Received: October 3, 2024. Revised: October 27, 2024.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Semantic Segmentation of Pedestrian Groups Based on Directional-oriented Density Features with Shallow U-net Architecture

Hanugra Aulia Sidharta[1] Berlian Al Kindhi[2] Eko Mulyanto[1,3] Mauridhi Hery Purnomo[1,3]*

[1]Deparment of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia
[2]Department of Electrical Automation Engineering,
Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia
[3]Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia
* Corresponding author's Email: hery@ee.its.ac.id

Abstract: Pedestrians typically form a small group with another pedestrian when they are traveling in the same direction and toward the same destination. While becoming members of a group of pedestrians, they dynamically interact with other pedestrians in the same group and outside the group. In order to understand pedestrian interaction, it is necessary to distinguish between pedestrians in a group and pedestrians not in a group. This can be achieved by performing semantic segmentation based on the JAAD dataset, which is suitable for observing pedestrian walking behavior with abundant annotation. In this research, we propose to perform semantic segmentation by utilizing directional-oriented density features. Density features are calculated by utilizing each joint relationship, while pedestrian direction can be predicted by calculated dot product based on shoulder, neck and hip joint. Segmentation is performed by employing a shallow U-network architecture, with fewer layers than the original U-network architecture. Compared with the original U-net architecture and its derived, our proposed method not only outperforms but also achieves stable performance from the early epoch in 6th epoch, and reaches a score over 0.97 during prediction, demonstrating its impressive performance.

Keywords: Semantic segmentation, Group of pedestrians, Density features, Directional oriented.

1. Introduction

One of the most critical events related to pedestrian safety is when pedestrians are crossing the road, as they are in direct contact with other road users. Pedestrians, cyclists and motorcyclists can be categorized as vulnerable road users (VRU) due to their lack of protection. This vulnerability makes pedestrians the most at risk of harm when involved in a traffic collision. To mitigate this problem, many automakers are incorporating advanced pedestrian safety features into their latest vehicles, such as automatic emergency braking, forward collision warning, and pedestrian detection systems. However, these safety features have some limitations, including their sensitivity to lightning conditions, their reliance on sensor readings, and their inability to handle complex situations. Therefore, the implementation of an Advanced Driver Assistance System (ADAS) is necessary to effectively respond to the complex and ever-changing nature of pedestrian behavior.

Understanding pedestrian behavior is challenging because each individual may have a unique pattern. Pedestrian behavior can be observed through their walking patterns, including speed, direction, and formation. Hu performs pedestrian re-identification of walking patterns based on the relationship of postural structures through hypergraph analysis [1]. Noh identifies and categorizes pedestrian risk clusters during traffic crossings based on individual walking patterns [2]. Neogi analyzed the interaction between pedestrians and car drivers by developing a contextual model based on each frame in the crossing situation [3]. analyzes crowd movement patterns by

pedestrian microsimulation and examines their egress walking behavior [4].

The difficulty of understanding pedestrian behavior increases when pedestrians walk together as a group. As social beings, pedestrians constantly interact with others in the same group. While walking within the same group, pedestrians typically maintain their walking speed to maintain group cohesion [5, 6]. Zaki's research also supports this claim, mentioning that certain leaders can influence group behavior and walking patterns [7]. When walking in a group, pedestrians form a distinct V pattern, which can be observed in their density data [8].

Detecting groups of pedestrians can be achieved by combining the pedestrian feature space and constructing a graph to measure their correlation [9]. A group of pedestrians can also be viewed through multi-camera data to analyze the pedestrian field topology and further investigate their joint motion and mechanisms [9]. Cheng uses a density-based technique by integrating density-based clustering from each pedestrian spatial area [10]. Thus, density data can be used as a feature for observing pedestrians and their groups.

The crowd of pedestrians can also be analyzed pixel-wise with pedestrian segmentation techniques. The group of pedestrians can also be achieved by examining the spatiotemporal relationship of pedestrian density [11]. Pedestrian segmentation can also be achieved by employing the Histogram of Oriented Gradients (HOG) and inputting it into the Mask R-CNN architecture [12]. Wang employs infrared imagery to generate heatmap pedestrian masks, which are utilized for pedestrian segmentation tasks [13].

In this experiment, we employ 2D pose estimation to extract pedestrian behavior. The 2D pose can be employed as a means of human action recognition (HAR) through observation of joint relationships [14, 15], the prediction of pedestrian crossing intentions [16, 17], and even can be performed in real-time scenario The advantages of 2D pose estimation include the ability to perform thorough observation of pedestrians through relational joint observation and joint movement, as well as the extraction of spatio-temporal features.

This experiment aims to leverage the advantage of 2D pose estimation by providing detailed pedestrian joint estimation. This joint data is then utilized to predict pedestrian direction. The density map feature is extracted as a group of pedestrian features, with each pedestrian direction considered when constructing this feature.

The contributions of this study can be summarised as follows:

- This research proposes that a direction-aware feature has been extracted as a group of pedestrians' features. This feature was obtained by performing a dot calculation from the shoulder and hip joints, with the neck joint serving as the reference node.
- A density map feature is being constructed by calculating the relationship between joints within and between pedestrians.

The paper is structured as follows: Section 2 explores the research foundation through equivalent research. Section 3 describes the suggested approach and dataset. Section 4 contains this study's results, analysis, comparison within the similar U-net approach, and discussion, while section 5 presents the conclusion and future work.

2. Related work

Implementing group pedestrian segmentation involves a multitude of challenges, from managing occlusions to dealing with fluctuating pedestrian densities. In the complex environments where pedestrians often gather closely together, segmenting overlapping pedestrians in dense crowds is a particularly demanding task. It necessitates advanced approach that can effectively capture both appearance and contextual relationships among individuals, a task made difficult by viewpoint distortions and variations in pedestrian scale. The limitations of current methodologies in high-density situations often lead to problems such as overlooked detections and erroneous positives in group segmentation tasks.

Observing groups of pedestrian interaction can be performed by detecting pedestrians and employing cluster-based analysis [18], tracking the movement of each pedestrian [19], recognition through each pedestrian walking similarities pattern [5]. ch pedestrian's behavioral patterns in the temporal domain is computationally intensive. To address this issue, the group of pedestrians can be evaluated based on their density and spatial proximity. This density characteristic, which can be observed during queues in traffic lights [20], road side [21], and detecting density based on their cellular data [22], is proposed as an innovative approach to understanding pedestrian groups. The features are retrieved by analyzing pedestrian direction using the results of 2D pose estimation.

Another approach for understanding pedestrian interaction is performing a group of pedestrian segmentation. Pedestrian segmentation can be executed by establishing adaptive regions of interest using stereo images [23], utilising circular shortest paths via infrared images [24], and implementing

intensity inhomogeneity [13]. Despite being designed for medical applications [25], the U-net architecture demonstrates notable efficiency even with constrained data availability [26]. As result, numerous segmentation tasks are executed through the development of U-net architecture. Wang is enhancing U-net by incorporating a residual block as the convolutional backbone [27, 28], enriching segmentation accuracy through building attention unit to focus on essential regions [29], and refining feature extraction through recurrent layer and residual connection. In this experiment, we propose a shallow U-net to aim at a group of pedestrian segmentation by distinguishing pedestrians to facilitate a comprehensive understanding of their behavior.

3. Proposed method

This research is composed of five sections, as illustrated in Fig. 1, to achieve group pedestrian segmentation. The process begins with a monocular image as the input data, which is then processed with 2D pose estimation to obtain the joint prediction for each pedestrian. The direction of each pedestrian is calculated, and the group of pedestrians' features are extracted using this data. Finally, the segmentation task is performed using the U-net architecture.

