# SA-BiGCN: Bi-Stream Graph Convolution Networks With Spatial Attentions for the Eye Contact Detection in the Wild

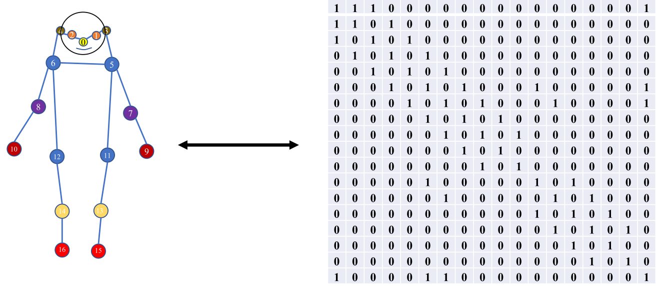
Yancheng Ling, Zhenliang Ma, *Member, IEEE*, Bangquan Xie, *Member, IEEE*,
Qi Zhang, and Xiaoxiong Weng

***Abstract*— Eye contact is essential in transmitting information and intention in the wild environment (e.g., urban streets or parking lots) with mixed vehicles and pedestrians. Compared with the vision image data, the human skeleton data are deemed to be robust to unconstrained surroundings and illumination. However, the skeleton graph-based approaches are mainly used for the action recognition. It is challenging to directly apply them to the eye detection task, which is momentary and dynamic given the complex wild environment. This paper proposes a Bi-stream Spatial Attention Graph Convolution Network (SA-BiGCN) for eye contact detection in the wild. We design a directed, nose-centric skeleton graph to capture relevant and hierarchical information and their interactions. We also propose a Bi-stream graph convolution network model with spatial attention to dynamically extract and fuse skeleton joints and bones information. The model was validated by comparing with state-of-art models on three large-scale public datasets, including JAAD, PIE, and LOOK. The results highlight the accuracy and generalization performance of the proposed SA-BiGCN model in detecting the eye contact in the wild environment. The ablation analysis validates the importance of the skeleton graph design, the spatial attention mechanism in the feature fusion process, as well as the model robustness against noisy skeleton data in terms of part occlusions, block occlusions, random occlusions, and random deviations.**

***Index Terms*— Eye contact detection, skeleton graph, graph convolution networks, spatial attention.**

## I. Introduction

EYE contact is a salient visual signal for humans and is one of the most critical signals for communicating intentions. Eye contact plays a vital role in reality, such as identifying the face awareness [1], perceiving the human-robot interaction quality [2], [3], and detecting lies [4]. With the rapid development of intelligent vehicles, eye contact detection becomes increasingly important in understanding the intention of surrounding pedestrians in the autonomous driving environment [5], [6], [7].



Natural skeleton graph

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Uni-labeling adjacency matrix

Fig. 1. The natural connections of human body joints and the corresponding uni-labeling adjacency matrix.

Many studies on eyes contact detection use vision images taken close to a person with a clear facial appearance [8], [9], [10], [11], [12]. However, pedestrian eye detection in the wild uses distant images or videos from vehicle sensors which brings great challenges to the problem given the image quality, unconstrained surroundings and illuminations [6], [7]. To address that, Belkada et al. [7] developed a key points-based eye contact detection model. Compared with Red-Green-Blue (RGB) image-based approaches [13], [14], the key-points based model achieves the best results on three tested public data sets [7], [14], [15]. The main reason is that the key-points based skeleton data are more robust against variations of image backgrounds.

Belkada et al. [7] used a simple fully-connected network with residual blocks as the main structure and utilized the normalized coordinate and confidence score in the image plane as the input. However, the proposed model treats human joints as independent features and limits its capability to fully use inherent graphically structural information. The fully-connected network and Convolutional Neural Network (CNN) models are capable of extracting information from euclidean data but are limited in capturing the inherent structure between human joints in the non-euclidean skeleton data. The Graph Convolution Network (GCN) was able to address that challenge. For example, Yan et al. [16] proposed the

Manuscript received 6 March 2023; revised 17 June 2023; accepted 25 August 2023. Date of publication 11 September 2023; date of current version 2 February 2024. The Associate Editor for this article was L. M. Bergasa. *(Corresponding author: Zhenliang Ma.)*

Yancheng Ling, Bangquan Xie, and Xiaoxiong Weng are with the School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510000, China (e-mail: lingyancheng@126.com; ctbqx51@mail.scut.edu.cn; ctxxweng@scut.edu.cn).

Zhenliang Ma and Qi Zhang are with the Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, 11428 Stockholm, Sweden (e-mail: zhema@kth.se; qzhan@kth.se).

This article has supplementary downloadable material available at https://doi.org/10.1109/TITS.2023.3310194, provided by the authors.

Digital Object Identifier 10.1109/TITS.2023.3310194

© 2023 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

spatial-temporal GCN (ST-GCN) model for action recognition using the sequential skeleton data and achieved the best performance on public data sets, including Kinetics [17] and NTU-RGB+D [18]. Following that, several model variants were proposed to recognize actions focusing on effectively representing and learning dependencies of spatial features and long-term temporal features [19], [20].

Compared with the action recognition task, the eye contact detection task has its unique challenges: (1) the eye contact is momentary, and its detection is based on a single image rather than a sequence of images in the action recognition; (2) the detection context is complex and the feature importance varies under different scenarios. For example, the head direction is important if a pedestrian is right or left against a vehicle, while the eye gaze direction is vital if the pedestrian faces straight to the vehicle. Therefore, it is not reasonable to directly apply the action recognition model to the eye contact detection problem.

The skeleton graph and its adjacent matrix are crucial for GCN based information aggregation and update [21]. Most studies used the natural connections of human body joints to design the graph and construct the corresponding uni-labeling adjacency matrix (Fig. 1.). The graph based on the natural connections of the whole human body joints can capture internal structures between joints, which is suitable for the action recognition tasks. However, eye contact detection has different characteristics for joints [7]. For example, the joints corresponding to the head (e.g., eyes and ears) have a critical impact on eye contact detection, while the joints for the legs and hands may not. It is important to design a special graph to extract more abundant information from the upper part of the body. In addition, unlike the skeleton sequence (capturing complex spatial-temporal joints correlations), the single skeleton image concentrates more on the joint spatial fusion in which the joints may contribute differently to eye contact detection. However, no study was found on dynamically representing and fusing the joints depending on the detection context, as discussed before.

In addition, the bones information of the skeleton graph was reported significantly contribute to the action recognition [22] since the lengths and moving directions of bones are naturally informative and discriminative for different actions. Similarly, the bones also add additional information for eye contact detection. For example, the lengths of the head and body bones would be visually different depending on whether a person looks at a camera. Therefore, it is potentially useful to incorporate the bone information in the skeleton graph representation, but yet receives no attention in the literature.

To address these gaps, we propose a Bi-stream graph convolution network model with spatial attention for momentary eye contact detection in the wild. It uses a novel-designed skeleton graph to generate the uni-labeling adjacent matrix for information aggregation and update. It dynamically extracts and fuses the human skeleton joints and bones information using the SA-BiGCN model. The main contributions are:

1) Proposing a Bi-stream graph convolution network model with spatial attention for momentary eye contact detection in the wild. It dynamically extracts and fuses skeleton joints and bones information.
2) Proposing a directed, nose-centric skeleton graph to capture relevant and hierarchical information and their interactions for eye contact detection, including the body posture, head posture, and eye gaze direction.
3) Validating the model on three large-scale public data sets for eye contact detection in the wild and conducting ablation studies on various aspects, including the skeleton graph topology, the components of the SA-BiGCN model, and the robustness against noisy datasets.

The remaining paper is structured as follows. Section II reviews related studies for skeleton data based recognition tasks. Section III proposes the SA-BiGCN framework and details its methodology. Section IV validates the model performance and explores its characteristics. Section V summarizes the main conclusions and future research directions.

## II. Related Work

We reviewed the skeleton data-based recognition tasks, such as actions, hand gestures, and eye contact. The synthesis focuses on GCN-based recognition models from aspects of skeleton graph and adjacency matrix designs and GCN attention mechanisms. We also discuss in detail the closely related studies for eye contact detection tasks.

