# Context Aware Multimodal Fusion YOLOv5 Framework for Pedestrian Detection under IoT Environment

*Yuan SHU [1,2], Youren WANG [1], Min ZHANG [2], Jie YANG [2], Yi WANG [2], Jun WANG [2], Yunbin ZHANG [3]*

[1] College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, Jiangsu, P.R. China
[2] Nanjing Vocational Institute of Railway Technology, Nanjing 210031, Jiangsu, P.R. China
[3] Jinan Electric Service Section, China Railway Jinan Bureau Group Co., Ltd, Jinan 250023, Shandong, P.R. China

shuyuan@whu.edu.cn, wangyrac@nuaa.edu.cn, zm90zhangmin@163.com, yangjyy221@163.com, wangyi8098@nuaa.edu.cn, wangjunsnn@163.com, zhangyunbinAa@163.com

Submitted February 24, 2024 / Accepted December 9, 2024 / Online first March 21, 2025

**Abstract.** *Pedestrian detection based on deep networks has become a research hotspot in the field of computer vision. With the rapid development of the Internet of Things (IoT) and autonomous driving technology, the deployment of pedestrian detection models on mobile devices places higher demands on the accuracy and real-time performance of detection. In addition, fully integrating multimodal information can further improve the robustness of the model. To this end, this article proposes a novel multimodal fusion YOLOv5 network for pedestrian detection. Specifically, to improve the performance of multi-scale pedestrian detection, we enhance contextual awareness abilities by embedding the multi-head self-attention (MSA) mechanism and graph convolution operations in the existing YOLOv5 framework. In addition, we can fully explore the real-time advantages of the YOLOv5 framework in pedestrian detection tasks. To improve multimodal information fusion, we introduce the joint cross-attention fusion mechanism to enhance knowledge interaction between different modalities. To validate the effectiveness of the proposed model, we conduct a large number of experiments on two multimodal pedestrian detection datasets. All the results confirm that our proposed model obtains the highest performance in terms of multi-scale pedestrian detection. Moreover, compared to other multimodal deep models, our proposed model still shows superior performance.*

## Keywords

Pedestrian detection, IoT, deep learning, multimodal fusion, YOLOv5

## 1. Introduction

Deep learning-based object detection algorithms have obtained high detection accuracy in fields such as unmanned driving, video fire detection, safety monitoring, and drone scene analysis [1–3]. Pedestrian detection is an important branch of object detection and is widely used in fields such as intelligent security, vehicle-assisted driving, and intelligent transportation [4], [5]. It has enormous social and economic value and has become a research hotspot in computer vision [6]. Existing convolution neural networks (CNNs)-based pedestrian detection is mainly divided into one-stage models and two-stage models. Two-stage models divide pedestrian detection into two parts for processing, i.e., pedestrian identification and localization. The regions with convolutional neural networks (R-CNN) which was proposed by Girshick, R. et al. [7] extracted a set of candidate boxes from the target object, and finally uses a support vector machine classifier (SVM) to predict pedestrian targets by exploiting pedestrian features [8]. To improve the efficiency of candidate box generation, Li et al. [9] proposed a novel fast R-CNN framework in which multiple built-in sub-networks were introduced. Then, outputs from all the sub-networks were adaptively combined to generate the final pedestrian detection results. To further improve the speed of pedestrian detection, Yu et al. [10] proposed an effective faster RCNN model by combining feature concatenation and hard negative mining strategies to boost performance. Although these two-stage pedestrian detection algorithms have achieved good performance in detection accuracy, due to the high computational complexity, these models are difficult to effectively deploy in scenes with high real-time requirements.

With the rapid progress of the IoT and electronic technology, pedestrian detection based on mobile devices plays an important role in daily life which is shown in Fig. 1. To improve the effectiveness of deep network deployment, some requirements, i.e. real-time performance, computational complexity, and memory usage of the model are becoming more important. Some one-stage detection networks, such as single shot mul-tiBox detector (SSD) [11], you only look once (YOLO) [12], [13], and variants YOLO networks [14], effectively improve real-time performance in object detection

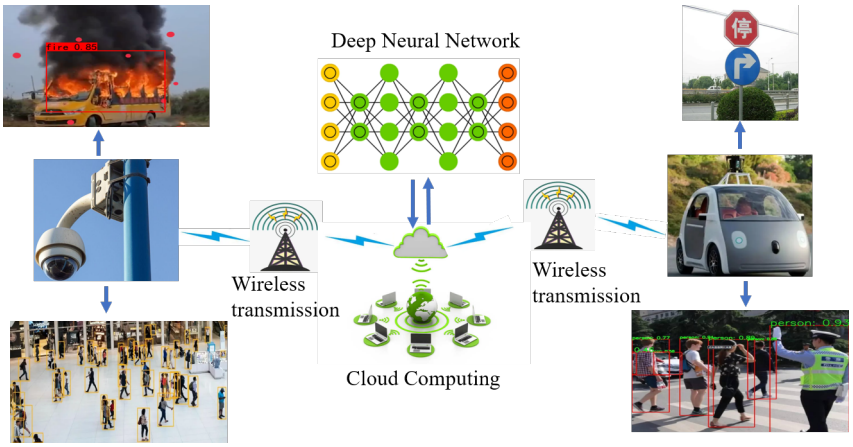


**Fig. 1.** Various applications of YOLO framework based on IoT technology.

tasks. In the task of pedestrian detection, Lan et al. [15] proposed a new YOLO-R network structure based on the YOLOv2 framework. The YOLO-R network alleviated the appearance of pedestrians by adding three passthrough layers, thereby improving the accuracy of pedestrian detection. Jiang et al. [16] proposed a YOLOv5s-based framework that combines image and video to address the issues of complex object scenes and low video resolution. To improve low real-time performance, Li et al. [17] proposed a variant YOLO algorithm by using multi-scale detection maps. Although these YOLO-based models achieve higher real-time performance, the accuracies of these models need to be improved, especially for small-scale pedestrian detection. To this end, Hsu et al. [18] proposed a novel YOlOv3-based model named ratio-and-scale-aware YOLO (RSA-YOLO) which designed a novel scale-aware mechanism by exploiting multiresolution fusion mechanism for small pedestrian detection. Hsu et al. [19] adopted different networks to extract features. Moreover, these extracted features were adaptively fused to generate the final detection results. To effectively explore the global contextural information, Hou et al. [20] designed the M-YOLO model to preserve multi-scale information and enhance the expressiveness of features. To make full use of multi-scale contextural information, Xue et al. [21] proposed a multi-scale pedestrian detection method named MRFC which exploited the global and local attention and multi-scale receptive field context mechanisms for pedestrian detection. However, the real-time performance of network MRFC is not as good as the YOLO structure. Therefore, effectively integrating contextual information with existing YOLO frameworks can provide strong guarantees for the accuracy and timeliness of multi-scale pedestrian detection.