3.1 Monocular dataset

The study of pedestrian behavior is complex, as it is subject to many influences, including internal decision-making processes and external factors. The internal variables may be influenced by several factors, including the direction the pedestrian is moving, their intended destination, and their physical appearance. In contrast, several variables may influence exterior factors, including pedestrian relationships, traffic conditions, weather, and obstacles. A multitude of datasets are currently being developed to observe pedestrians. These include the KITTI vision dataset [30], the ETH Pedestrian Dataset [31], the TJU-DHD (Tianjin University-Diverse Human Dataset) [32], the JAAD (Joint Attention in Autonomous Driving) dataset [33], and the PIE (Pedestrian Intention Estimation) Dataset [34]. However, only the JAAD and PIE datasets are equipped with behavioral annotations. The JAAD dataset offers comprehensive behavioral annotations for pedestrian behavior and interactions, while the PIE dataset focuses on pedestrian behavioral annotations.

The JAAD dataset is recorded using a high-quality monocular dashcam positioned on the dashboard of a moving vehicle. The dataset provides authentic scenarios of pedestrians interacting with their environment by capturing real-time events during driving. JAAD meticulously adjusts the video length from 5 to 10 seconds to achieve a regular data distribution and mitigate data imbalance. The resulting dataset consists of 346 short videos annotated by experts to provide accurate and reliable information. The data may be used to study various aspects, including pedestrian intention while crossing,

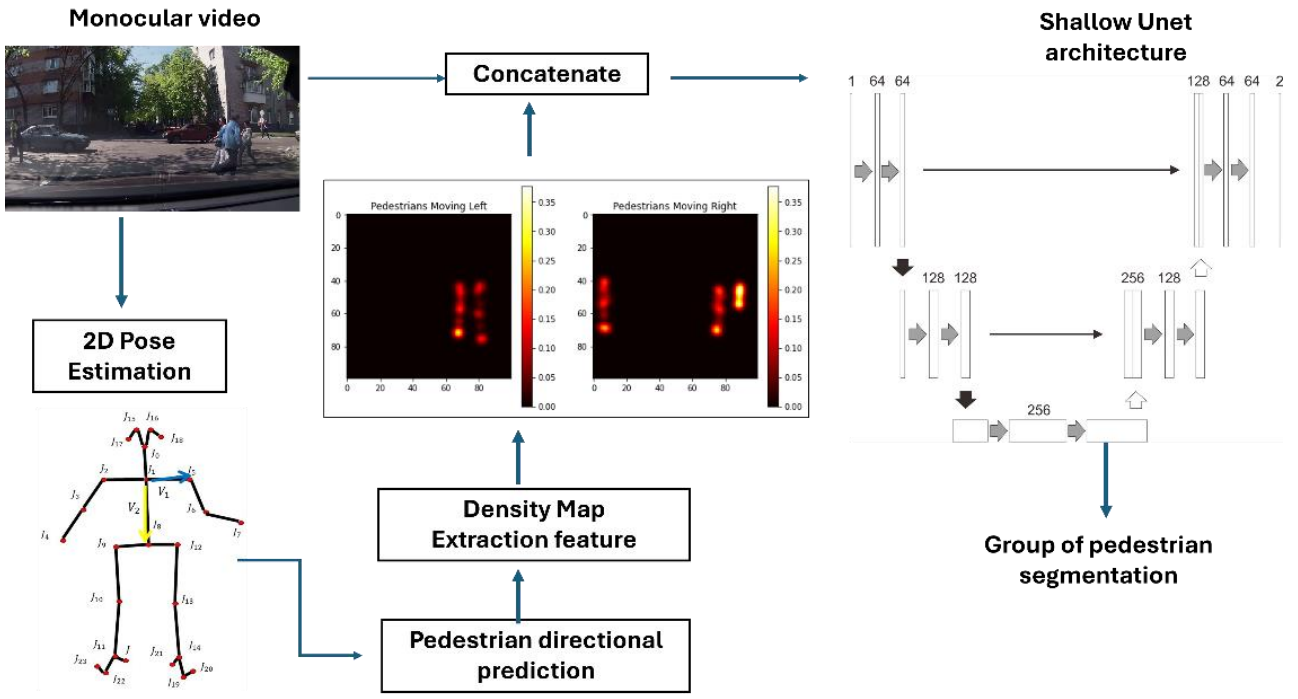


Figure 1. Proposed block diagram for semantic segmentation group of pedestrians

pedestrian interaction with drivers, and pedestrian behavior while walking on or crossing the road.

3.2 2D Pose estimation

Extracting pedestrian behavior from a monocular video is inherently challenging due to the absence of depth information in the data. Due to the dynamic nature of the environment, in which the distance between vehicles and pedestrians is subject to change, the images of pedestrians are susceptible to scaling issues. Furthermore, a multitude of pedestrian scenarios result in occlusion. It is, therefore, essential to accurately identify and extract the behavioral characteristics of each pedestrian. One of the most frequently utilized methodologies is deploying a pedestrian detector, such as YOLO [35] or multispectral pedestrian detection [36]. However, this solution presents a challenge regarding resource consumption, as it necessitates an additional step following the recognition of each pedestrian. This study uses two-dimensional pose estimation to accurately identify pedestrians' positions and anticipate each individual's joint movements.

A thorough examination of the JAAD dataset revealed that many scenes depicted vehicles approaching pedestrians. These scenarios are observed at zebra crossings, intersections, parking lots, and indoor areas. The number of pedestrians depicted in each dataset image ranges from one to thirteen. However, the dataset does not include images of large crowds. Consequently, we propose the utilization of Open Pose as a framework for 2D pose estimation, as Open Pose is particularly well-suited for scenarios involving the estimation of 2D poses of multiple pedestrians. Pose estimation is achieved using bottom-up methodologies that yield detailed joint information for pedestrians [37].

The open pose can extract 135 key points representing human joints. It provides an output format BODY25, which includes 25 human joints and extensive information about hand joints, foot joints, and facial expressions. This research used the BODY25 format, where each pedestrian is represented by 25 joint locations in 2D Cartesian coordinates. Each joint is accompanied by a confidence degree.

3.3 Pedestrian direction prediction

Per our observation, pedestrians tend to cross the road together. This behavior is frequently observed when they are waiting by the roadside. Pedestrians tend to form groups with other individuals headed to the same destination while awaiting the clearance of traffic to cross the road.

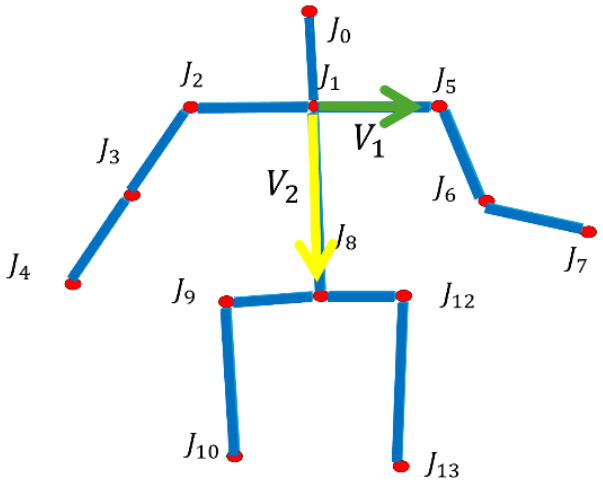


Figure. 2 Pedestrian direction prediction based on V_1 and V_2 dot product from 25 joint node

Upon the arrival of a traffic halt, the pedestrians cross the road in unison at a consistent walking speed. Pedestrians on the left side of the road cross in the direction of the right side, whereas those on the right side cross in the direction of the left. Consequently, the direction of each pedestrian is in opposition to their initial position.