### A. Skeleton Graph and Adjacency Matrix Designs

The GCN-based models have been widely used for skeleton data-based recognition tasks, such as action recognition [16], [19], [20], [22], [23], [24], gait recognition [25], [26], hand gesture recognition [27], [28], etc. For example, Yan et al. [16] proposed the spatio-temporal GCN model (ST-GCN) for action recognition using the natural connections of human body joints and three different strategies to design the adjacency matrix. Following that, several variants of the ST-GCN model have been proposed to design the skeleton graph and adjacency matrix better to facilitate effective learning in the recognition [19], [22], [23], [24]. For example, Shi et al. [22] developed an adaptive GCN to adaptively learn the skeleton graph's topology, which automatically learns the skeleton graph's spatial structure (rather than setting it manually). Li et al. [23] introduced an encoder-decoder model to adaptively learn the action-specific latent dependencies and construct the data-driven topology of the skeleton graph. They also extended existing skeleton graphs to represent higher-order dependencies of structural links. Finally, they combined the actional links and structural links into a generalized skeleton graph. Liu et al. [24] proposed disentangling and unifying graph convolutions to capture long-range joint dependencies and complex spatial-temporal dependencies for action recognition. Shi et al. [19] proposed a directed acyclic graph to represent the skeleton data, which could incorporate the joint and bone data more effectively. They developed a directed graph neural network(DGNN) to extract the feature from joints and bones and used a data-driven method to obtain an adaptive graph specifically. The improved graph and the adjacency matrix are far from the original ones in ST-GCN. These works highlight the critical importance of designing the skeleton graph and adjacency matrix in improving recognition performance by effectively capturing the intrinsic structures between joints.

### B. Attention Mechanism in GCN-Based Recognition Models

The attention mechanism has been widely used in GCN models to improve the model performance in capturing the complex and long-range dependencies [29], [30], [31], [32], [33]. For example, Xie et al. [29] introduced the spatial-temporal attention mechanism and designed an attention adjacency matrix (AAM) to capture implicit joint correlations and improve its generalizability ability to diverse skeletons. Heidari et al. [30] proposed a temporal attention model to select the most informative skeletons of actions and improve computational efficiency. Ahmad et al. [31] introduced the spatial attention model to eliminate the irrelevant skeleton joints and reserve relevant ones for action detection. Xing et al. [32] proposed an adaptive spatial attention model to capture dynamic relations between joints, supplementing the natural connections of human skeleton joints. Qiu et al. [33] introduced the temporal attention model and developed a spatial-temporal segments attention-based method to extract the correlation of different joints between consecutive video frames. Different from existing studies using attentions to extract spatial-temporal relationships between skeleton joints, we use attentions to dynamically fuse joints and bones information to improve eye contact detection performance.

### C. Eye Contact Detection in the Wild

Unlike the traditional eye contact detection task at a close distance or under controlled laboratory environments, eye contact detection in the wild is more challenging at a long distance and with unconstrained surroundings. Existing studies are categorized into image or skeleton based models depending on the input data. For example, Rasouli et al. [14] used the pedestrians' cropped images as inputs and applied the AlexNet model [34] to extract image features and make classifications. Mordan et al. [13] introduced multi-task fields for pedestrian eye contact detection by fusing both the pedestrian appearance and the environmental information. However, the image-based method is limited by the image quality and complex background information in the natural environment. Belkada et al. [7] established a public data set for the eye contact in the wild and introduced the skeleton-based method for eye contact detection. However, they used a fully connected neural network to extract features from skeleton joint coordinates directly and thus could not fully utilize the intrinsic human joint relationships in the skeleton graph. Recently, Hata et al. [6] considered the motion information for eye contact detection in the wild. They used the skeleton sequences as inputs and introduced the MS-G3D model [24] to extract spatial-temporal features and make classifications.

## III. Methodology

### A. Problem Definition and Framework

Eye contact detection is formulated as a binary classification task based on the human skeleton graph. Denote the human skeleton graph as G = (V, E), where $V = \{v_1, v_2, \ldots, v_N\}$ is the set of nodes, $E = \{e_1, e_2 \ldots, e_N\}$ is the set of edges. The eye contact detection problem is formally defined as: *Given the skeleton data extracted from distant images in the wild, infer whether a person has eye contact with other objects.*

The eye contact detection in the wild is challenging due to distance and unconstrained surroundings. Compared to directly using images, the skeleton data is less affected by noise in the environment. The paper proposes a Bi-stream spatial attention GCN-based methodology for the eye contact detection. Fig. 2. shows the overall architecture of SA-BiGCN. We employ the pose detector OpenPifPaf [35] to extract the keypoints of the human body from the pedestrian crop. Based on the keypoints, We adopt the joint stream and bone stream as our two streams for eye contact detection. The two streams use the same network structure but use different input data. In the joint stream, the input comprises the joints' coordinates and confidence scores in each frame. In the bone stream, we use the bones' coordinates and confidence scores in each frame as the input. For each stream, the process begins by initializing the skeleton graph. Next, we employ the Spatial Graph Convolutional Network (SA-GCN-Model) to extract features and generate the probability of eye contact. Finally, we combine the weighted probabilities of joint and bone to determine the ultimate probability of establishing eye contact.

The SA-BiGCN model comprises three core modules: skeleton graph design, Feature extracting and eye contact probability, and Bi-stream fusion.

1) Skeleton graph design. It designs different variants of the naturalistic skeleton graph to capture the problem-specific graph representation for the eye contact detection in the wild.
2) Feature extracting and eye contact probability. It takes inputs of the joints or bones (coordinates and confidence score) and uses the SA-GCN-Model to generate the probability of eye contact for the joint or bone stream. It dynamically fuses hidden representations features from different joints or bones using spatial attentions.
3) Bi-stream fusion. It consists of the joint stream and bone stream, which receive inputs of joint and bone coordinates, along with their corresponding confidence scores, which then estimate the probability for each respective stream. The network subsequently generates a weighted average probability for both joints and bones, ultimately determining the final detection results.

### B. Skeleton Graph Design

The skeleton graph is crucial for GCN model [19], [22] [23], [24]. We design five different skeleton graphs for eye contact detection (Fig. 3.), including (a) the natural connection graph (NCG), (b) the adaptive connection graph (ACG), (c) the upper body natural connection graph (uNCG), (d) the upper body connection with nose-centric graph (uNCG-N), and (e) the upper body natural connection with nose-centric and intermediate joints graph (uNCG-NI). We use the color and number of the skeleton to differentiate each point (representing key joints of the human skeleton). The color is used to distinguish between point pairs or individual points, while the

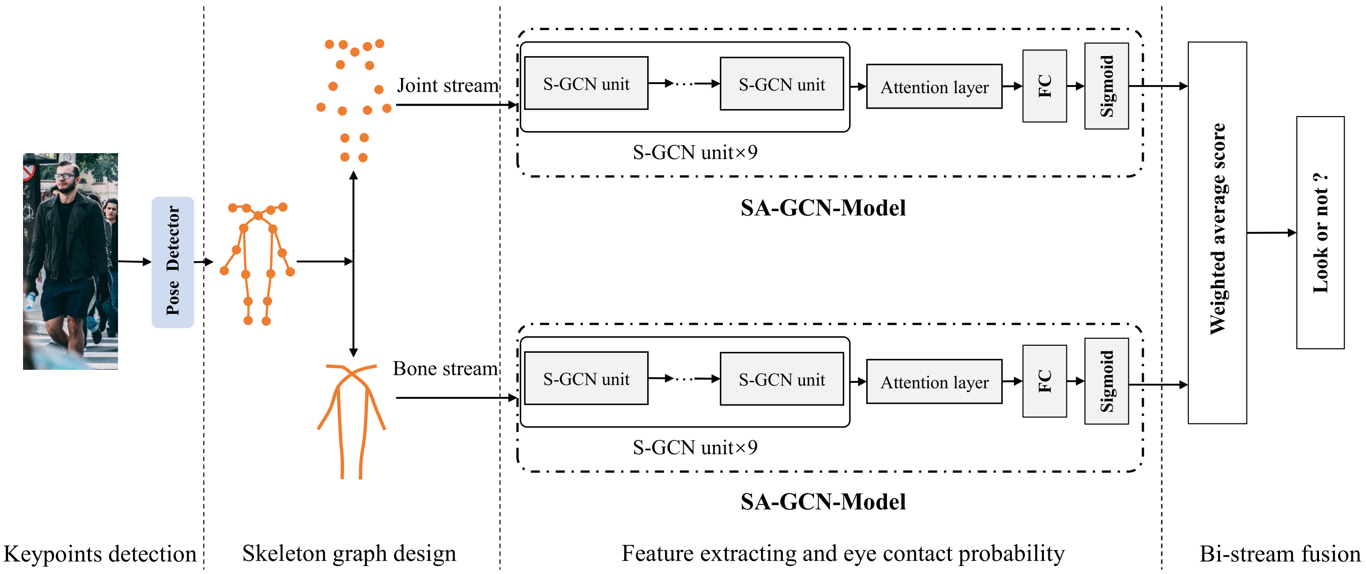


Fig. 2. The SA-BiGCN model for eye contact detection. The pose detector OpenPifPaf [35] is used to extract 17 keypoints from the pedestrian crop. Subsequently, the joint and bone data are derived from these keypoints through a reprocessing step. The SA-GCN-Model performs feature extraction and produces the probability for eye contact separately for both the joint and bone stream data. The joint probability and bone probability are weighted and combined to calculate the final probability of eye contact.

number is used to differentiate each point within a pair. The colors and numbers for the same skeleton joints are consistent in all sub-figures. For example, the upper shoulder joint points are assigned a blue color and the left joint is numbered 5 while the right one 6. For the skeleton graph with less number of joints (e.g., Fig.3(c)), we used the number with a prime to differentiate between points that share the same numerical label but are distinct entities (e.g., 7 in Fig.3(a) versus 7’ in Fig.3(c)).