Despite the limited resources of IoT devices, the collection of multimodal data is more convenient. The fusion of multimodal data can fully compensate for the shortcomings of a single modality and effectively improve the robustness of the detection system. In multimodal pedestrian detection, Xue et al. [22] designed a novel multi-modal attention fusion YOLO (MAF-YOLO) in which the compressed Darknet53 framework was constructed to extract multimodal features. Then, the modal weighted fusion module was proposed to enhance the detection accuracy. Dasgupta et al. [23] used the deformable ResNeXt-50 encoders for feature extraction from RGB and thermal images. Moreover, a multimodal feature embedding module (MuFEm) was proposed to improve feature fusion between different modalities. Kolluriet et al. [24] proposed an effectively deep model named IPDC-HMODL which combined hybrid metaheuristic optimization with a deep learning algorithm for multimodal pedestrian detection. The IPDC-HMODL used the YOLOv5 for pedestrian feature extraction. The entire model achieved good real-time detection performance. To explore the potential of modality-specific features, Lee et al. [25] proposed the cross-modality attention transformer (CAT) model in which the multimodal fusion transformer (MFT) was designed to perform feature fusion. Li et al. [26] designed a novel visible and thermal image fusion approach which adopted some specific mechanisms, i.e., gated unit, parameter optimization, soft treatment, and parameter adaption, to enhance the sensor fusion performance. Although these multimodal pedestrian detection models achieve good robustness, how to effectively fuse information from different modalities has always been a focus of multimodal pedestrian detection.

Inspired by the above research, we propose a new multimodal fusion YOLOv5 network for pedestrian detection. This network can fully utilize the real-time performance of the YOLO framework in pedestrian detection tasks, providing convenience for the mobile deployment. In addition, by introducing a multimodal fusion mechanism based on joint cross-attention fusion mechanism, the robustness is effectively improved. Specifically, we embed the multi-head attention module and the graph convolution module into the existing YOLOv5 framework. The proposed model can effectively mine global and local contextual relationships, effectively improving the accuracy of pedestrian detection. To improve the fusion of multimodal information, we exchange the learned features between different modalities to improve information fusion between modalities. Compared to existing multi-modal fusion deep models, our main contributions are summarized as follows:

1. We design a novel and effective variant YOLOv5 framework for pedestrian feature extraction. To effectively alleviate the problem of insufficient accuracy in multi-scale object detection in YOLOv5, we fully explore the contextual information of pedestrian targets by embedding the multi-head attention module and graph convolution layers, where the multi-head attention module can mine the global contextual information, and the graph can exploit the local contextual information.
2. We propose an effectively multi-modal fusion deep model for multi-scale pedestrian detection, in which we design a novel variant YOLOv5 framework to extract pedestrian features, and we introduce the exchanging-based fusion module in which the cross-attention fusion mechanism is designed to exchange knowledge between different modalities. The exchanging-based fusion module can effectively explore the inter-modal dependencies to improve the robustness of pedestrian detection.
3. We conduct extensive experiments on two multimodal pedestrian detection datasets. The proposed model can effectively exploit the inter-modality and intra-modality relationships. Moreover, the experiments show the proposed model still achieves a higher recognition rate compared to existing other deep multimodal fusion networks.

The rest of the paper is organized as follows. Section 2 reviews the related works. In Sec. 3, the main method of the YOLOv5 network, multi-head attention mechanism, and graph convolution operation are described in detail. The experiments and results are presented in Sec. 4. Section 5 gives the conclusions.

# 2. Related Works

## 2.1 Deep Learning-Based Pedestrian Detection

In the early research work of pedestrian detection, a large number of hand-crafted features, such as scale-invariant feature transform (SIFT) [27] and histogram of directional gradients (HOG) [28], were extracted from the target area and sent to different classifiers, such as SVM [29] and Ada Boost algorithms, for final pedestrian detection. In pedestrian detection task, feature selection is particularly important. To this end, the fusion of multiple hand-crafted features effectively improves the accuracy of pedestrian detection [30]. Although these models have achieved good performance in pedestrian detection tasks, their accuracy in pedestrian detection has significantly decreased in complex scenarios [31].

With the rapid development of deep networks and artificial intelligence theory, pedestrian detection based on deep learning has been widely applied and studied. The pedestrian detection methods based on convolution neural networks can be divided into two-stage detection and one-stage detection. In the two-stage detection, representative methods include R-CNN, Fast R-CNN, and Fast R-CNN [32]. In the one-stage based pedestrian detection, representative methods include SSD [33] and YOLO series [34], [35]. Due to the superior detection accuracy of the two-stage model compared to the one-stage model, the overall computational complexity of the two-stage model makes it difficult to effectively deploy to resource-constrained mobile devices. The one-stage models can achieve a good balance between detection accuracy and real-time performance, where YOLO series [36] have obtained the best performance in practical applications. Sukkar et al. [37] proposed a real-time pedestrian detection model based on YOLOv5. To further improve the performance of YOLOv5, Lv et al. [38] designed some specific mechanisms, i.e., the L1 regularization is introduced before the BN layer, context information was exploited to extend the receptive fields of different sizes. In this paper, we designed an efficient deep model by embedding the multi-head attention mechanism and graph convolution operation in YOLOv5 for multi-scale pedestrian detection.

## 2.2 Multimodal Fusion-Based Pedestrian Detection

With the increasing complexity of pedestrian detection scenarios, using only single-modal data cannot achieve satisfactory results [39]. In recent years, with the rapid development of various sensor technologies, the collection of multimodal data has become simpler. Multimodality object detection has attracted widespread attention [40], [41]. By using multimodality information, it is possible to fully explore the complementarity between modalities and effectively improve the robustness. In pedestrian detection tasks, more and more research work is using multimodal information to construct more efficient pedestrian detection systems.

To improve the mobile deployment of multimodal models, Dradrach et al. [42] designed pedestrian features based on the YOLOv5s framework from infrared and visible light images. The frame rate of the entire model is approximately 7 FPS. Wang et al. [43] proposed a novel multimodal YOLOv3 model named MDY for pedestrian detection on embedded devices. The MDY used the optimized anchor frames and added the small target detection branches to improve detection accuracy. Cao et al. [44] designed a simple and efficient YOLOv4 model to extract multimodal features from the color stream and thermal stream. Moreover, different fusion mechanisms, i.e., early fusion, halfway fusion, late fusion, and direct fusion were further validated. To alleviate the imbalance problem during feature fusion, Das et al. [45] proposed a novel training setup with a regularizer in the feature extraction model. To dynamically fuse multimodal information, Li et al. [46] proposed a novel confidence-aware method, which can generate a reliable result according to the confidence of different modalities. Moreover, Dempster's combination rule is introduced to produce the final output. To alleviate the redundant information during multimodal fusion, Wang

et al. [47] designed a novel multimodal pedestrian detection model named RISNet which suppresses cross-modal redundant information between RGB and infrared features. Although these deep models have achieved higher performance by using multimodal information, building efficient fusion modules has become an urgent issue for multimodal pedestrian detection tasks.