The direction of pedestrians can be predicted by examining the dot product of two vectors, V_1 and V_2, which are obtained from the 2D pose estimation results of each pedestrian. The value of V_1 is determined by subtracting J_5 from J_1, while V_2 is calculated by subtracting J_8 from J_1. V_1 represents the vector observation from the neck to the shoulder, while V_2 represents the observation from the neck to the hips. The direction of a pedestrian can be determined by combining the dot product of two vectors. Fig. 2 illustrates how V_1 acts as a global reference vector can be utilized, as demonstrated in Eq. (1), which depicts the direction from the neck to the right shoulder. A positive dot product result indicates that the predicted direction of the pedestrian is aligned with the global "right" direction. Conversely, a negative value suggests that the predicted direction is opposite. Conversely, a negative value suggests that the predicted direction is opposite. The dot product calculation can be expressed using the following equation, where V_1 and V_2 are the input vectors and θ is the reference for direction.

$$V_1 \cdot V_2 = \|V_1\|\|V_2\| \cos\theta \quad (1)$$

3.4 Density map extraction feature

The proximity of pedestrians can be used to identify a group of pedestrians.

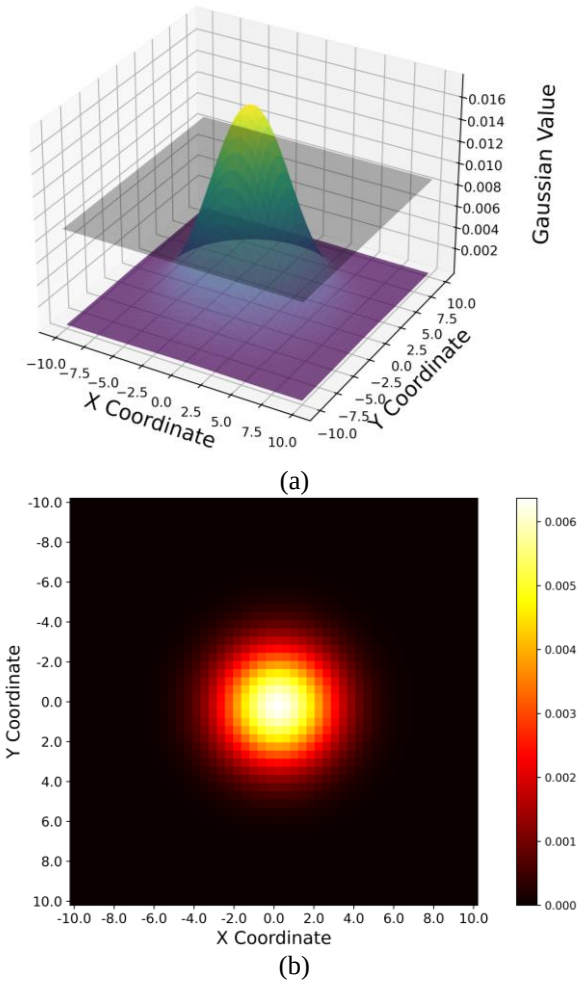


Figure. 3 The construction of density features based on gaussian filter with: (a) 3D plane visualization of gaussian filter and (b) 2D plane visualization gaussian filter based on its centroid

If two or more pedestrians are nearby, they can be considered part of the same group. However, this approach may result in ambiguity due to the possibility of different starting points. While near another pedestrian, they cannot be classified as part of the same group of pedestrians because they are moving in opposite directions. Thus, it is important to extract a group of pedestrians while considering their direction.

$$G(x,y) = \frac{1}{2\,\pi\sigma^2}\, e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2)$$

Algorithm 1 is developed to extract the pedestrian feature. This approach aims to derive a density map that includes a group of pedestrian characteristics. A density map aims to identify and examine the spatial patterns of pedestrian proximity within a specified area. The decomposition of $density_L$ and $density_R$ is accomplished using a calculation derived from Eq. (1). Pixel density characteristics can be obtained by analyzing the inter- and intra-joint relationships among pedestrians. The noise in the data can be reduced by applying a Gaussian filter to each dataset.

Each pedestrian joint is represented by a point defined by x and y coordinates in a 2D coordinate system. If the density feature is calculated using this dot data, the result will be a null of data. Therefore, the variance size is induced by modifying based on a Gaussian filter kernel. Filter kernel adjustment can be accomplished by adjusting the value of σ, as demonstrated in Eq. (2) and Algorithm. 1 in lines 11 and 12. The construction of the density map is illustrated in Fig. 3, comprising a 3D image in Fig. 3(a) and a 2D image in Fig. 3(b). This is crucial in modulating the adjacent pixel, leading to a more refined data output. The primary consequence of the increase σ is its effect on the area width. As σ grows, the value diverges further from the data center. A density map was created by intersecting a 3D image with a 2D plane, as illustrated in Fig. 3(a), and this 2D plane can be represented as a 2D density map feature, as shown in Fig. 3(b).

The advantages of utilizing 2D pose estimation is the data of pedestrians is provided detailed in form of joint data, by utilizing this data then there are no required performing pedestrians detector. Therefore, we can easily acquire each pedestrian bounding box, by performing calculations regarding the maximum and minimum of x and y coordinates. While group of pedestrians bounding box can be obtained by calculating proximity of each pedestrian bounding box in same frame. Ee can develop Algorithm 2 to compute the bounding box for individual pedestrians as well as groups of pedestrians.

3.5 Shallow U-net architecture

The extraction group of pedestrian features resulted in nine features, which were composed $density_L$, $density_R$, $pedBB_L$, $pedBB_R$, $groupBB_L$, $groupBB_R$, with additional 3 layer of original image. To ensure data integrity, all features are required to have the same data shape, with dimensions of $256\,x\,25$. This consistency is crucial, as deep learning architectures are designed to handle data with high integrity.

Algorithm 1 Building density map feature

Input: 2D pose coordinate
Output: $density_L, density_R$
Parameter: 25 joint coordinates

1: Set n_ped = number of pedestrian
2: Set $density_L$ dimension = [256, 256]
3: Set $density_R$ dimension = [256, 256]

4: **for** i = 0, in range i < n_ped **do**
5: $dir = \|V_1\|\|V_2\| \cos\theta$
6: **if** dir < 0 then
7: calculate $density_L$ based on each joint
8: **else**
9: calculate $density_R$ based on each joint
10: **end for**
11: Gaussian filter ($density_L$, σ)
12:Gaussian filter ($density_R$, σ)

Algorithm 2 Building bounding box feature

Input: 2D pose coordinate
Outpu $groupBB_L$, $groupBB_R$, $pedBB_L$, $pedBB_R$
Parameter: 25 joint coordinates
1: Set n_ped = number of pedestrian
2: Set $groupBB_L, groupBB_R$
3: Set $pedBB_L, pedBB_R$
4: **for** i = 0, in range i < n_ped **do**
6: **if** dir < 0 then
7: calculate $pedBB_L$ based on each joint
8: **else**
9: calculate $pedBB_L$ based on each joint
10: **end for**
11: **for** dt_BB, in $pedBB_L$ **do**
12: **if** dt_BB intersect then
13: $groupBB\_L$ append(dt_BB)
14: **for** dt_BB, in $pedBB_R$ **do**
15: **if** dt_BB intersect then
16: $groupBB_R$ append(dt_BB)

The U-Net architecture is employed due to its reputation for high-performance segmentation, even with limited data. This architecture is commonly used for performing segmentation in medical images with high complexity and various data patterns. It is derived from the autoencoder concept, with layers performing the functions of encoding, bottlenecks, and decoding. This architecture is modified with skip connections between each layer to improve model performance by preserving spatial information and improving gradient flow.