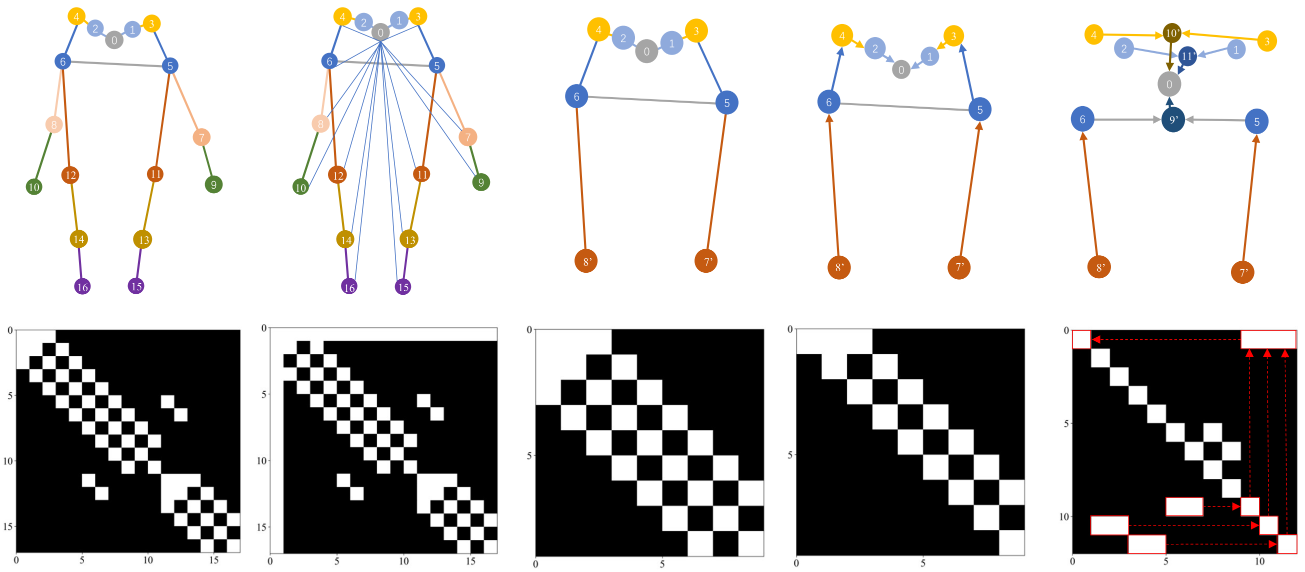
The NCG is formed by the natural physical connection of the human skeleton. The ACG is generated by adding adaptive and data-dependent connections in the natural connection graph. Compared to the NCG, the ACG contains additional connections between joints that are not physically adjacent. The uNCG removes the edges and points of the human legs and hands in the NCG. The uNCG-N adds directed links to the uNCG and connects all edges directly to the nose joint. Based on the uNCG-N graph, the uNCG-NI further adds three intermediate joints for the ear, eye, and shoulder, respectively. The intermediate joint is the middle point of the symmetric joints (ear/eye/shoulder). For example, the intermediate joint-9’ is the middle point between joint-5 and joint-6.

The corresponding uni-labeling adjacency matrices of these graphs are shown in the bottom row of Fig. 3. Compared to the uni-labeling adjacency matrix of the NCG, the adjacency matrix of the proposed uNCG-NI graph is unsymmetrical and directed, capturing the heterogeneous importance of joints and their interaction directions in the representation. In addition, the uNCG-NI contains additional nodes aggregating and transferring local information to important nodes for the problem of interest. That is, it has a hierarchical information aggregation and exchange as red rectangles and arrows shown in Fig. 3(e).

As discussed, the eye contact detection task has its own characteristics in which the joints corresponding to the head (e.g., eyes and ears) have a critical impact, while the importance of these for legs and hands is marginal [7]. Therefore, the NCG skeleton graph could be inefficient for the eye contact detection task. Compared to the NCG skeleton graph, the proposed uNCG-NI skeleton graph has the following characteristics and advantages: (1) it removes irrelevant joints for eye contact detection, which decreases the potential influence of redundant information in the learning and detecting processes. (2) It adds intermediate joints for important joints (ears, eyes, and shoulders), which enriches the hierarchical information aggregation and exchange between important joints. The hypothesis is that the added intermediate joints could help the model fuse the symmetrical joint information and contribute to posture measuring. (3) It is a directed and acyclic graph and gathers all information to the center nose joint. As mentioned in [36] and [37], the GCN-based model becomes over smoothing as the increase in the number of network layers since the node representation in the same connected component tends to converge to the same value for information aggregation and update. The directed and acyclic graph makes the information flow in one direction and relieves the over-smoothing issue in GCN learning.

### C. Feature Extracting and Eye Contact Probability

After initializing the graph, the SA-GCN-Model is employed to extract features and generate the probability of positive class for the joint or bone data. Fig. 4 shows the backbone structure of the SA-GCN-Model, which consists of $L_n$ S-GCN unit, a normalization layer, an attention layer, a fully connected layer (FC), and a Sigmoid function. The number of each S-GCN unit output channel is 64, 64, 64, 128, 128, 128, 256, 256, and 256, respectively. It takes inputs of joints (bones) coordinates and confidence scores and extracts their hidden representations using sequentially connected



(a) (b) (c) (d) (e)

Fig. 3. Skeleton graph representations. The first row shows five different skeleton graphs (a) the natural connection graph (NCG), (b) the adaptive connection graph (ACG), (c) the upper body natural connection graph (uNCG), (d) the upper body natural connection with nose-centric graph (uNCG-N), (e) the upper body natural connection with nose-centric and intermediate joints graph (uNCG-NI). The bottom row shows the corresponding uni-labeling adjacency matrices. The red rectangle and arrows in (e) represent the information aggregation and interaction flows.

basic modules. The hidden representations of various nodes (joints and bones) are fused using attention weights in the attention layer to obtain the final representation. This final representation is then passed through the fully connected layer and Sigmoid function to generate the probability of eye contact.

*1) The S-GCN Unit:* The S-GCN unit contains a series of connected basic modules. The basic module consists of the Conv-s, Bias, and a ReLU activation function (Fig. 4). In the proposed SA-GCN model, we add the residual connection to the output of spatial GCN to stabilize the training. Besides, we use the learnable edge importance weighting $\mathbf{W_e}$ (Eq. (1)) to capture the importance of different edges in the graph. The Conv-s is the core layer aggregating and updating information for each vertex. Given the original vertexe state matrix $\mathbf{X} \rightarrow \mathbb{R}^{N \times C_0}$, the updated vertex hidden state matrix $\mathbf{H}_{\mathbf{L_n}}^{\mathbf{Conv-s}} \rightarrow \mathbb{R}^{N \times C_{L_n}}$ is calculated as:

$$\mathbf{H}_{\mathbf{L_n}}^{\mathbf{Conv-s}} = \begin{cases} \mathbf{W_e^{L_n}} \odot \mathbf{AXW_c^{L_n}}, & L_n = 1 \\ \mathbf{W_e^{L_n}} \odot \mathbf{AH_{L_n-1}^{Conv-s}W_c^{L_n}}, & L_n >= 2 \end{cases} \qquad (1)$$

where $\mathbf{W_e^{L_n}} \rightarrow \mathbb{R}^{N \times N}$ is the learnable edge weights, $\mathbf{A} \rightarrow \mathbb{R}^{N \times N}$ is the adjacent matrix of the spatial graph, and $\mathbf{W_c^{L_n}} \rightarrow \mathbb{R}^{C_{L_n-1} \times C_{L_n}}$ a learnable weight matrix.