### 2.3 Transformer-Based Object Detection

Although traditional object detection networks based on convolutional operations have achieved great success, convolutional operations can effectively focus on the aggregation of local information through parameter sharing, but this also leads to insufficient modeling of global information. In object detection tasks, multi-scale object detection requires attention to global feature information. Recently, Transformer-based models which use the multi-head self-attention mechanism can model global dependencies [48]. The first Transformer-based model for object detection is DETR [49]. In the DETR detector, the Transformer model can analyze the relationship between the target and global features, predict a set of objects in parallel at once, and allow for modeling their relationships. Secondly, DETR adopts a set-based global loss approach and uses binary matching to make one-on-one predictions between predictions and real boxes. The main advantage of DETR is that it simplifies the detection process and eliminates reliance on hand-crafted modules and operations, such as region proposal networks (RPN) and non-maximum suppression (NMS) commonly used in object detection. To alleviate convergence and feature space resolution issues, Zhu et al. [50] proposed a new object detection model named Deformable DETR, which uses a deformable attention module that only cares about a small group of key sampling points around the reference point, without considering the size of the feature map. By combining multi-scale features, deformable attention is extended to multi-head deformable attention modules to achieve multi-scale feature maps. To reduce the computational complexity of the Transformer model, Reese et al. [51] proposed the sparse DETR which only updates some encoder tokens. The experimental results show that sparse DETR effectively reduces the computational complexity without significantly reducing detection performance. Recently, more and more Transformer-based models have been proposed, such as Anchor DETR [52], Dynamic DETR [53] and Conditional DETR [54]. These models fully utilize the advantages of the multi-head self-attention mechanism in long-range dependency modeling. In this paper, we design a multi-modal fusion framework by embedding the multi-head self-attention mechanism in the existing YOLOv5 for pedestrian detection.

## 3. The Proposed Multimodal Fusion YOLOv5 Framework

The development of deep learning technology has provided great convenience for the application of pedestrian detection, among which pedestrian detection based on the YOLO framework has achieved good comprehensive performance in pedestrian detection tasks. However, in complex scenarios, there are still significant challenges for multi-scale pedestrian detection by using the YOLOv5 framework. To this end, this article embeds the Transformer model and graph convolution operation into the existing YOLOv5 model, fully mining the contextual information of pedestrian targets, and constructing an efficient multi-scale pedestrian detection model. To further improve the robustness of pedestrian detection, this paper constructs a more robust pedestrian detection model by fusing multimodal information. The entire model is shown in Fig. 2. The entire framework includes input, feature extraction, joint cross-attention multimodal fusion and pedestrian prediction modules.

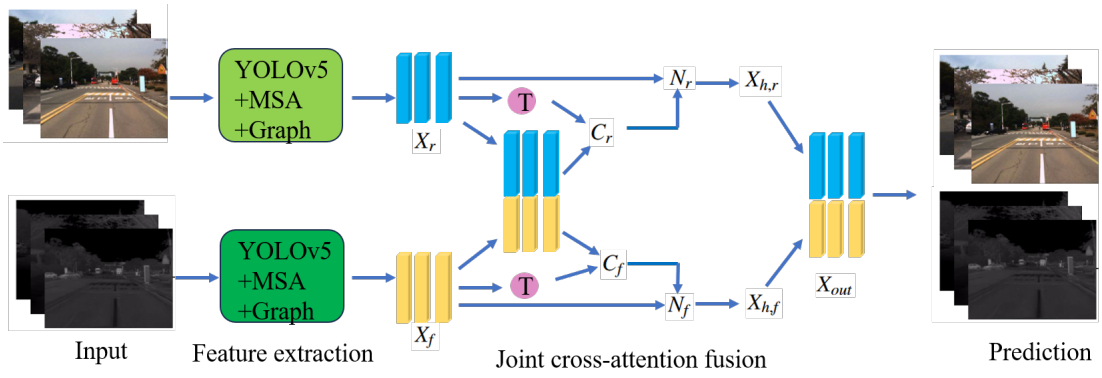


**Fig. 2.** The overview of the proposed joint cross-attention multimodal fusion YOLOv5 framework for pedestrian detection, in which we exploit multimodal information as input. Then we embed MSA and graph convolution in YOLOv5 for feature extraction. Moreover, we introduce the joint cross-attention mechanism for multimodal information fusion, where $\mathbf{X}_r$ and $\mathbf{X}_f$ denote the deep features from different modalities, $\mathbf{C}_r$ and $\mathbf{C}_f$ denotes the joint relation matrices of different modalities, $\mathbf{N}_r$ and $\mathbf{N}_f$ denote the attention weights of different modalities, $\mathbf{X}_{h,r}$ and $\mathbf{X}_{h,f}$ denote the weighted features of different modalities, T denote the transposition operation, $\mathbf{X}_{\text{out}}$ denotes the fused features from different modalities.

### 3.1 Feature Extraction Based on YOLOv5

Considering the real-time and accuracy of pedestrian detection, the YOLO-based detection models achieve high performance [55]. In addition, YOLO-based models also provide great convenience for mobile deployment. The existing YOLOv5 model has good accuracy and speed in pedestrian detection tasks, so this paper uses the YOLOv5 framework which is shown in Fig. 3 as the backbone for feature extraction. The existing YOLOv5 backbone network consists of different modules, including focus, consist conv-bn-SiLU (CBL) module which is lightweight convolutional block, Resunit, spatial pyramid pooling (SPP) and cross-stage partial (CSP) which is divides to two kinds models, CSP1_X and CSP2_X to extract features of varying granularity from the image. In addition, some operations, including CONV, Concat and upsampling are introduced, in which, CONV denotes tradition convolution operation, Concat denotes concatenation operation of tensor features, and upsampling is used for feature sampling. To increase the capacity of the entire dataset and reduce the consumption of GPU memory, the YOLOv5 framework adopts Mosaic data augmentation operations at the input end. The main idea of the Mosaic data augmentation method is to concatenate four randomly cropped images onto one image as training images. The dataset is enriched through this method and the robustness of the network is better. For adaptive anchor box calculation, in the YOLOv5 algorithm, to enable the network to learn better detectors, all pedestrians in the video use the default label box spacing. During training, a prediction box is output based on this, which is convenient for comparing the initial box with the prediction box to calculate the difference. Using the k-means clustering method, the evaluation indicator for defining distance is changed to the intersection and union ratio (IOU) between the anchor and bounding box, which improves the numerical value. For adaptive image scaling, in the YOLOv5 algorithm, the image input network needs to be uniformly scaled to a single size. Using Letterbox adaptive scaling technology during scaling can eliminate redundancy and improve computational speed.