In this study, we utilize the shallow U-Net architecture to segment a group of pedestrians. Our approach involves employing three layers of the U-Net, with filter size configurations that commence with 64 in the initial layer, 128 in the second layer, and 256 in the third layer bottleneck layer. We ensure that the dimensions of the filters are harmonized across both the encoder and decoder layers. Compared with the original U-Net architecture, our shallow U-Net architecture has fewer layers.

4. Experimental result and discussion

This experiment aims to perform semantic segmentation on a group of pedestrians. The segmentation process employs joint density calculations based on the results of 2D pose estimation. This approach allows for observing the relationship between the joints of individual pedestrians and those of other pedestrians in the scene. To differentiate between pedestrians moving in different directions, this experiment employs a method of predicting the direction of pedestrian movement based on the coordination of their shoulder and hip joints with their neck as the reference point.

4.1 Group of pedestrian feature extraction

The group of pedestrian features is composed of 9 layers, which consist of $density_L$, $density_R$, $pedBB_L$, $pedBB_R$, $groupBB_L$, $groupBB_R$, with additional three layers of original image. The illustration of this feature can be observed in Fig. 4. The data presented therein was extracted from video_149. The video was selected for analysis due to its suitability for illustrating the formation of groups of pedestrians. In this video, vehicles move closer to the intersection, where pedestrians are waiting on both sides. The pedestrians on both sides await the vehicle's arrival and maintain visual contact with it. A group of pedestrians forms on the right side, comprising two individuals, while a single pedestrian is waiting on the left side. Upon the vehicle's halt, the pedestrians on both sides commenced traversing the roadway. As pedestrians cross the road, they interact with each other, which can result in forming a group. The employment of group features based solely on density data, without consideration of pedestrian direction, may result in ambiguity regarding the group's integrity. Therefore, the integrity of the group of pedestrians can be maintained by constructing a density data set based on their direction of movement.

Fig 4 portrays a scenario where a single pedestrian traverses the crossing from the left lane while two pedestrians proceed in a group from the right side. Fig. 4(a) illustrates the actual image data of the event, while Figs. 4(b) and 4(c) together provide a density features of both directions. By examining both density features, it becomes apparent that the head, hands, hips, and legs of each pedestrian contribute to the generation of relational data based on their proximity.

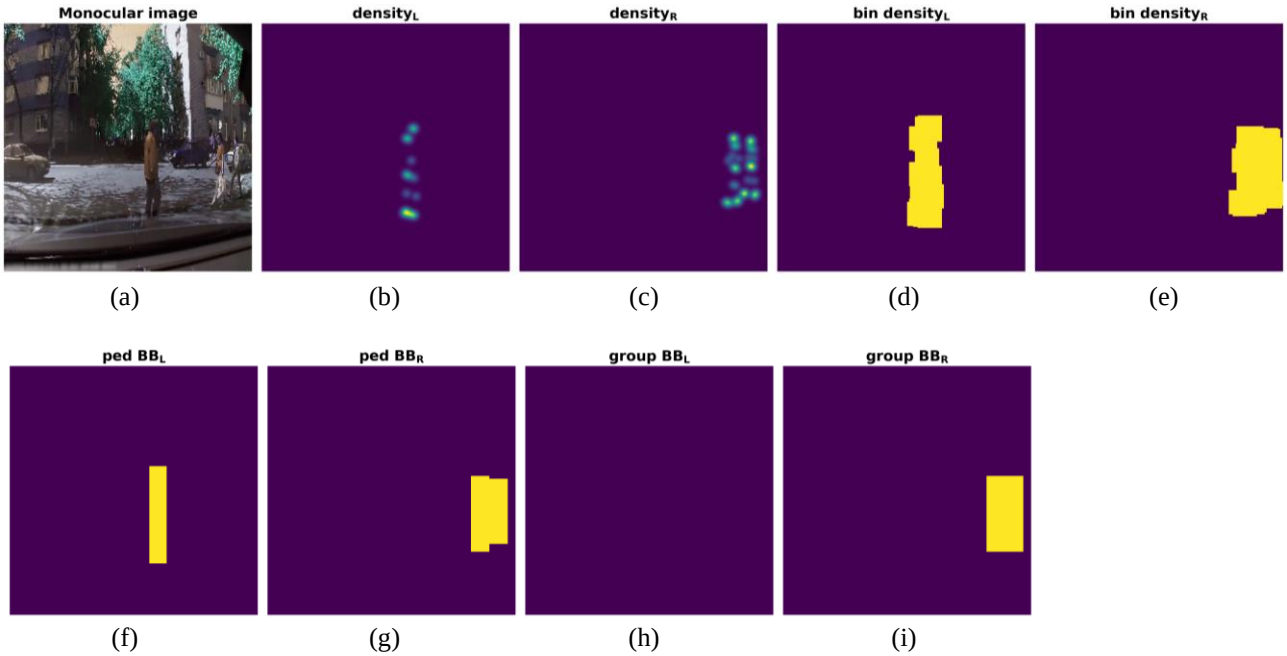


Figure. 4 Group of pedestrian feature extraction result: (a) Monocular image, (b) density$_L$, (c) density$_R$, (d)bin density$_L$, (e)bin density$_R$, (f)ped BB$_L$, (g)ped BB$_R$, (h)group BB$_L$, and (i)group BB$_R$