*2) The Attention Layer:* The attention mechanism [38] is widely used for spatial-temporal information learning tasks, for example, learning the graph edge importance and constructing the self-adapting skeleton structure, as well as learning the interactions between video frames in the action recognition task. Eq. (1) introduces the learnable edge weighting $\mathbf{W_e}$ to learn the importance of skeleton edges in the spatial GCN module. To improve the model's robustness, we further introduce the attention mechanism to fuse the spatial joint and bone features. The output hidden feature tensor of the Layer normalization in Fig. 4 is denoted as $\mathbf{H} \rightarrow \mathbb{R}^{N \times C'}$. Fig. 5. shows the proposed attention layer, consisting of four operations:

(1) Channel conversion: The conversion tensor $\mathbf{H_d} \rightarrow \mathbb{R}^{N \times C''}$ is calculated as:

$$\mathbf{H_d} = \mathbf{HW_d} \qquad (2)$$

where $\mathbf{W_d} \rightarrow \mathbb{R}^{C' \times C''}$ represents learnable weights.

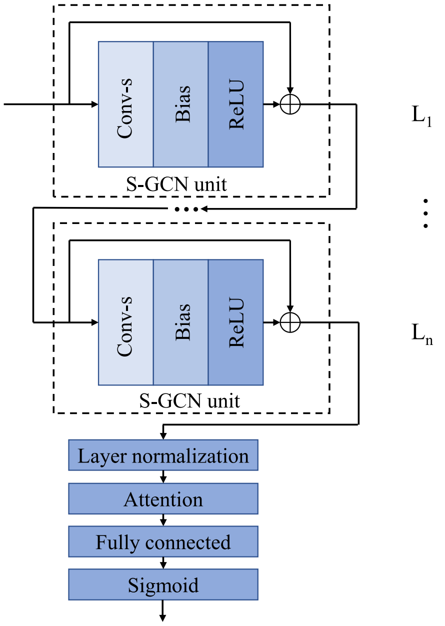


Fig. 4. The backbone of the SA-GCN model. It consists of $L_n$ S-GCN unit layers and a normalization layer, an attention layer, a fully connected layer, and a Sigmoid function. Each S-GCN unit layer contains the Conv-s, Bias, and a ReLU activation function.

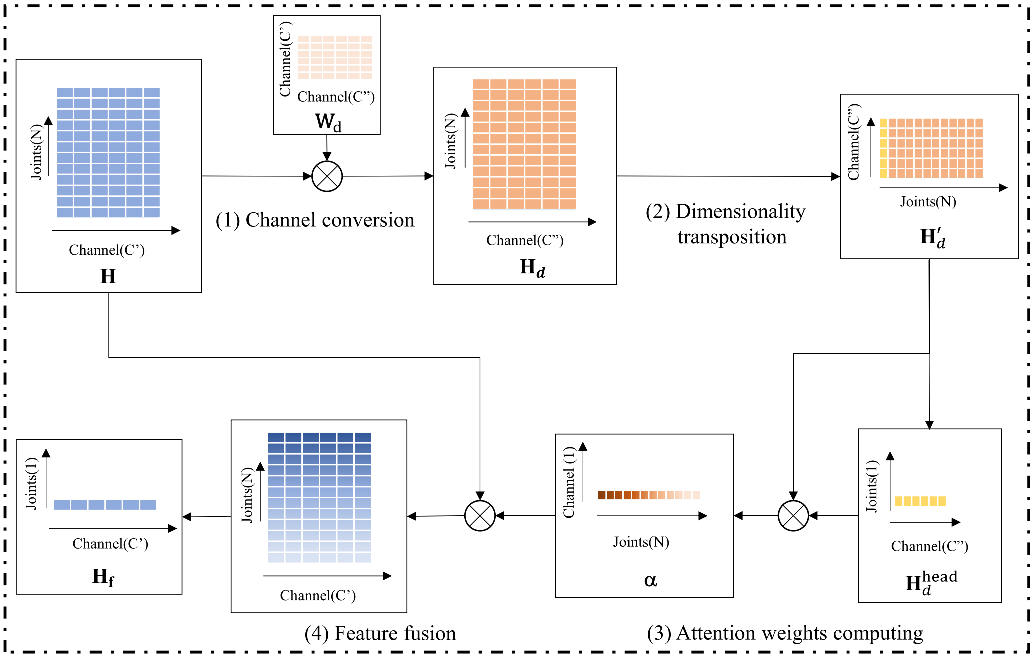


Fig. 5. The spatial attention for the skeleton joint and bone information fusion. It includes four operations:(1) Channel conversion, (2) Dimensionality transposition, (3) Attention weights computing, and (4) Feature fusion.

(2) Dimensionality transposition: The transposition tensor $\mathbf{H'_d} \rightarrow \mathbb{R}^{C'' \times N}$ of $\mathbf{H_d}$ is calculated as:

$$\mathbf{H'_d} = (\mathbf{H_d})^T \tag{3}$$

(3) Attention weights computing: Define the head tensor $\mathbf{H_d^{head}} \rightarrow \mathbb{R}^{1 \times C''}$, as the feature vector of the first vertex (joint or bone). The attention tensor $\alpha \rightarrow \mathbb{R}^{1 \times N}$ is calculated as:

$$\mathbf{H_d^{head}} = (\text{expand\_dims}(\mathbf{H'_d}[:, 0], \text{axis} = 1))^T \tag{4}$$

$$\alpha = \mathbf{H_d^{head}} \mathbf{H'_d} \tag{5}$$

(4) Feature fusion: Given attentions $\alpha$ and hidden feature tensor $\mathbf{H}$, the fusion tensor $\mathbf{H_f} \rightarrow \mathbb{R}^{1 \times C'}$ is calculated as:

$$\mathbf{H_f} = \alpha \mathbf{H} \tag{6}$$

*3) Output Layer and Loss Function:* Given $\mathbf{H_f}$, the probability of eye contact $p$ is estimated using a fully connected layer followed by a Sigmoid function.

$$p = \frac{1}{1 + e^{-\mathbf{H_f W_p}}} \tag{7}$$

where $\mathbf{W_p} \rightarrow \mathbb{R}^{C' \times 1}$ represents learnable parameters of the fully connected layer mapping $\mathbf{H_f}$ to the eye contact probability.

As discussed, the task of detecting pedestrian eye contact is formulated as a binary classification problem. Thus, the standard Binary Cross-Entropy loss function is used for the model training.

$$loss = -l \times log(p) - (1 - l) \times log(1 - p) \tag{8}$$

where $l$ represents the true binary label (0 or 1) and $p$ the probability of eye contact.

### D. Bi-Stream Fusion

The bone information has been proven to be important for skeleton-based action detections [22]. Inspired by existing methods [22], [27], we proposed a Bi-stream model to fuse the joints and bones information for eye contact detection. The Bi-stream model shares the same network structure and processes the bones and joints separately. We used the pose detector OpenPifPaf [35] to extract keypoints, which provides keypoint coordinates (x, y) along with corresponding confidence scores (denoted as s). The confidence score of a pose keypoint represents the level of confidence or certainty associated with the estimated position of that keypoint. It indicates the reliability of the estimated keypoint location and ranges from 0 to 1(the higher the value, the higher the confidence/reliability of the estimated keypoint location). For the joint stream, we use the coordinates and confidence scores of key joints as the input, denoted as $J = (x_i, y_i, s_i)$. To increase the model generalization, we normalize the key joints:

$$\begin{cases} x'_i = \frac{x_i - x_{left}}{w_{box}}, \\ y'_i = \frac{y_i - y_{left}}{h_{box}}, \\ s'_i = s_i, \end{cases} \tag{9}$$

where $(x_i, y_i)$ are original pixel joint coordinates and $(x'_i, y'_i)$ the normalized coordinates. $x_{left}$ and $y_{left}$ are the coordinates of the top left corner pedestrian box. $w_{box}$ and $h_{box}$ are the width and height of the pedestrian box, respectively. $s'_i$ and $s_i$ are the transformed and original confidence scores of joint $i$.