The backbone network of YOLOv5 is mainly responsible for extracting image features during the detection process. It adopts fewer layers and fewer parameters, but has the same feature expression ability as larger networks. The SPP module performs four different scales of maximum pooling on the input, and then concatenates the pooled results as fused feature outputs. This structure allows the network to receive input images of random sizes, making it robust for detecting large and small targets. The Focus module performs slicing operations on images, to sacrifice a small amount of computation during the downsampling process to concentrate the width and height information on the channel dimension, so that more comprehensive and sufficient feature information can be obtained when performing convolution operations to extract features. CSP1_X and CSP2_X modules can help the YOLOv5 structure achieve richer gradient combinations, thereby reducing computational complexity. The output part of YOLOv5 mainly adopts the complete intersection over union loss function, which can alleviate the problem of different relative positions when the current detection box and target box do not coincide. The output end is used to complete the output of target detection results, where the number of branches varies depending on the algorithm, usually including classification branches and regression branches. The loss function is the classification loss function and regression loss function.

### 3.2 Contextual Information Modeling Based on Transformer and Graph Convolution Operation

Although the YOLOv5 model has improved in speed and flexibility, there are still shortcomings in performance, especially its detection accuracy for small and medium-sized targets is not high. To improve the performance of multi-scale pedestrian detection, the contextual information of pedestrian features has a positive impact on the improvement of detection performance. The multi-head self-attention mechanism can effectively model the global dependencies of features.

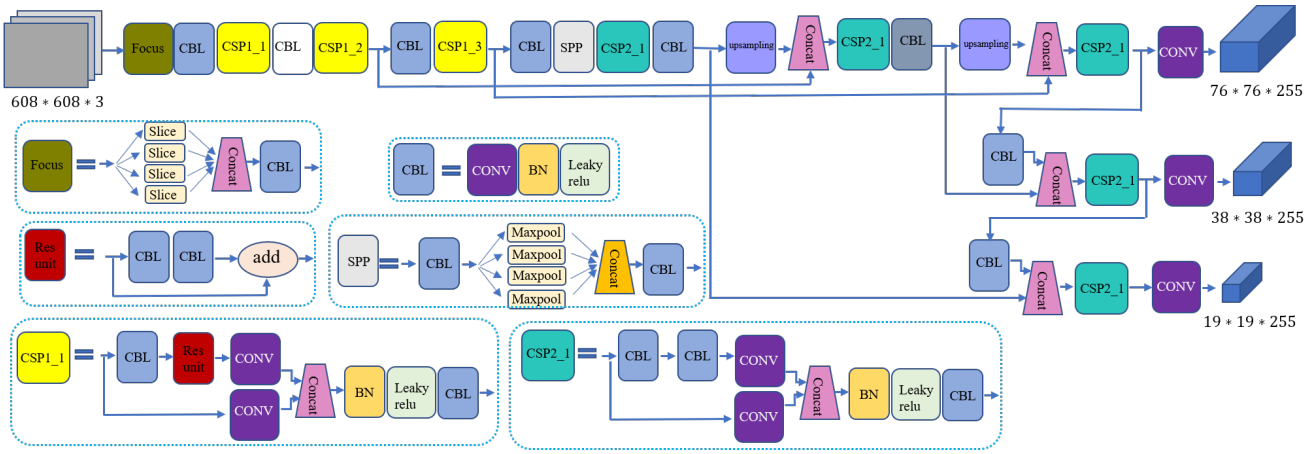


**Fig. 3.** Overview of the YOLOv5 framework.

In addition, the local contextual information of features plays an important role in small-scale object detection, and graph convolution operations effectively improve the modeling of local dependencies by aggregating local features. Therefore, this paper proposes a new variant YOLOv5 framework for multi-scale pedestrian detection by embedding Transformer and graph convolution operations in the existing YOLOv5 framework.

Each element has a specific position in sequence data, and the sequential relationship between positions has a significant impact on the processing of data. The self-attention mechanism allows the model to consider the relationships among elements when processing a sequence. This mechanism can help the model better understand the contextual information in the sequence, thus processing the sequence data more accurately. In the self-attention mechanism, the model calculates the correlation degree of different elements. The correlation degree reflects the inter-relationships between elements, such as in language models, which can reflect the semantic correlation between words. To improve the modeling of global contextual feature information, we replace CSP1_x with the multi-head self-attention mechanism in the existing YOLOv5 framework. The original image is feature extracted through an improved backbone network, convolutional module, and a series of CBL, CSP1, and SPP structures. The extracted features are then processed through the MSA.

After the feature extraction from the SPP layer, the location information, linear correlation information and category marks are used for constructing the embedded block vector. We denote all the patch features as $\mathbf{X} \in R^{N\times L}$, where $L$ is the dimension of patch features. Then, three projected patch-embedded matrices, i.e., query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$ are learned from $\mathbf{X} \in R^{N\times L}$ are defined as follows:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_{\mathrm{Q}}, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_{\mathrm{K}}, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_{\mathrm{V}} \tag{1}$$

where $\mathbf{W}_{\mathrm{Q}}, \mathbf{W}_{\mathrm{K}}, \mathbf{W}_{\mathrm{V}}$ denote three learnable matrices.

The multi-head self-attention mechanism improves the spatial resolution, concurrency, and computational efficiency of the attention mechanism. The MSA is defined as follows:

$$\mathrm{MSA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{concat}\left[\mathrm{OneHead}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i)\right]_{h=1,\ldots,H}\mathbf{W}_{\mathrm{o}} \tag{2}$$

where concat denotes the concatenation operation, and $\mathbf{W}_{\mathrm{o}}$ is the learnable matrix. The $\mathrm{OneHead}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i)$ which is denoted as one-head self-attention is calculated as follows:

$$\mathrm{OneHead}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i) = \mathrm{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^{\mathrm{T}}}{\sqrt{L}}\right)\mathbf{V}_i \tag{3}$$

where $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i$ denote the query, key and value matrices in the $i$th head in the MSA. In each MSA, we adopt the residual connection. The output of the transformer encoder $\mathbf{X}_{\mathrm{o}}$ is defined as follows:

$$\mathbf{X}_{\mathrm{o}} = \mathbf{X} + \mathrm{MSA}(\mathbf{Q},\mathbf{K},\mathbf{V}). \tag{4}$$