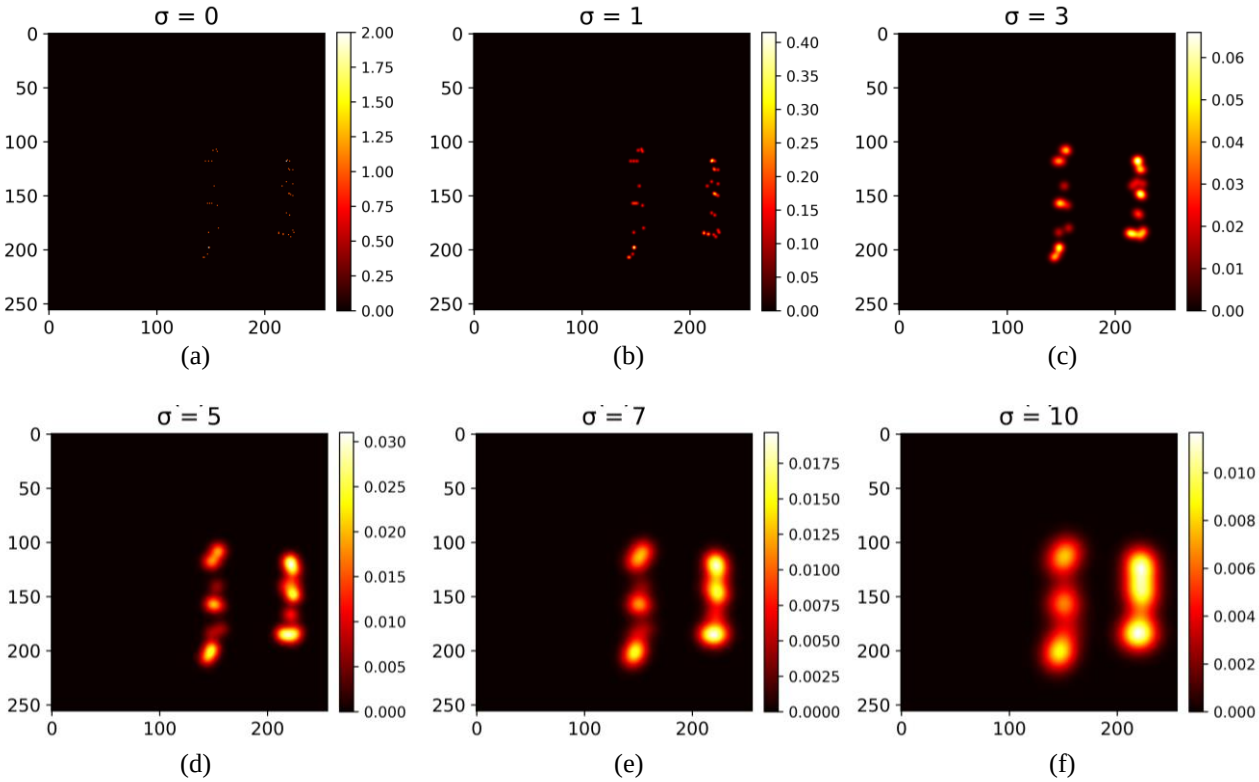
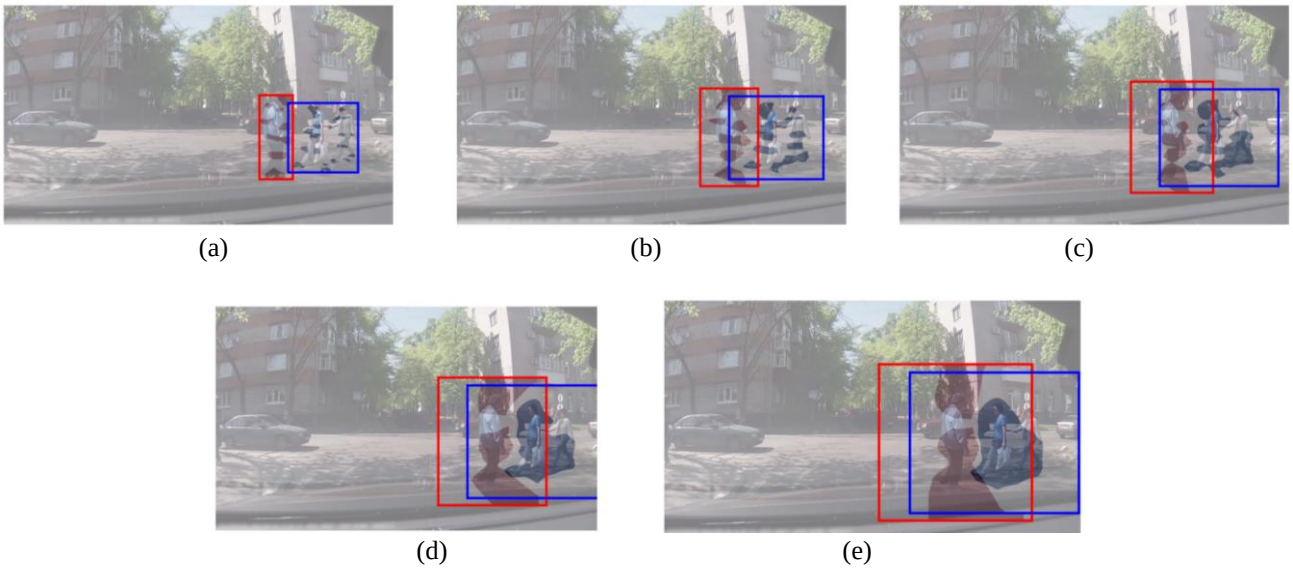


Figure. 5 Density map comparation based on different gaussian margin: (a)$\sigma = 0$, (b)$\sigma = 1$, (c)$\sigma = 3$, (d)$\sigma = 5$, (e)$\sigma = 7$, and (f)$\sigma = 10$

In the $density_R$ representation, the second pedestrian on the right displays a higher density than the first pedestrian in the same group. One disadvantage of employing density as a spatial feature is that pedestrians far from the camera may experience size integrity issues. To address this limitation, binarization of the density data is employed, as illustrated in Figs. 4(d) and 4(e). The $bin\ density_L$ and $bin\ density_R$ are calculated by normalizing the data based on both density features using binary data. The binarized data is employed as target data masking.



(a) (b) (c)

(d) (e)

Figure. 6 Overlay of pedestrian density features with a directional grouping and boundary box: (a)$\sigma = 1$, (b)$\sigma = 3$, (c)$\sigma = 5$, (d)$\sigma = 7$, and (e)$\sigma = 10$

Pedestrian density features can be obtained by utilizing the Algorithm. 1. However, the original Algorithm. 1 resulted in zero data due to each pedestrian's joint coordination being represented by a dot at the respective coordinate of each joint location. To address this challenge, we modified the Algorithm. The modifications were made to lines 11 and 12 of the Algorithm when applying the Gaussian filter, as these lines were found to be the key areas affecting the data clarity.

As shown in Fig. 5, the impact of varying margins on density features is significant. Our series of observations with varying σ value, including 0, 1, 3, 5, 7, and 10, clearly demonstrate the importance of choosing the right σ value. Fig. 5(f) reveals the ambiguity in the pedestrian data, emphasizing the need for careful consideration of σ value to ensure data clarity. Based on visual inspection of Figs. 5(a) and 5(b), insufficient data are found, which may lead to ambiguous results. The optimal feature is produced with margins 3, 5, and 7, as seen in Figs. 5 (c) to 5(e).

A group of pedestrian bounding boxes is built by utilizing Algorithm.2. A group of pedestrian bounding boxes can be obtained by detecting each individual pedestrian bounding box. Each pedestrian bounding box is detected by considering the area of each 2D joint.

The results of the pedestrian bounding boxes are vividly depicted in Figs. 3(f) and 3(g), providing a clear visual representation. As shown in Fig. 3(g), the height difference between the two pedestrians is clearly discernible. The differences, caused by their respective positions, indicate varying data depths. The bounding box of the group of pedestrians is clearly visible in Figs. 3(h) and 3(i). However, there is no data in Fig. 3(h) due to the presence of pedestrians forming the related pedestrian group. In contrast, Fig. 3(i) clearly illustrates the occupancy of the group of pedestrians.

The outcome of Algorithm. 1 and Algorithm. 2 can be presented in Fig. 6 by overlaying density features and the bounding boxes of pedestrian groups. Fig. 6 was constructed by applying various σ values following Eq. (2). Fig. 6 illustrates the impact of shifting on σ value on density feature and bounding box dimensions. An increase in σ value will result in more extensive data and lead to ambiguity. This image further confirms the findings of directed pedestrian group detection. A group of pedestrians is divided into two directions, designated by distinct colors: a red bounding box for those traveling to the right and a blue bounding box for those moving to the left. Despite the increase in feature size corresponding with σ shifting, the direction-oriented group for pedestrian bounding box identification demonstrates excellent performance.