For the bone stream, each skeleton bone is bound with two joints and thus the bones information could be represented using joints. We define the source joint of a bone as the one close to joint-0 in the uNCG-NI skeleton graph, and the target joint as the one far away from joint-0. For each bone $B = (u_i, v_i, s_i)$, we calculate the bone attributes as:

$$\begin{cases} u_i = x_i.d - x_i.s, \\ v_i = y_i.d - y_i.s, \\ s_i = \frac{s_i.d + s_i.s}{2}, \end{cases} \tag{10}$$

where $u_i$, and $v_i$ are the normalized pixel coordinates of a bone. $x_i.s$, $x_i.d$, $y_i.s$, and $y_i.d$ are normalized coordinates of source and target joints, respectively. $s_i$ is the confidence score of bone $i$. $s_i.d$ and $s_i.s$ are the confidence scores of source and target joints, respectively.

Note that the skeleton graph data has no cycles thus the bone-0 could not be assigned with source and target joints. We set the bone-0 as an empty bone with the corresponding bone attributes as (0, 0, 0).

The SA-BiGCN model consists of two streams, including joint and bone streams (Fig. 2). The SA-GCN model takes inputs of joint and bone coordinates, and their corresponding confidence scores and outputs the probabilities of the joint stream $p_j$ and the bone stream $p_b$. The final probability of eye contact $p_f$ is calculated as:

$$p_f = \beta p_j + (1 - \beta) p_b \tag{11}$$

where $\beta$ represents the weighting factor. By adjusting the value of $\beta$, we can control the influence of each stream in the final probability of eye contact.

The final detection result $d$ for each image crop is true if $p_f$ is greater than a preset threshold value $d_{\text{threshold}}$; false otherwise. Algorithm 1 shows the main process and algorithms of the SA-BiGCN model.

**Algorithm 1** SA-BiGCN Model

**Input:** The pedestrian image crop set $\mathbb{I}$, labeled image set $\mathbb{L}$ (eye contact or not), graph adjacent matrix **A**, learnable parameters $\mathbf{W_e^{L0}}, \ldots, \mathbf{W_e^{LN}}$, $\mathbf{W_c^{L0}}, \ldots, \mathbf{W_c^{LN}}$, $\mathbf{W_d}$, and $\mathbf{W_p}$, the fusion weighting factor $\beta$, the detection threshold $d_{\text{threshold}}$
**Output:** Eye contact or not $d$ of pedestrian image crop $i \subseteq \mathbb{I}$

```
Define the Class FeatureExtraction&ContactProbability $p$=SA-GCN-Model(X) using the process in Fig. 4
Extracting keypoints using OpenPifPaf [35] (keypoints coordinates (x, y) and confidence score (s))
Generate inputs data for Joint and Bone streams
    • Joints input data $\mathbf{J}$ using Eq. (9)
    • Bone input data $\mathbf{B}$ using Eq. (10)
Train the SA-GCN-Model for joint and bone streams by minimizing Eq. (8) using the labeled set $\mathbb{L}$
Calculate eye contact probability
    • $p_j \leftarrow$ SA-GCN-Model($\mathbf{J}$)
    • $p_b \leftarrow$ SA-GCN-Model($\mathbf{B}$)
Calculate fused eye contact probability $p_f \leftarrow \beta p_j$+(1)-$\beta) p_b$
Detect eye contact, $d \leftarrow$ True if $p_f \geq d_{\text{threshold}}$; $d \leftarrow$ False, otherwise
return eye contact detection $d$
```

## IV. Experiments and Results

### A. Data and Benchmark Models

We compared our method with state-of-the-art models on three standard benchmark datasets: JAAD [14], PIE [15] and LOOK [7].

*1) JAAD:* The JAAD data set contains 346 video clips, and each clip lasts 5-10 seconds. There are 390,000 instances of pedestrians among which 17,000 instances have eye contact with drivers. The JAAD dataset consists of 686 unique pedestrians and has been widely utilized in various autonomous driving research studies.

*2) PIE:* The PIE is a real-world data set for pedestrian intention detection. It contains 6 hours of continuous pedestrian footage downtown and 700,000 annotated pedestrian instances, in which it contain 1,842 unique pedestrians and 180 pedestrians have eye contact with drivers.

*3) LOOK:* The LOOK data set is the first specialized data set for eye contact detection in the wild. It has 57K annotated pedestrian instances containing 7,944 unique pedestrians, which makes it the most diverse dataset for eye contact detection in the wild.

Regarding the experiment setting (e.g., training and testing split), we used the same setting as that in [7]. Specifically, the JAAD dataset was officially split as: 177 videos for training, 29 for validation, and 117 for testing. For the PIE dataset, we employed set01, set02, and set04 for training, set05 and set06 for validation, and set03 for testing. The LOOK dataset was officially split as approximately 45,738 samples for training and validation (90% for training and 10% for validation), and 10,378 samples for testing.

As is mentioned in [7], JAAD, PIE, and LOOK data are unbalanced toward a majority of people that do not have eye contacts with drivers. For fair comparisons, we used the same balanced test set generated using the procedure in [7] and [13]. The key steps of the procedure include:

- Obtaining indices of positive and negative samples.
- Extracting positive samples and counting the number of positive samples.
- Shuffling negative sample indices and selecting the same number of negative sample indices as positive samples.
- Extracting negative samples using negative sample indices, and concatenating both positive and negative samples to create a balanced dataset.

This process ensures that we have an equal representation of positive and negative samples, facilitating a more balanced and fair analysis in our approach.

We compared with seven benchmark models, including:

- **Rasouli** [14]. It uses the AlexNet architecture [34] as the backbone layer and is followed by fully connected layers. It takes the head crops of pedestrians as inputs.
- **MTL-Fields** [13]. It uses a Multi-Task Learning (MTL) model as the backbone. It uses the full image as the input.
- **ResNeXt-E** [39]. It uses ResNeXt-50 as the backbone layer and inputs the pedestrian eye crops.
- **ResNeXt-H** [39]. It uses ResNeXt-50 as the backbone layer and inputs the pedestrian head crops.
- **ResNeXt_FC-KH** [7]. It uses ResNeXt-50 as the backbone layer to extract features from the head crops,

TABLE I
MODEL SETTINGS

| | | |
|---|---|---|
| Data | Data split type | Default |
| | Data set | LOOK, JAAD, PIE |
| Model | Backbone | GCN |
| | Layers of backbone($L_n$) | 9 |
| | Number hidden units | 64,64,64,128,128,128,256,256,256 |
| | Normalization layer | Layer normalization |
| | Fusion layer | Attention |
| | Bi-stream Fusion weights | 0.6(joints),0.4(bones) |
| | $d_{\text{threshold}}$ | 0.5 |
| Training | Batch size | 32 |
| | Epochs | 300,200 |
| | Learning rate | 0.001,0.05 |
| | Optimizer | Adam |
| | Loss function | BCE Loss |
| Testing | Batch size | 32 |

and employs a fully connected block to extract features from the keypoints. In the later layers, it merges the features from both streams and generates predictions based on the fused features.

- **ResNeXt_FC-KE** [7]. It uses ResNeXt-50 as the backbone layer for feature extraction from the eye crops, and incorporates a fully connected block to extract features from the keypoints. In the early layers, it combines the features from both streams and produces predictions based on the fused features.
- **Belkada** [7]. It uses a simple fully-connected network with residual blocks ResNet-18 [40] as the backbone layer. It takes the key points of pedestrians as input.

We used average precision (AP) as the performance metric. It is calculated by the area under the Precision-Recall curve [41]. It provides a comprehensive performance evaluation of an algorithm by considering both the precision and recall scores.

### B. Model Settings

We implemented the SA-BiGCN model using PyTorch. Table I shows the model parameter settings. We set the Binary Cross Entropy as the loss function and the Adam optimizer for the model training with a weight decay of 5e-4. The SA-BiGCN model was pre-trained without the attention layer for 300 epochs (learning rate 0.001), and further trained with an attention layer for 200 epochs by fixing the backbone layer parameters (learning rate 0.05). The skeleton joint and bone streams are trained separately with a batch size of 32. We set the weights as 0.6 (joints) and 0.4 (bones) for the final Bi-stream fusion using a grid search of different weight values (See Appendix A). The same model configurations are used for the three datasets tested in the study.