Although the transformer encoder can obtain global contextual information, there is a lack of local contextual information, which plays an important role in small-scale object detection. To further mine the local textual information of in each object, we embed graph convolution operation. To improve the modeling of local dependencies, we fed the output of the Transformer module into the graph convolution operation module. Then the proceed features from MSA have performed feature fusion through FPN and PAN structures, and pass the CSP2 and convolution structures to obtain the final prediction layer which can produce three kinds of size prediction. First, all patch embeddings from the transformer encoder are adopted to construct pedestrian graph data, which can be denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $v_i \in \mathcal{V}$ and $e_{i,j} = (v_i, v_j) \in \mathcal{E}$ are the set of vertices and edges, respectively. $\mathbf{A}$ is the adjacent matrix which is defined as follows:

$$A_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

In this paper, we exploit the structural contextual information by performing graph convolution operation [56]. First, we calculate the attention weights of neighbors based on their features. The attention weights $\alpha_{uv}$ are defined as follows:

$$\alpha_{uv} = \frac{\exp(h_{uv})}{\sum_{w\in\mathbf{N}_v}\exp(h_{wv})}, \tag{6}$$

$$h_{uv} = \sigma(A_{uv}\cdot\mathbf{a}^{\mathrm{T}}[w^{\mathrm{s}}x_u || W^{\mathrm{s}}x_v])\forall(u,v)\in\mathcal{G} \tag{7}$$

where $\mathbf{N}_v$ denote the neighbor nodes of node $v$, $W^{\mathrm{s}} \in R^{F\times D}$ denote the learnable parameters, $\sigma(\cdot)$ denote the activation function, $\mathbf{a}$ denote the weight vector parameterizing the attention function implemented as a feed-forward layer. The output of graph convolution is defined as follows:

$$z_v = \sigma\left(\sum_{w\in\mathbf{N}_v}\alpha_{uv}W^{\mathrm{s}}x_u\right). \tag{8}$$

By constructing graph data and introducing a structural attention mechanism, we can node embedding through self-attention aggregation of adjacent node embedding, which can be seen as a single message passing between direct neighbors. This operation can effectively model local contextual information. By integrating the Transformer module with graph convolution operations, the model's modeling of global and local context dependencies can be effectively improved. The overview of our proposed variant YOLOv5 framework is shown in Fig. 4. The entire embedded module effectively improves the performance of the model for multi-scale pedestrian detection. We will further validate the effectiveness of our proposed model in subsequent experiments.

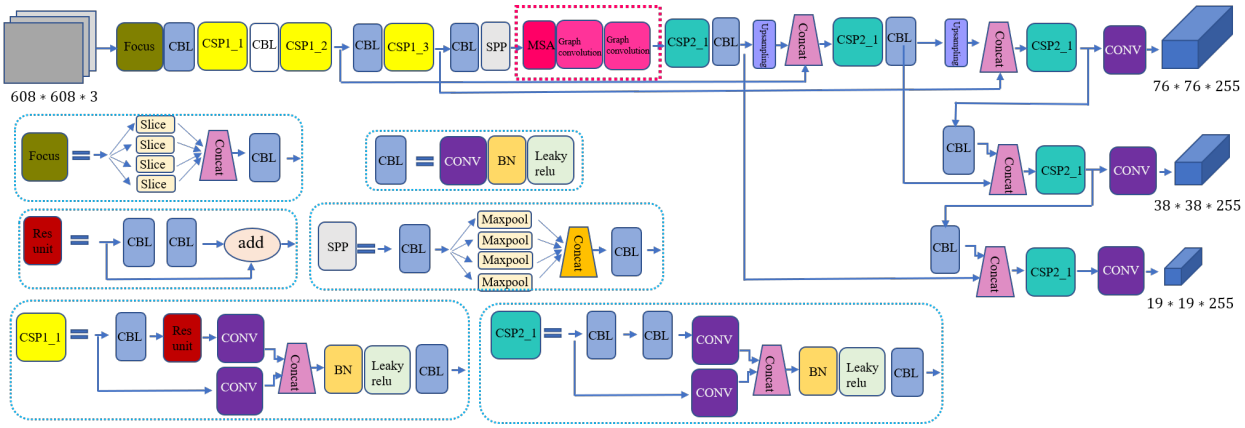


**Fig. 4.** The overview of our proposed YOLOv5 framework in which we replace the CSP1_X with MSA. Moreover, after the MSA, we embed two graph convolution modules.

### 3.3 Multimodel Feature Fusion Based on Joint Cross-Attention Mechanism

With the rapid development and deployment of IoT devices, the collection, processing, and fusion of multimodal information has become a mainstream method in the field of computer vision. Moreover, multimodal pedestrian detection can effectively alleviate the performance degradation caused by light and multi-scale dense targets. In this paper, we introduce the cross-attention mechanism [57] for pedestrian detection. Specifically, we use the proposed variant YOLOv5 framework to extract deep features from different modalities $\mathbf{X}_r \in R^{d_r \times L}$ and $\mathbf{X}_f \in R^{d_f \times L}$.

Firstly, we construct a joint feature representation by directly concatenating the features from two modalities. The joint feature representation is denoted as $\mathbf{J} = [\mathbf{X}_r; \mathbf{X}_f] \in R^{d \times L}$ where $d = d_r + d_f$. We calculate the two joint relation matrices of different modalities as follows:

$$\mathbf{C}_r = \tanh\left(\frac{\mathbf{X}_r^{\mathrm{T}} \mathbf{W}_{jr} \mathbf{J}}{\sqrt{d}}\right), \tag{9}$$

$$\mathbf{C}_f = \tanh\left(\frac{\mathbf{X}_f^{\mathrm{T}} \mathbf{W}_{jf} \mathbf{J}}{\sqrt{d}}\right) \tag{10}$$

where $\mathbf{C}_r$ and $\mathbf{C}_f$ denote different joint relation matrices, and $\mathbf{W}_{jr}$ and $\mathbf{W}_{jf}$ denote two learnable parameters. The joint correlation matrices provide a semantic measure of inter-modal and intra-modal relevance.

Then, we calculate attention weights for different modalities. The attention weights of different modalities are calculated as follows:

$$\mathbf{N}_r = \mathrm{ReLu}(\mathbf{W}_r \mathbf{X}_r + \mathbf{W}_{cr} \mathbf{C}_r^{\mathrm{T}}), \tag{11}$$

$$\mathbf{N}_f = \mathrm{ReLu}(\mathbf{W}_f \mathbf{X}_f + \mathbf{W}_{cf} \mathbf{C}_f^{\mathrm{T}}) \tag{12}$$

where $\mathbf{N}_r$ and $\mathbf{N}_f$ denote two attention weights, $\mathbf{W}_{cr}$,$\mathbf{W}_r$, $\mathbf{W}_{cf}$,$\mathbf{W}_f$ denote learnable weight matrices. Then, we use the attention weight to calculate the output of weighted features.