4.2 Imbalance data issue

Based on the observations in the JAAD dataset, we found that the event of pedestrians forming a group of pedestrians is limited. JAAD dataset provides 346 videos with a length of 5 to 10 seconds. Thus, all the data has a total of 82,032 and involving 2,786 pedestrians. Based on observations, this number is narrowed to 15,220 frames for data with pedestrians and decreases significantly to 3,535 frames with a group of pedestrians context. Based on calculations of Algorithm 2, 941 frames were detected with pedestrians with direction from left and

pedestrians with direction from right. Thus, the enormous difference between the number of frames with a group of pedestrians and those without a group of pedestrians could result in an imbalanced data issue. To mitigate this issue and maintain normal data distribution, we perform balancing data processing. Data processing can be fulfilled by carefully inspecting data distribution on every video and performing frame filtering.

Video_149 is utilized to visualize the data balancing procedure. This video contains a total of 299 frames, and by employing Algorithm. 2, the result is that 198 frames do not contain a group of pedestrians, while 101 frames contain a group of pedestrians. To balance the data then, we use the number frame containing a group of pedestrians as a reference and choose data with a group of pedestrians with an additional 10% data margin. Thus, the nongroup of pedestrians is shrunk into 111 frames. By employing this approach then, the total frame of video_149 becomes 212 frames. We can obtain 7.423 from the previous version with only 3,535 frames by utilizing this approach.

4.3 Segmentation model performance

In order to prevent overfitting due to the data sequence, this experiment separated data into training, validation, and testing data. Additionally, data shuffling is employed to ensure the integrity of the experimental process. The data set is divided into three subsets: 70% of the data is allocated to the training set, 15% is designated for the validation set, and the remaining 15% is reserved for the testing set. The training and validation data are employed during the training sessions, whereas the testing data are utilized for the prediction sessions.

The model performance will be evaluated using three metrics: loss, Intersection of Union (IoU), and Dice coefficient. The loss metric assesses the performance of the trained model by comparing the discrepancy between the predicted and ground truth values. While the IoU assesses the overlap between the prediction and the ground truth, the Dice coefficient serves a similar purpose, with an additional calculation that weighs positive true instances, thereby rendering the Dice coefficient more sensitive than the IoU.

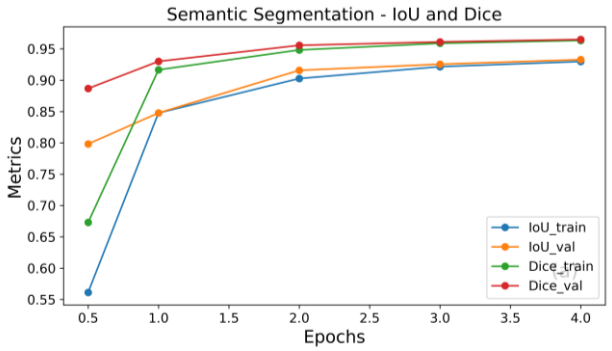
In this experiment, the model's performance will be evaluated twice by calculating each σ value and subsequently analyzing the results in more detail based on each epoch. Additional observations will be conducted by comparing the performance of our proposed shallow U-Net architecture with that of the original U-Net architecture. The data of each performance metric is presented in Table 1. The data in Table 1 is an average of all the data points from all the epochs and is equipped with the standard deviation. The simulation was conducted with identical epochs of 8 and a batch size of 16.

Based on data from Table. 1, Shallow U-net performs better than U-net. However, both models share similar characteristics due to utilizing the same backbone concept. Performance on σ=1 is adversely affected by the insufficient density of the data, which results in feature ambiguity. While performance improves from σ=3 onwards, there is a slight decline from σ=10. This evidence supports our initial hypothesis, as illustrated in Fig. 5. The larger size results in data ambiguity due to the border between each joint becoming vaque and unclear.

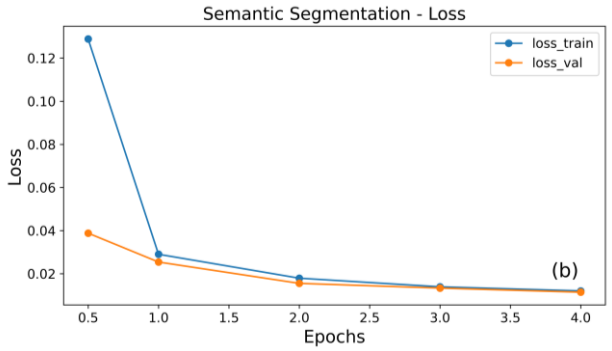
The most notable performance stability is achieved with σ=3, σ=5, and σ=7.

Table 1. Performance comparison of shallow U-net and U-net with various gausssian margin (metric shown in average value ± standard deviation)

		σ variance				
Model	**Metric**	**1**	**3**	**5**	**7**	**10**
S U-net	**Loss**	0.0166 ± 0.0229	0.0265 ± 0.0137	0.0195 ± 0.0237	0.0381 ± 0.0415	0.0643 ± 0.0624
	Val Loss	0.0091 ± 0.0052	0.0223 ± 0.0057	0.0122 ± 0.0058	0.0254 ± 0.0126	0.0448 ± 0.0352
	IoU	0.7939 ± 0.1667	0.8986 ± 0.0479	0.9414 ± 0.0557	0.9174 ± 0.0744	0.9093 ± 0.0785
	Val Iou	0.8377 ± 0.0739	0.9133 ± 0.0164	0.9579 ± 0.0165	0.9372 ± 0.0288	0.9347 ± 0.0397
	Dice	0.8626 ± 0.1540	0.9379 ± 0.0342	0.9666 ± 0.0376	0.9531 ± 0.0491	0.9485 ± 0.0505
	Val Dice	0.9085 ± 0.0482	0.9482 ± 0.0091	0.9783 ± 0.0088	0.9672 ± 0.0158	0.9657 ± 0.0222
U-net	**Loss**	0.0245 ± 0.0416	0.0321 ± 0.0240	0.0304 ± 0.0439	0.0403 ± 0.0562	0.0524 ± 0.0628
	Val Loss	0.0102 ± 0.0068	0.0230 ± 0.0064	0.0172 ± 0.0122	0.0228 ± 0.0166	0.0313 ± 0.0160
	IoU	0.7571 ± 0.2361	0.8804 ± 0.0735	0.9156 ± 0.0961	0.9174 ± 0.0989	0.9203 ± 0.0853
	Val Iou	0.8276 ± 0.0941	0.9068 ± 0.0240	0.9405 ± 0.0400	0.9494 ± 0.0298	0.9485 ± 0.0258
	Dice	0.8230 ± 0.2403	0.9255 ± 0.0536	0.9501 ± 0.0661	0.9499 ± 0.0704	0.9533 ± 0.0582
	Val Dice	0.9009 ± 0.0644	0.9445 ± 0.0135	0.9686 ± 0.0225	0.9736 ± 0.0164	0.9732 ± 0.0140



(a)



(b)

Figure. 7 U-net architecture model performance during training: (a)Semantic Segmentation – IoU and Dice and (b)Semantic Segmentation - Loss

Table 2. Comparison of prediction performance between Shallow U-net, U-net, Attention U-net and R2 U-net

Metric	**S U-net**	**U-net**[25]	**Atten Unet**[29]	**R2 Unet**[28]
IoU Score	0.9716	0.9671	0.9428	0.1766
Dice Coeff	0.9855	0.9832	0.9705	0.2960
Avg Prec	0.9864	0.9847	0.9747	0.5089
Recall	0.9911	0.9855	0.9514	0.8121

All performance metrics for both models exhibit satisfactory results, with values exceeding 0.90. Both models demonstrated the capacity to comprehend the data and perform semantic segmentation with commendable efficacy. Nevertheless, a more detailed examination of the performance data reveals that the shallow U-net model exhibits superior performance, demonstrating more stable performance across all measurements and attaining the highest performance levels.