### C. Results

Table II shows the model comparison results. Generally, the proposed SA-BiGCN model achieves the best performance on tested datasets by fully utilizing the skeleton structure information and dynamically fusing the joints and bone information under different detection contexts. In comparison to the pure image-based model, the methods that combine joints and image crops achieve better performance. However, the joint-based approach exhibits even better performance compared to the combined methods. It could be because of the low quality of images (even blurred) captured by the cameras for distant pedestrians in the wild. Compared with the best benchmark models, the SA-BiGCN model improves by 1.6% on JAAD, 2.6% on LOOK, and 0.1% on PIE when trained and tested on the same dataset.

As expected, the model performance decreases when training and testing on different datasets. For example, the average AP decrease is about 3.9% for SA-BiGCN (See the JAAD column, the AP decreases about 1.1% and 0.9% for the SA-BiGCN model when trained in LOOK and PIE dataset.), 4.1 % for Belkada, 4.3% for MTL-Field and 5.1% for Rasouli. It indicates the SA-BiGCN model has a relatively better generalization ability than its peer models. Also, the SA-BiGCN model consistently achieves the best performance when training and evaluating on different datasets.

Fig. 6. shows some examples of failed eye detection by the SA-BiGCN model. The reason could be due to the illumination (Fig. 6(a) and 6(b)). It is hard to recognize the posture of a pedestrian due to a lack of lighting or backlight. In Fig. 6(c) and 6(d), the sunglasses could prevent the model from obtaining the detail of the eyes, resulting in failure detection. Also, the remote distance would also lead to a failed detection given the pedestrian being too small to recognize (Fig. 6(e) and 6(f)). The failure detection in Fig. 6(g) and 6(h) is the occlusion of obstacles, in which the model could not get the full features of pedestrians, particularly the head.

### D. Ablation Study

To verify and explore the importance of each component in SA-BiGCN, we conducted the ablation analysis using the LOOK, PIE, and JAAD dataset. For discussion convenience, we presented results only for the LOOK data (see Appendix B for results of other datasets). The LOOK data set is specialized for eye contact detection, which is the most diverse dataset in pedestrians and environments. We test all model variants on the original LOOK data set without the balance operations.

*1) Impact of the Skeleton Graph:* We tested five different skeleton graphs and two different adjacency matrix construction strategies (used in ST-GCN) for the SA-BiGCN model. Table III presents a comparison of model performances using different skeleton graphs and adjacency matrices. It is observed that the model with NCG(UL) performs much better than the model with NCG(SCP). It indicates that the uni-labeling adjacency matrix construction strategy is favored over the spatial configuration partitioning strategy. The same results are observed for other skeleton graphs in the studied eye detection task in the wild, though the SCP was reported to be effective in action recognition given its ability to capture concentric and eccentric motion patterns. The reason could be that eye contact detection is more dependent on the imperceptible changes of eyes and head rather than the body movement patterns.

The SA-BiGCN models have poor performance with NCG(UL) and ACG(UL) skeleton graphs. Eye contact is a binary classification problem based more on head joints [7].

TABLE II
MODEL PERFORMANCE COMPARISONS

| Training Dataset | Method | Input | Eye Contact Classification (AP) ↑ | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | JAAD | LOOK-KITTI | LOOK-JRDB | LOOK-nuScenes | LOOK | PIE |
| JAAD | Rasouli [14] | Head crops | 75.4 | 65.9 | 87.2 | 78.7 | 77.3 | - |
| | MTL-Fields [13] | Images | 82.6 | 89.7 | 82.1 | 92.0 | 87.9 | - |
| | ResNeXt-E* [39] | Eye crops | 76.8 | 76.2 | 84.1 | 85.2 | 81.7 | 75.1 |
| | ResNeXt-H* [39] | Head crops | 79.1 | 72.7 | 92.2 | 87.0 | 88.7 | 79.2 |
| | ResNeXt_FC-KH* [7] | Keypoints+head crops | 80.3 | 76.3 | 92.6 | 88.1 | 89.6 | 80.0 |
| | ResNeXt_FC-KE* [7] | Keypoints+eye crops | 84.5 | 88.3 | 94.1 | **92.5** | 93.2 | 83.0 |
| | Belkada [7] | Keypoints | 85.9 | 91.6 | 94.8 | 91.0 | 92.5 | 84.5 |
| | Ours | Keypoints | **87.2** | **93.1** | **96.7** | 82.5 | **95.6** | **86.8** |
| LOOK | Rasouli [14] | Head crops | 71.0 | 76.8 | 89.5 | 82.9 | 83.1 | - |
| | MTL-Fields [13] | Images | 80.7 | 95.1 | 95.2 | 93.4 | 94.6 | - |
| | ResNeXt-E* [39] | Eye crops | 74.6 | 80.5 | 93.7 | 88.1 | 91.5 | 77.4 |
| | ResNeXt-H* [39] | Head crops | 71.9 | 85.3 | 95.1 | 90.4 | 93.9 | 78.1 |
| | ResNeXt_FC-KH* [7] | Keypoints+head crops | 76.5 | 88.4 | 95.9 | 92.6 | 94.9 | 81.8 |
| | ResNeXt_FC-KE* [7] | Keypoints+eye crops | 81.0 | 92.5 | 96.3 | 95.9 | 95.8 | 84.1 |
| | Belkada [7] | Keypoints | 86.0 | 96.4 | 97.1 | 95.1 | 96.2 | 86.4 |
| | ours | Keypoints | **86.1** | **98.0** | **99.0** | **96.3** | **98.8** | **86.6** |
| PIE | ResNeXt-E* [39] | Eye crops | 73.2 | 69.7 | 91.9 | 77.5 | 87.6 | 78.7 |
| | ResNeXt-H* [39] | Head crops | 72.8 | 74.3 | 89.0 | 79.3 | 86.5 | 80.2 |
| | ResNeXt_FC-KH* [7] | Keypoints+head crops | 77.7 | 75.1 | 91.9 | 84.3 | 89.4 | 82.8 |
| | ResNeXt_FC-KE* [7] | Keypoints+eye crops | 80.4 | 86.9 | 95.3 | 91.8 | 94.3 | 85.1 |
| | Belkada* [7] | Keypoints | 84.3 | 93.6 | 96.2 | 92.8 | 95.5 | 88.7 |
| | ours | Keypoints | **86.3** | **94.5** | **96.9** | **94.3** | **96.5** | **88.8** |

[1] The contrasted results are mainly cited from [7]
[2] Note that the '*' means the contrasted results are tested by ourselves according to the open source code (https://github.com/vita-epfl/looking).

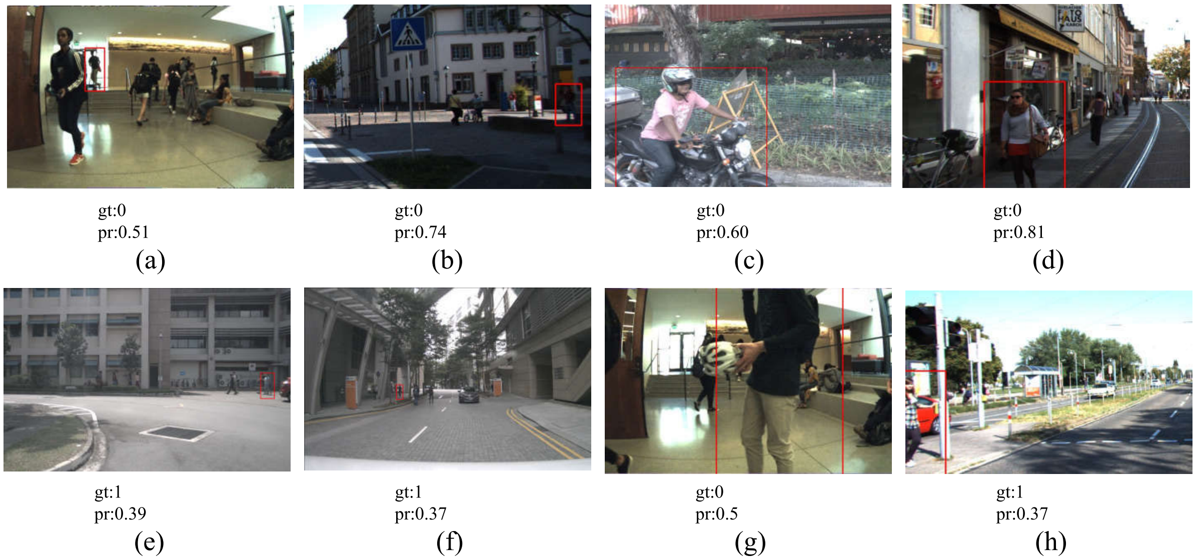

gt:0
pr:0.51
(a)
gt:0
pr:0.74
(b)
gt:0
pr:0.60
(c)
gt:0
pr:0.81
(d)
gt:1
pr:0.39
(e)
gt:1
pr:0.37
(f)
gt:0
pr:0.5
(g)
gt:1
pr:0.37
(h)

Fig. 6. Typical examples of detection failures. '1' represents eye contact, while '0' is without eye contact. '0.5' is the classification threshold. 'gt' is the ground truth classification value, and 'pr' is the predicted value. The illumination, shading of sunglasses, remote distance, and occlusion of obstacles are main reasons for failure detections.

