-230.

[6] Zhou Y, Tan G, Zhong R, et al. PIT: Progressive interaction transformer for pedestrian crossing intention prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(12): 14213-14225.

[7] Chen X, Zhang S, Li J, et al. Pedestrian crossing intention prediction based on cross-modal transformer and uncertainty-aware multi-task learning for autonomous driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(9): 12538-12549.

[8] Alofi A, Greer R, Gopalkrishnan A, et al. Pedestrian safety by intent prediction: A lightweight lstm-attention architecture and experimental evaluations with real-world datasets[C]//2024 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2024: 77-84.

[9] Elgazwy A, Elgazzar K, Khamis A. Predicting pedestrian crossing intentions in adverse weather with self-attention models[J]. IEEE Transactions on Intelligent Transportation Systems, 2025.

[10] Liu H, Liu C, Chang F, et al. Long–Short Observation-driven Prediction Network for pedestrian crossing intention prediction with momentary observation[J]. Neurocomputing, 2025, 614: 128824.