An in-depth performance analysis is performed by focusing with σ=5 on shallow U-net architecture. The performance of the shallow U-Net architecture model during the training session can be observed in Fig. 7. In this experiment, we utilize an early stopping function to prevent overfitting, enhance generalization, and conserve computational resources. Early stopping function can be maintained by closely observing the validation loss. The performance metric in Fig. 7 indicates that the model demonstrates proficiency in learning the data. A breakdown of the loss metric in Fig. 7(b) reveals a positive trend in the data. The loss during training and the training itself improve from the earliest epoch and continue until epoch 5. After that, the trend appears to be leveling off. The IoU and Dice coefficient in Fig. 7(a) illustrate a similar pattern. The model encounters difficulties in the early epochs for both training and validation. However, the trend continues to increase, indicating that the model is learning the data correctly and reaching a plateau from epoch six onwards.

To evaluate our model's predictive capability, we compare it to the shallow U-net, the original U-net [25], Attention U-net [29] and R2 U-net [28] which employ the same U-net architecture as their backbone. Table. 2 compares all methods, this comparison is performed using IoU score, dice coefficient, average precision, and recall. All the prediction comparisons are performed by utilizing the same data. Table. 2 compares all methods; this comparison is performed using IoU score, dice coefficient, average precision, and recall. All the prediction comparisons are performed by utilizing the same data. Based on Table 2, it is shown that shallow U-net, U-net, and attention U-net perform flawlessly, reaching performance scores above 0.95 for IoU score, dice coefficient, average precision, and recall. The performance indicates that the model is learning the data very well and could perform prediction with the best performance. However, attention U-net performance is behind the original U-net and shallow U-net. This may be caused by the attention mechanism increasing model complexity, and attention gates tend to struggle with performing fine-grained segmentation.

Based on comparison between the original U-net and shallow U-net shows that both models perform very well, and both model performances have similar performance. However, our proposed model, which has less layer advantage, has more agile data learning advantages.

Based on the data in Table 2, we can observe that the recurrent residual U-Net is unable to achieve the same level of performance as the other method. R2 is experiencing difficulties in terms of IoU, Dice coefficient, and average precision. In contrast, the recall performance is relatively robust, with a value exceeding 0.80. It appears that the R2 U-Net is not an optimal choice for segmenting groups of pedestrians. Performance issues may be caused by residual connections increasing model complexity, thus

causing vanishing gradients, slower convergence, and slower learning time. Thus, the resulting R2 U-Net learning rate falls behind other methods.

The shallow U-net, as a proposed method, has shown remarkable performance across all metrics. The recall score, reaching 0.9911, is a testament to the model's exceptional ability to accurately identify true positive segmentations, leaving a strong impression of its capabilities.

5. Conclusion

In this research, we demonstrate semantic segmentation for a group of pedestrians from the JAAD dataset. We propose a direction-oriented density feature as a group of pedestrian features. The density feature is constructed by calculating pedestrian joint proximity and correlating it with another pedestrian. The density feature is constructed with several scenarios by adjusting the σ value in the Gaussian filter. Pedestrian direction is predicted using a dot product based on the shoulder, neck, and hip joint and the neck to shoulder joint as a directional reference. We perform depth modification by constructing a shallow U-net architecture.

There are six σ value scenarios, and by comparing all of the performances, the best performance is achieved with σ=5. Both shallow and original U-net architecture are performing well and have similar performance due to having the same backbone. However, based on detailed performance observation, our shallow U-net is performing better and achieving stable performance during the early sixth epoch. There is no overfitting and underfitting problem. By comparing with another model that employs U-net as its backbone, our proposed model could perform prediction flawlessly with all performance of IoU metric, Dice coefficient, average precision, and recall with a score over 0.97.

We are expected to understand dynamic pedestrian interaction more effectively by employing semantic segmentation on a group of pedestrians. By understanding pedestrian interaction patterns then, pedestrian safety can be ensured.

Conflicts of Interest

The authors guarantee that they do not have any conflicting interests.

Author Contributions

HS is accountable for the conceptualization, original draft preparation, and methodology of the project. BK is charged with the responsibility of conducting formal analysis, developing methodology, and ensuring the validity of the research process. EM will be responsible for the conceptualization, investigation, and visualization of the project, as well as providing supervision. MHP is responsible for the conceptualization of the project, as well as the writing, review, and editing of the final document.

Acknowledgments

The author would like to give special thanks to the Ministry of Finance Republic of Indonesia, for LPDP Scholarship with the Beasiswa Pendidik 2020 scheme.

References

[1] X. Hu, D. Wei, Z. Wang, J. Shen, and H. Ren, "Hypergraph video pedestrian re-identification based on posture structure relationship and action constraints", *Pattern Recognit*, Vol. 111, p. 107688, 2021, doi: 10.1016/j.patcog.2020.107688.

[2] B. Noh, W. No, J. Lee, and D. Lee, "Vision-based potential pedestrian risk analysis on unsignalized crosswalk using data mining techniques", *Appl. Sci*, Vol. 10, No. 3, 2020, doi: 10.3390/app10031057.

[3] S. Neogi, M. Hoy, K. Dang, H. Yu, and J. Dauwels, "Context model for pedestrian intention prediction using factored latent-dynamic conditional random fields", *arXiv*, pp. 1-12, 2019, doi: 10.1109/tits.2020.2995166.

[4] X. Shi *et al.*, "Verifying the applicability of a pedestrian simulation model to reproduce the effect of exit design on egress flow under normal and emergency conditions", *Phys. A Stat. Mech. its Appl.*, Vol. 562, p. 125347, 2021, doi: 10.1016/j.physa.2020.125347.

[5] A. K. Chandran, L. A. Poh, and P. Vadakkepat, "Identifying social groups in pedestrian crowd videos", In: *Proc. of ICAPR 2015 - 2015 8th Int. Conf. Adv. Pattern Recognit*, pp. 1-6, 2015, doi: 10.1109/ICAPR.2015.7050677.

[6] X. Li, S. Xiong, P. Duan, S. Zheng, B. Li, and M. Liu, "A Study on the Dynamic Spatial-Temporal Trajectory Features of Pedestrian Small Group", In: *Proc. of 2015 2nd Int. Symp. Dependable Comput. Internet Things, DCIT 2015*, pp. 112-116, 2016, doi: 10.1109/DCIT.2015.9.

[7] M. H. Zaki and T. Sayed, "Automated Analysis of Pedestrian Group Behavior in Urban Settings", *IEEE Trans. Intell. Transp. Syst.*, Vol. 19, No. 6, pp. 1880-1889, 2018, doi: 10.1109/TITS.2017.2747516.

[8] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing,

and G. Theraulaz, “The walking behaviour of pedestrian social groups and its impact on crowd dynamics”, *PLoS One*, Vol. 5, No. 4, pp. 1-7, 2010, doi: 10.1371/journal.pone.0010047.

[9] K. Cheng, Q. Liu, R. Tahir, L. Wang, and M. Li, “Logical Topology Inference via CPGCN Joint Optimizing With Pedestrian Re-Id”, *IEEE Trans. Neural Networks Learn. Syst.*, Vol. 34, No. 8, pp. 5099-5111, 2023, doi: 10.1109/TNNLS.2021.3125368.