The weighted features of different modalities are calculated as follows:

$$\mathbf{X}_{h,r} = \mathbf{W}_{nr} \mathbf{N}_r + \mathbf{X}_r, \tag{13}$$

$$\mathbf{X}_{h,f} = \mathbf{W}_{nf} \mathbf{N}_f + \mathbf{X}_f \tag{14}$$

where $\mathbf{X}_{h,r}$ $\mathbf{X}_{h,f}$ denote two weighted features, and $\mathbf{W}_{nr}$, $\mathbf{W}_{nf}$ denote learnable weight matrices.

Finally, the fused features from different modalities are calculated as follows:

$$\mathbf{X}_{\mathrm{out}} = [\mathbf{X}_{h,r}; \mathbf{X}_{h,f}]. \tag{15}$$

By introducing a joint cross-attention fusion model, we can effectively utilize complementary modal relationships. The cross-attention mechanism can effectively improve the performance of multimodal feature fusion by using joint feature representations and the correlation between individual modalities to calculate cross-attention weights. To effectively model local and global dependencies of features, we add one MSA layer and two graph convolution layers after the SPP layer of the YOLOv5 framework. The MSA layer has four OneHead modules, and the dimension of each OneHead is 32. The input and output dimensions of the graph convolution layer are the same as the MSA layer.

## 4. Experiments and Analysis

In this section, we introduce the preprocessing and implementation details of experiments. Then, we compare our proposed multimodal fusion model with existing multimodal fusion mechanisms. In addition, we further validate the effectiveness of our proposed model with different backbones. Finally, we conduct ablation studies to validate different modules in our proposed model.

### 4.1 Datasets and Experimental Settings

The KAIST dataset is the most commonly used public dataset for pedestrian detection [58]. This dataset is a visible infrared pedestrian image collected by KAIST University in South Korea in 2015. The KAIST dataset captured various conventional traffic scene images at different periods, including campus, street, and rural areas. It consists of a total of 95328 paired visible and infrared image pairs, each with an image size of $640 \times 512$, including a total of 103128 targets from 1182 pedestrians. Paired datasets are uniformly collected during the day and night, so that the obtained dataset contains complete brightness conditions. The dataset is divided into a total of 12 folders, named set00 to set11 in sequence. Among them, the first 6 folders are training sets, containing 50172 images; The last six folders are the test set, containing 45156 images. The collection time for set06, set07, and set08 is daytime, while the collection time for set09, set10, and set11 is nighttime. Due to certain issues with raw data annotation and for a more fair comparison, this paper uses the improved annotation as the training label according to the reference [59]. The test set for this dataset includes a total of 2252 pairs of visible and infrared data, of which 1455 pairs come from daytime and 797 pairs come from nighttime. Each test image is obtained by capturing one frame every 20 frames in a continuous video according to the reference [60]. CVC-14 is composed of visible and infrared images collected by a car-mounted camera. The environment for collecting the dataset is on the streets during the day and at night. The entire dataset contains 7085 frames of visible and infrared image pairs. The visible image is presented as a single-channel grayscale image, with each pair of images having a size of $640 \times 471$. There are a total of 1433 pairs of test sets for this dataset. When labeling pedestrians, we use visible light mode labeling as the final standard [60]. Due to the issue of modal misalignment in the CVC-14 dataset, independent annotations were made on visible and infrared images. This dataset is more challenging and poses a certain challenge to the network's ability to handle modal misalignment. Some samples of these two datasets are shown in Fig. 5.

In this experiment, the missing rate ($MR^{-2}$) was selected as the evaluation indicator for algorithm performance. $MR^{-2}$ uses false positive per image (FPPI) as the horizontal axis and log (MR) as the vertical axis of the curve. Uniformly select 9 FPPIs within the range of $[0.01, 1]$ to obtain their corresponding log (MR) values, and average these vertical coordinate values. Finally, restore the above average values to the percentage form of MR through exponential operation, and obtain the MR2 indicator used to quantify the MR-FPPI curve. The smaller the indicator $MR^{-2}$, the higher the detector performance. For the fairness of the experiment, we set the intersection over union (IoU) value to be greater than or equal to 0.5, which determines that the pedestrian has been successfully detected. Otherwise, it is considered a detection failure. At the same time, pedestrians who have repeatedly been detected will also be considered as detection failures. Due to inconsistent image sizes between the two datasets, it is necessary to adjust the image size of the CVC-14 dataset to $640 \times 512$. To improve the stability of training, the network fixes the first 10 layers of weights like the standard Faster R-CNN, and its values always remain consistent with the pre-training weight values on ImageNet. We use the stochastic gradient descent (SGD) method for training. The initial training learning rate is 0.001, and the batch size is set to 4. After 3 iterations (Epoch), the learning rate decreases to 0.1 times the original. The CPU of our computer is based on 3.6 GHz Intel i7-9700K. The operating system is Windows 10. All deep models run on Anaconda 3.0 software. All code in this paper is written based on Pytorch 1.5, and Python 3.7. The version of CUDA is 10.1.

### 4.2 Effectiveness of Different Deep Models

In this section, we compared our proposed model with other models on KAIST dataset. We adopt three scenarios for pedestrian detection: all pedestrians (All), daytime pedestrians (Day), and nighttime pedestrians (Night). When evaluating All, Day, and Night, we set the pedestrian target pixel to be higher than 55. Moreover, we only detect pedestrians who are not obstructed or partially obstructed. The results of different model detection are shown in Tab. 1.