The internal connection between body joints (e.g., both hands) could add marginally useful information for eye detection. The uNCG(UL) performs better than the NCG(UL) since the arms and legs contribute little useful information for eye contact. Compared with the uNCG(UL), the uNCG-N(UL) has a significant improvement. It verifies the hypothesis that the directed information transfer could decrease the over-smoothing issue in model training. As expected, the uNCG-NI(UL) performs best for the direction and hierarchical fusion. It improves the detection performance of the NCG(UL) model by 1.7% in final results, which increase by 1.7% in joints and 1.8% in bones, respectively.

*2) Impact of the Spatial Attention Layer:* Table IV shows the model comparison results with and without attention. We used four different methods to obtain the final features from different joints and bones. The 'Single joint or bone' means that we used the joint-0 or bone-0 as the final feature for the classification. The 'Sum,' 'Average,' and 'Attention' means that we used different methods to fuse the features from joints and bones as the final feature for the classification. The

TABLE III
MODEL PERFORMANCE WITH DIFFERENT SKELETON GRAPHS

| Graph and strategy | Eye Contact Classification (AP) ↑ | | |
|---|---|---|---|
| | LOOK data set | | |
| | joint | bone | two streams |
| NCG(SCP) | 75.2 | 86.7 | 84.3 |
| NCG(UL) | 94.5 | 94.4 | 94.9 |
| ACG(UL) | 93.9 | 92.1 | 94.0 |
| uNCG(UL) | 94.9 | 95.7 | 95.6 |
| uNCG-N(UL) | 95.6 | 95.2 | 96.0 |
| uNCG-NI(UL) | **96.2** | **96.2** | **96.6** |

[1] The SCP represents the spatial configuration partitioning strategy, and the UL represents the uni-labeling strategy [16].

TABLE IV
MODEL PERFORMANCE WITH DIFFERENT FEATURE FUSION MODELS

| methods | Eye Contact Classification (AP) ↑ | | |
|---|---|---|---|
| | LOOK data set | | |
| | joint | bone | two streams |
| Single joint or bone | 96.1 | 96.0 | 96.1 |
| Fusing joints or bones by sum | 96.1 | 96.1 | 96.1 |
| Fusing joints or bones by average | 96.1 | 96.1 | 96.1 |
| Fusing joints or bones by attention | **96.2** | **96.2** | **96.6** |

experimental results show that the Sum and Average method has the same performance as the single information model. The model with an attention layer has an improvement of about 0.5% in detection compared with the model without an attention layer.

*3) Impact of the Noisy Skeleton Data:* In real-life scenarios, it is inevitable to encounter incomplete and erroneous skeleton data. To further verify the advantage of SA-BiGCN, we designed four distinct types of noisy skeleton data, as described in the works of [42] and [43] (illustrated in Fig. 7). The noisy skeleton data comprises four distinct types of noise, namely:

- **Part occlusions**: This type of noise involves occlusions where either the leg or arm is partially obstructed (the joints of the leg or arm are set to zero).
- **Block occlusions**: The body is occluded, and only the keypoints of the head are obtained.
- **Random occlusions**: Some keypoints are randomly occluded with varying probabilities, such as 10% or 20%.
- **Random deviation**: Each keypoint experiences a deviation with different levels, such as 1% or 2%. This variation introduces slight inaccuracies in the joint positions.

We added the new comparison results of the two models in Table V for the constructed noisy test dataset generated using the LOOK dataset. The results show that the performance of Belkada model significantly deteriorates for the noisy dataset, especially when faced with block occlusions and random occlusions introducing incompleteness and noise to the skeleton data. In comparison, our proposed SA-BiGCN model achieves a reasonable detection performance under noises (over 70% accuracy) and performs better than Belkada across all scenarios. The observed discrepancy in performance can be attributed to the fully-connected network's strong dependence on specific data characteristics, such as the symmetry of the x-axis. In contrast, our proposed model excels in capturing the spatial structure inherent in the skeleton data, making it less sensitive to partial, incomplete, or deviated data. The SA-BiGCN model does not rely on single keypoints but instead the wholistic skeleton structure and it also does not need keypoints derived from arms or legs which can also contribute to its enhanced robustness and effectiveness in handling incomplete and noisy skeleton data.

TABLE V
THE COMPARISON RESULTS ON DIFFERENT TYPES OF NOISY DATA

| METHOD AND DATA SET / NOISY TYPE | | | LOOK dataset | |
|---|---|---|---|---|
| | | | BELKADA [7] | OURS |
| Original | | | 92.5 | 96.6 |
| Occlusion | Part occlusions | arm | 76.1 | 96.6 |
| | | leg | 80.9 | 96.6 |
| | Block occlusions | body | 53.8 | 70.8 |
| | Random occlusions | 0.1 | 59.4 | 78.4 |
| | | 0.2 | 49.7 | 72.6 |
| Deviation | Deviation percentage | 0.01 | 88.2 | 96.3 |
| | | 0.02 | 84.9 | 95.5 |

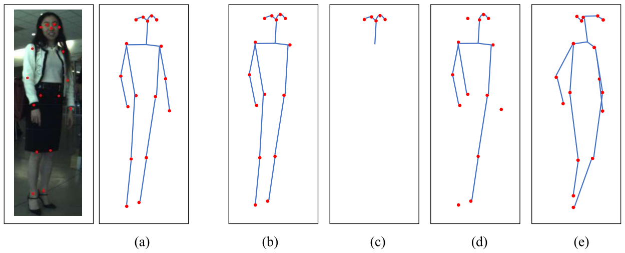


Fig. 7. Examples of the noisy skeleton data. (a) the normal skeleton data.(b) part occlusion. (c) block occlusion. (d) random occlusion and (e) random deviation.

## V. CONCLUSION

The eye contact transmits information and intention in the wild environment with mixed vehicles and pedestrians. This paper propose a GCN-based model for eye contact detection, namely, a Bi-stream spatial attention graph convolution network. It dynamically extracts and fuses skeleton joints and bones information under different detection contexts. We also design a directed, nose-centric skeleton graph for the eye contact detection which avoids redundant skeleton information and enriches the hierarchical information aggregation and exchange between important joints.

We validate the model performance by comparing benchmark models on three public datasets, including JAAD, PIE, and LOOK. The results highlight the accuracy and generalization performance of the proposed SA-BiGCN model in detecting the eye contact in the wild environment. The SA-BiGCN model improves the best benchmark model performance by 1.6% on JAAD data, 0.1% on PIE, and 2.6% on LOOK. The detection failures are mainly caused by the image quality, such as illumination, shading of sunglasses, remote distances, and occlusion of obstacles. The ablation analysis provides evidence supporting the significance of the skeleton graph design and the spatial attention mechanism in the feature fusion process. Notably, the proposed uNCG-NI skeleton graph demonstrates a notable improvement in model performance, surpassing the natural skeleton graph by 1.7% on the LOOK dataset. The experiment on noisy data further verifies the robustness of SA-BiGCN. Future work can further improve the eye contact detection performance by adding consecutive frames of skeleton graphs.

## REFERENCES

[1] T. Stein, A. Senju, M. V. Peelen, and P. Sterzer, "Eye contact facilitates awareness of faces during interocular suppression," *Cognition*, vol. 119, no. 2, pp. 307–311, May 2011.

[2] H. Kiilavuori, V. Sariola, M. J. Peltola, and J. K. Hietanen, "Making eye contact with a robot: Psychophysiological responses to eye contact with a human and with a humanoid robot," *Biol. Psychol.*, vol. 158, Jan. 2021, Art. no. 107989.