[10] H. Cheng, Y. Li, and M. Sester, “Pedestrian group detection in shared space”, *IEEE Intell. Veh. Symp. Proc.*, Vol. 2019-June, No. Iv, pp. 1707-1714, 2019, doi: 10.1109/IVS.2019.8813849.

[11] Z. Wang, C. Cheng, and X. Wang, “A Fast Crowd Segmentation Method”, In: *Proc. of ICALIP 2018 - 6th Int. Conf. Audio, Lang. Image Process.*, pp. 242-245, 2018, doi: 10.1109/ICALIP.2018.8455441.

[12] Y. Yang and L. Lin, “Automatic Pedestrians Segmentation Based on Machine Learning in Surveillance Video”, In: *Proc. of 2019 IEEE Int. Conf. Comput. Electromagn. ICCEM 2019 - Proc.*, pp. 1-3, 2019, doi: 10.1109/COMPEM.2019.8779084.

[13] Y. Wang and X. Bai, “Intensity Inhomogeneity Suppressed Fuzzy C-Means for Infrared Pedestrian Segmentation”, *IEEE Trans. Intell. Transp. Syst.*, Vol. 20, No. 9, pp. 3361-3374, 2019, doi: 10.1109/TITS.2018.2875159.

[14] Y. Tang, Y. Wei, X. Yu, J. Lu, and J. Zhou, “Graph Interaction Networks for Relation Transfer in Human Activity Videos”, *IEEE Trans. Circuits Syst. Video Technol*, Vol. 30, No. 9, pp. 2872-2886, 2020, doi: 10.1109/TCSVT.2020.2973301.

[15] W. Yang, J. Zhang, J. Cai, and Z. Xu, “Shallow graph convolutional network for skeleton-based action recognition”, *Sensors (Switzerland)*, Vol. 21, No. 2, pp. 1-14, 2021, doi: 10.3390/s21020452.

[16] M. I. Perdana, W. Anggraeni, H. A. Sidharta, E. M. Yuniarno, and M. H. Purnomo, “Early Warning Pedestrian Crossing Intention from Its Head Gesture using Head Pose Estimation”, In: *Proc. of 2021 Int. Semin. Intell. Technol. Its Appl. Intell. Syst. New Norm. Era, ISITIA 2021*, pp. 402-407, 2021, doi: 10.1109/ISITIA52817.2021.9502231.

[17] S. Zhang, M. Abdel-Aty, Y. Wu, and O. Zheng, “Pedestrian Crossing Intention Prediction at Red-Light Using Pose Estimation”, *IEEE Trans. Intell. Transp. Syst.*, Vol. 23, No. 3, pp. 2331-2339, 2022, doi: 10.1109/TITS.2021.3074829.

[18] W. Ouyang, X. Zeng, and X. Wang, “Single-pedestrian detection aided by two-pedestrian detection”, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 37, No. 9, pp. 1875-1889, 2015, doi: 10.1109/TPAMI.2014.2377734.

[19] M. Chen, S. Banitaan, and M. Maleki, “Enhancing Pedestrian Group Detection and Tracking Through Zone-Based Clustering”, *IEEE Access*, Vol. 11, No. November, pp. 132162-132179, 2023, doi: 10.1109/ACCESS.2023.3336592.

[20] S. Islam, Z. H. Khan, T. A. Gulliver, K. S. Khattak, and W. Imran, “Pedestrian Traffic Characterization Based on Pedestrian Response”, *IEEE Access*, Vol. 10, No. October, pp. 118397-118408, 2022, doi: 10.1109/ACCESS.2022.3220321.

[21] K. Murayama, K. Kanai, M. Takeuchi, H. Sun, and J. Katto, “Deep Pedestrian Density Estimation for Smart City Monitoring”, In: *Proc. of Image Process. ICIP*, Vol. 2021-Septe, pp. 230-234, 2021, doi: 10.1109/ICIP42928.2021.9506522.

[22] J. Huo, X. Fu, Z. Liu, and Q. Zhang, “Short-Term Estimation and Prediction of Pedestrian Density in Urban Hot Spots Based on Mobile Phone Data”, *IEEE Trans. Intell. Transp. Syst.*, Vol. 23, No. 8, pp. 10827-10838, 2022, doi: 10.1109/TITS.2021.3096274.

[23] M. Wu, S. K. Lam, and T. Srikanthan, “Stereo based ROIs generation for detecting pedestrians in close proximity”, In: *Proc. of 2014 17th IEEE Intell. Transp. Syst. ITSC 2014*, pp. 1929-1934, 2014, doi: 10.1109/ITSC.2014.6957988.

[24] X. Bai, P. Wang, and F. Zhou, “Pedestrian Segmentation in Infrared Images Based on Circular Shortest Path”, *IEEE Trans. Intell. Transp. Syst.*, Vol. 17, No. 8, pp. 2214-2222, 2016, doi: 10.1109/TITS.2016.2516342.

[25] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, *Medical Image Computing and Computer-Assisted Intervention*, Vol. 9, 2015, pp. 234-241. doi: 10.1007/978-3-319-24574-4_28.

[26] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, “Deep multimodal fusion for semantic image segmentation: A survey”, *Image Vis. Comput.*, Vol. 105, p. 104042, 2021, doi: 10.1016/j.imavis.2020.104042.

[27] P. Wang and X. Bai, “Thermal Infrared Pedestrian Segmentation Based on Conditional GAN”, *IEEE Trans. Image Process.*, Vol. 28, No. 12, pp. 6007-6021, 2019, doi: 10.1109/TIP.2019.2924171.

[28] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation", 2018, [Online]. Available: http://arxiv.org/abs/1802.06955

[29] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas", no. Midl, 2018, [Online]. Available: http://arxiv.org/abs/1804.03999

[30] Y. Pang, J. Cao, Y. Li, J. Xie, H. Sun, and J. Gong, "TJU-DHD: A Diverse High-Resolution Dataset for Object Detection", *IEEE Trans. Image Process.*, Vol. 30, pp. 207-219, 2021, doi: 10.1109/TIP.2020.3034487.

[31] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite", In: *Proc. of IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, pp. 3354-3361, 2012, doi: 10.1109/CVPR.2012.6248074.

[32] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking", In: *Proc. of IEEE Int. Conf. Comput. Vis.*, No. Iccv, pp. 261-268, 2009, doi: 10.1109/ICCV.2009.5459260.

[33] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior", In: *Proc. of 2017 IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017*, Vol. 2018-Janua, pp. 206-213, 2017, doi: 10.1109/ICCVW.2017.33.

[34] M. Cordts *et al.*, "The Cityscapes Dataset for Semantic Urban Scene Understanding", In: *Proc. of IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Vol. 2016-Decem, pp. 3213-3223, 2016, doi: 10.1109/CVPR.2016.350.

[35] C. R. Dow, H. H. Ngo, L. H. Lee, P. Y. Lai, K. C. Wang, and V. T. Bui, "A crosswalk pedestrian recognition system by using deep learning and zebra-crossing recognition techniques", *Softw. - Pract. Exp.*, Vol. 50, No. 5, pp. 630-644, 2020, doi: 10.1002/spe.2742.

[36] R. Li, J. Xiang, F. Sun, Y. Yuan, L. Yuan, and S. Gou, "Multiscale Cross-modal Homogeneity Enhancement and Confidence-aware Fusion for Multispectral Pedestrian Detection", *IEEE Trans. Multimed.*, Vol. PP, pp. 1-13, 2023, doi: 10.1109/TMM.2023.3272471.

[37] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields", In: *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1302-1310, 2017. doi: 10.1109/CVPR.2017.143.