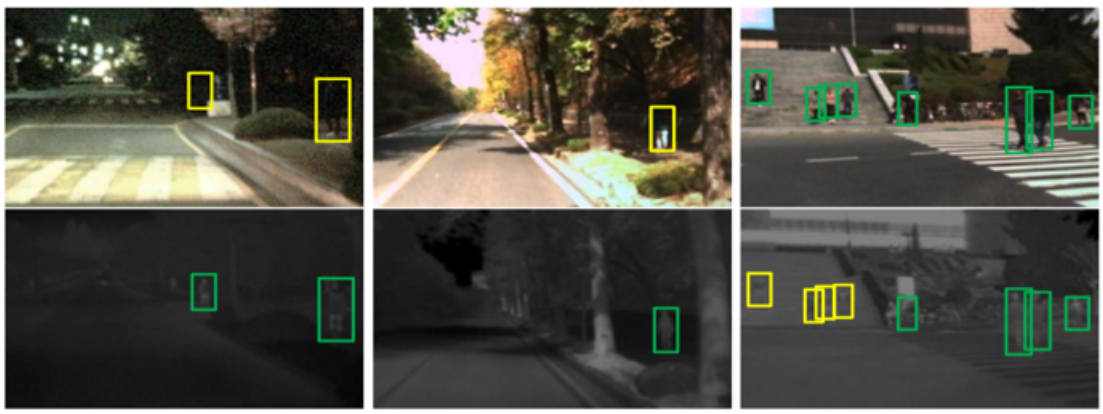

**Fig. 5.** Some samples of KAIST dataset.

As all the results are shown in Tab. 1, three indicator values obtained by our proposed model on three indicators are 7.14, 8.22 and 5.86, respectively. In terms of network structure selection, Fusion RPN+BF and Halfway Fusion are based on Faster R-CNN which lose some underlying details, resulting in poor detection performance on small targets. However, our proposed model integrates MSA and graph convolution operation to exploit context information, ultimately effectively improving the accuracy of small object detection. In terms of the use of modal information, MBNet has a lighting awareness feature combination module. During the day, its combined features mainly refer to visible light images, while at night, it mainly refers to infrared images. Bu, both visual and infrared information are adopted simultaneously in our proposed model. From the perspective of feature fusion, ACF adopts pixel-level feature fusion, while Halfway Fusion adopts intermediate-layer feature fusion. Compared to these models, our proposed model utilizes the cross-attention fusion mechanism to effectively exploit the inter-modality and intra-modality relationships. Compared to other multimodal fusion models, i.e., IATDNN+IASS [61], MSDS-RCNN [62], AR-CNN [63] and MBNet [59], our proposed model achieved the best performance on all three evaluation indicators.

We conducted extensive comparative experiments on the CVC-14 dataset. Similar to the KAIST dataset, we also compared the performance of various models under different lighting conditions. All experimental results are shown in Tab. 2. As all the results are shown in Tab. 2, the three indicator values of all models are significantly higher than those of the KAIST dataset. The changes in values further reflect the difficulty of pedestrian detection on CVC-14 dataset. To improve multi-scale pedestrian detection, MACF adopts the fast feature pyramid structure. But, the complexity of the model will significantly increase. In addition, this structure has a limited ability to model global context. However, due to the lack of spatial consistency in the fully convolutional networks used by MCIP and insufficient consideration of pixel-to-pixel relationships, it is difficult to effectively improve the performance of multi-scale pedestrian detection. Compared to AR-CNN, which only employs a weakly aligned cross-modal learning mechanism, the cross-attention fusion mechanism used in this paper achieves better performance for multimodal information fusion. Among all the comparative methods, i.e., MCIP [67], Halfway Fusion [60], UMSPD [68], AR-CNN [63] and MBNet [59], our proposed model achieves the highest performance.

To construct confusion matrices, the pedestrian (foreground) area and background area were defined as positive and negative data, respectively. We use confusion matrices to represent the detection performance of two categories: pedestrian and background, on two public datasets. All the results are shown in Fig. 6 and Fig. 7. Our proposed model achieves high performance in pedestrian and background recognition.

From the confusion matrix results on two datasets, it can be seen that the model proposed in this paper achieves good performance in pedestrian detection and background recognition accuracy. By fusing multimodal information, the robustness of the model is effectively improved. In addition, the model proposed in this article performs better for multi-scale object detection by fully mining contextual information.

| Methods | All | Day | Night |
|---|---|---|---|
| ACF [64] | 47.32 | 42.57 | 56.17 |
| Halfway Fusion [60] | 25.75 | 24.88 | 26.59 |
| Fusion RPN+BF [65] | 18.29 | 19.57 | 16.27 |
| IAF R-CNN [66] | 15.73 | 14.55 | 18.26 |
| IATDNN+IASS [61] | 14.95 | 14.67 | 15.72 |
| MSDS-RCNN [62] | 11.34 | 10.53 | 12.92 |
| AR-CNN [63] | 9.34 | 9.94 | 8.38 |
| MBNet [59] | 8.13 | 8.28 | 7.86 |
| Ours | 7.14 | 8.22 | 5.86 |

**Tab. 1.** Effectiveness of different deep models on KAIST dataset.

| Methods | All | Day | Night |
|---|---|---|---|
| MACF [64] | 60.1 | 61.3 | 48.2 |
| MCIP [67] | 47.3 | 49.3 | 43.8 |
| Halfway Fusion [60] | 37.0 | 38.1 | 34.4 |
| UMSPD [68] | 31.4 | 31.8 | 30.8 |
| AR-CNN [63] | 22.1 | 24.7 | 18.1 |
| MBNet [59] | 21.1 | 24.7 | 13.5 |
| Ours | 20.2 | 22.8 | 12.4 |

**Tab. 2.** Effectiveness of different deep models on CVC-14 dataset.

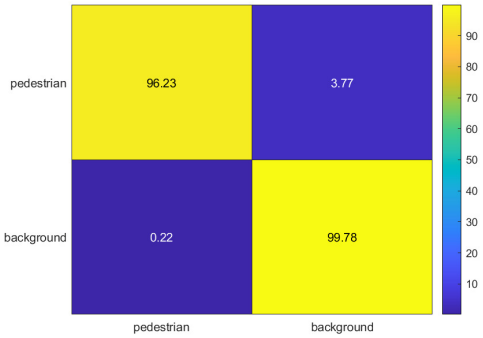


**Fig. 6.** Confusion matrices of KAIST datasets.

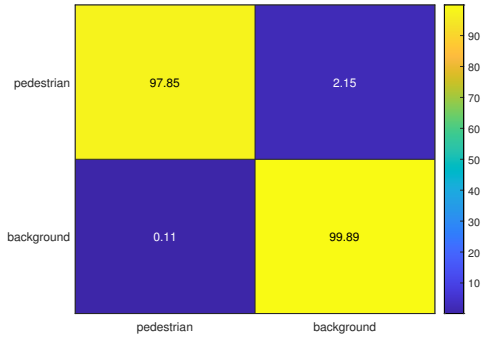


**Fig. 7.** Confusion matrices of CVC-14 datasets.

## 4.3 Effectiveness of Different Deep Models for Different Scale Pedestrian Detection

To verify the effectiveness of the model proposed in this article for multi-scale pedestrian detection, we set three scales of pedestrians, such near pedestrians, medium pedestrians, and far pedestrians. These three scales of pedestrian targets are all aimed at unobstructed pedestrians. To further quantify the effectiveness of pedestrian detection, we set the pixels of Near pedestrian targets to be higher than 115, Medium pedestrian targets to be higher than 45 pixels but lower than 115, and Far pedestrian targets to be lower than 45 pixels. The results of different models on the KAIST dataset are shown in Tab. 3.

As the results are shown in Tab. 3, we compare different models in terms of three kinds of scales. All the results confirm that our proposed model still obtains the highest performance. Compared to other models, our proposed method is more robust for pedestrian detection at different scales.

## 4.4 Effectiveness of Different Fusion Mechanisms

The use of multimodal information in pedestrian detection effectively improves the robustness of pedestrian detection. Different fusion mechanisms have different impacts on pedestrian detection. For this purpose, we compare the performance of different fusion mechanisms in this subsection.