[3] K. Kompatsiari, F. Ciardo, V. Tikhanoff, G. Metta, and A. Wykowska, "It's in the eyes: The engaging role of eye contact in HRI," *Int. J. Social Robot.*, vol. 13, no. 3, pp. 525–535, Jun. 2021.

[4] A. Vrij, S. Mann, S. Leal, and R. Fisher, "'Look into my eyes': Can an instruction to maintain eye contact facilitate lie detection?" *Psychol., Crime Law*, vol. 16, no. 4, pp. 327–348, May 2010.

[5] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and Ü. Özgüner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 2, pp. 221–230, Jun. 2022.

[6] R. Hata, D. Deguchi, T. Hirayama, Y. Kawanishi, and H. Murase, "Detection of distant eye-contact using spatio-temporal pedestrian skeletons," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 2730–2737.

[7] Y. Belkada, L. Bertoni, R. Caristan, T. Mordan, and A. Alahi, "Do pedestrians pay attention? Eye contact detection in the wild," 2021, *arXiv:2112.04212*.

[8] V. Onkhar, P. Bazilinskyy, J. C. J. Stapel, D. Dodou, D. Gavrila, and J. C. F. de Winter, "Towards the detection of driver-pedestrian eye contact," *Pervas. Mobile Comput.*, vol. 76, Sep. 2021, Art. no. 101455.

[9] Y. Mitsuzumi and A. Nakazawa, "Eye contact detection algorithms using deep learning and generative adversarial networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2018, pp. 3927–3931.

[10] Y. Omori, "Image augmentation for eye contact detection based on combination of pre-trained alex-net CNN and SVM," *J. Comput.*, vol. 15, no. 3, pp. 85–97, 2020.

[11] K. Arai, A. Yamashita, and H. Okumura, "Pedestrian safety with eye-contact between autonomous car and pedestrian," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 1–5, 2019.

[12] D. Zhang et al., "Onfocus detection: Identifying individual-camera eye contact from unconstrained images," *Sci. China Inf. Sci.*, vol. 65, no. 6, pp. 1–12, Jun. 2022.

[13] T. Mordan, M. Cord, P. Pérez, and A. Alahi, "Detecting 32 pedestrian attributes for autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11823–11835, Aug. 2022.

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

[15] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6261–6270.

[16] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.

[17] W. Kay et al., "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[18] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[19] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7904–7913.

[20] X. Zhang, C. Xu, and D. Tao, "Context aware graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14321–14330.

[21] Z. Fang, X. Zhang, T. Cao, Y. Zheng, and M. Sun, "A new adjacency matrix configuration in GCN-based models for skeleton-based action recognition," 2022, *arXiv:2206.14344*.

[22] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12018–12027.

[23] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3590–3598.

[24] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 140–149.

[25] X. Liu, Z. You, Y. He, S. Bi, and J. Wang, "Symmetry-driven hyper feature GCN for skeleton-based gait recognition," *Pattern Recognit.*, vol. 125, May 2022, Art. no. 108520.

[26] F. Zhou, X. Tu, Q. Wang, and G. Jiang, "Improved GCN framework for human motion recognition," *Sci. Program.*, vol. 2022, pp. 1–10, May 2022.

[27] W. Zhang, Z. Lin, J. Cheng, C. Ma, X. Deng, and H. Wang, "STA-GCN: Two-stream graph convolutional network with spatial–temporal attention for hand gesture recognition," *Vis. Comput.*, vol. 36, nos. 10–12, pp. 2433–2444, Oct. 2020.

[28] M. Shopon, A. S. M. H. Bari, and M. L. Gavrilova, "Residual connection-based graph convolutional neural networks for gait recognition," *Vis. Comput.*, vol. 37, nos. 9–11, pp. 2713–2724, Sep. 2021.

[29] J. Xie et al., "Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition," *Neurocomputing*, vol. 440, pp. 230–239, Jun. 2021.

[30] N. Heidari and A. Iosifidis, "Temporal attention-augmented graph convolutional network for efficient skeleton-based human action recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7907–7914.

[31] T. Ahmad, H. Mao, L. Lin, and G. Tang, "Action recognition using attention-joints graph convolutional neural networks," *IEEE Access*, vol. 8, pp. 305–313, 2020.

[32] H. Xing and D. Burschka, "Skeletal human action recognition using hybrid attention based graph convolutional network," 2022, *arXiv:2207.05493*.

[33] H. Qiu, B. Hou, B. Ren, and X. Zhang, "Spatio-temporal segments attention for skeleton-based action recognition," *Neurocomputing*, vol. 518, pp. 30–38, Jan. 2023.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[35] S. Kreiss, L. Bertoni, and A. Alahi, "PifPaf: Composite fields for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11969–11978.

[36] X. Xiong, W. Min, Q. Wang, and C. Zha, "Human skeleton feature optimizer and adaptive structure enhancement graph convolution network for action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 342–353, Jan. 2023.

[37] Chinese Academy of Cyberspace Studies yangshuzhen@cac. gov.cn, "Development of network information technology in the world," in *World Internet Development Report*. Berlin, Germany: Springer, 2021, pp. 25–48.

[38] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.

[39] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[41] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

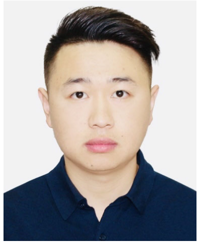
[42] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Richly activated graph convolutional network for robust skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1915–1925, May 2021.

[43] Y. Yoon, J. Yu, and M. Jeon, "Predictively encoded graph convolutional network for noise-robust skeleton-based action recognition," *Appl. Intell.*, vol. 52, pp. 2317–2331, Jun. 2022.

**Yancheng Ling** received the B.E. degree in traffic engineering and the Internet of Things engineering from East China Jiaotong University, Nanchang, China, in 2017, and the M.S. degree in traffic engineering from the South China University of Technology, Guangdong, China, in 2019, where he is currently pursuing the Ph.D. degree in traffic information engineering and control. He has been a Visiting Ph.D. Student with the Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, since 2022. Since 2017, his research interests mainly include mobile phone information collection and processing, intelligent transportation system (ITS), computer vision, video analysis, NLP, and traffic accident analysis.

**Zhenliang Ma** (Member, IEEE) received the B.Sc. degree in electrical engineering from Shandong University in 2009, the M.Sc. degree in information technology in 2012, and the Ph.D. degree in transportation engineering from The University of Queensland in 2015. He is currently an Assistant Professor in road traffic engineering and a Faculty Member of Digital Futures with the KTH Royal Institute of Technology. His research interests include statistics, machine learning, computer science-based modeling, simulation, and the optimization and control within the framework of selected mobility-related complex systems, which are intelligent transport systems (traffic/public transport/rails) and personal information systems (transport/energy).

**Bangquan Xie** (Member, IEEE) received the B.S. degree from the Changsha University of Science and Technology, China, in 2007, and the M.S. degree from the South China University of Technology, China, in 2013, where he is currently pursuing the Ph.D. degree. He has been a joint Ph.D. student with Clemson University International Center for Automotive Research (CU-ICAR), USA, since 2019. His research interests include sensor fusion, 3D detection, tracking and segmentation, multi-task learning, autoML, unsupervised and self-supervised, robot perception, autonomous driving, and automotive technology. He is the author of book *Detection and Maintenance of Electronic Control System for Automobile Body*. He was the youngest Senior Lecturer with the Guangdong Technician College of Transportation. He is an Automobile Inspection Assessor and the Senior Technician of Automotive Technology.



**Qi Zhang** received the B.E. and M.S. degrees in traffic information engineering and control from the Wuhan University of Technology. He is currently pursuing the Ph.D. degree with the Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm. His research interests include deep learning, urban public transport, individual mobility, and knowledge graph.

**Xiaoxiong Weng** received the bachelor's degree in industrial automation major from the Dalian University of Technology, China, the master's degree in automatic control theory and application major from Shanghai Jiao Tong University, China, and the Ph.D. degree in control theory and control engineering from the South China University of Technology (SCUT), China. She is currently a Professor with the School of Civil Engineering and Transportation, SCUT. She is the author of two books and multiple inventions. She has repeatedly undertaken the design and consulting of government management departments for the Guangzhou Asian Games, Shenzhen Universiade, Guangzhou University City, urban expressway system, subway, urban public transportation system, and other projects. Her research interests include intelligent transportation systems, computer vision, the dynamic modeling of urban traffic flow, data mining on transit systems, traffic signal control systems, transit commuter behavior analysis, and traffic accident analyzing.