All comparisons of different fusion mechanisms on the KAIST dataset are shown in Tab. 4. We compare different fusion mechanisms, i.e., Feature concatenation, Cross-attention [69], Leader-follower [70] and Joint cross-attention. The feature concatenation fusion mechanism shows the lowest performance. Compared to the feature concatenation mechanism, the cross-attention and leader-follower fusion mechanism can obtain higher performance. The cross-attention fusion mechanism can leverage the inter-modal relationships to extract the salient features across different modalities. The leader-follower fusion mechanism follows the model-level fusion strategy in which a two-layer LSTM network is used to capture the temporal dynamics. The joint cross-attention achieves the highest performance compared to other fusion mechanisms. In this paper, we adopt the joint cross-attention mechanism to fuse different modalities.

## 4.5 Effect of Different Modules on our Proposed Model

Due to the good real-time performance of the existing YOLO framework in object detection tasks, the performance of multi-scale object detection still needs further improvement. Therefore, this article embeds MSA and graph convolutional models in the YOLOv5 framework, and further improves the performance of the model for pedestrian detection by mining the contextual information of the object. Therefore, in this experiment, we further verify the impact of different modules on pedestrian detection performance.

| Methods | Near | Med. | Far |
|---|---|---|---|
| ACF [64] | 28.74 | 53.67 | 88.20 |
| Halfway Fusion [60] | 8.13 | 30.34 | 75.70 |
| Fusion RPN+BF [65] | 0.04 | 30.87 | 88.86 |
| IAF R-CNN [66] | 0.96 | 25.24 | 77.84 |
| IATDNN+IASS [61] | 0.04 | 28.55 | 83.42 |
| MSDS-RCNN [62] | 1.29 | 16.19 | 63.73 |
| AR-CNN [63] | 0 | 16.08 | 69.00 |
| MBNet [59] | 0.00 | 16.07 | 55.99 |
| Ours | 0.00 | 11.52 | 40.85 |

**Tab. 3.** Effectiveness of different deep models for different scale pedestrian detection on KAIST dataset.

| Fusion Module | All | Day | Night |
|---|---|---|---|
| Feature Concatenation | 10.05 | 12.56 | 9.52 |
| Cross-attention [69] | 8.26 | 10.25 | 6.28 |
| Leader-follower [70] | 8.82 | 11.34 | 6.82 |
| Joint cross-attention | 7.14 | 8.22 | 5.86 |

**Tab. 4.** Effectiveness of different fusion mechanisms on KAIST dataset.

| Models | All |
|---|---|
| YOLOv3 | 73.5 |
| YOLOv4 | 76.9 |
| YOLOv5tiny | 90.3 |
| YOLOv5 | 92.2 |
| YOLOv5+MSA | 93.3 |
| YOLOv5+MSA+Graph | 94.5 |

**Tab. 5.** Effectiveness of different models on KAIST dataset.

In our experiments, we adopt different YOLO frameworks, i.e., YOLOv3, YOLOv4, YOLOv5tiny and YOLOv5 as the backbone for feature extraction. In this experiment, we use precision as an evaluation indicator. As the results shown in Tab. 5, YOLOv5 obtains higher performance compared to other backbones. By introducing the MSA model, the precision of the model has been improved from 92.2% to 93.3%. Moreover, we embed the graph convolution operation, and the model precision has reached 94.5%. All the results confirm that the MSA and graph convolution can effectively improve the performance of pedestrian detection.

## 4.6 Effectiveness of Different Number Graph Convolution Layers

So, in our proposed model, we embed the graph convolution operation to mine the local contextual information. We conducted compared experiments to validate the effectiveness of different graph convolution layers on KAIST dataset. We use the YOLOv5+MSA as the backbone to extract pedestrian features and adopt precision as an evaluation indicator. All the results are shown in Fig. 8.

As the results are shown in Fig. 8, we change the number of graph convolution layers from 1 to 5. When the number of graph convolutional layers is set to 2, achieves the highest detection accuracy of 94.5%. However, as the number of graph convolutional layers increases, the detection accuracy of our proposed model decreases.

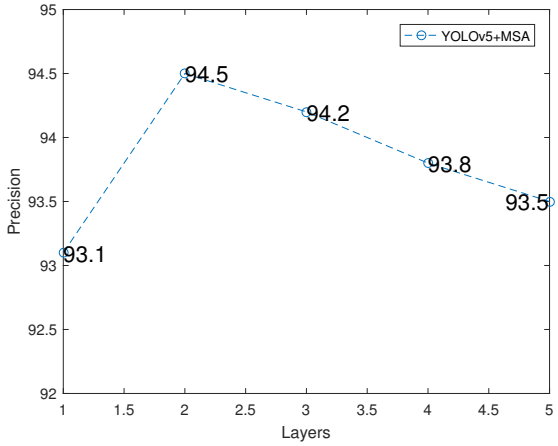


**Fig. 8.** The effectiveness of different number graph convolution layers.

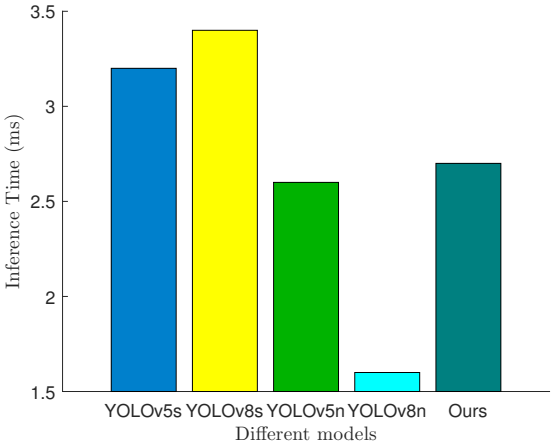


**Fig. 9.** The speed of different model on KAIST dataset.

For example, when the number of graph convolutional layers is 5, the recognition accuracy of our proposed model is 93.5%. In addition, adding graph convolutional layers will increase the parameters. So, we set the number of graph convolutional layers as 2 to achieve the highest detection accuracy without significantly increasing the model complexity.

### 4.7 Speed of Different Deep Models

The YLOLO series models outperform other deep models in object detection tasks. To evaluate the testing speed of different models, we compared the speed of different YOLO models with our proposed model. The YOLO models we use are mainly two versions of YOLOv5 and YOLOv8. We conduct experiments on KAIST dataset and all experiments are shown in Fig. 9. Among the models we compared, YOLOv8n achieved the fastest detection speed. When our proposed model uses YOLOv5n as a backbone network for feature extraction, the overall speed of the proposed model is slower than YOLOv5n. So by embedding the multi-head attention mechanism and graph convolution, the computational complexity of our proposed model is increased.

By adding graph convolution and multi-head self-attention layers, the computational complexity of the model will increase. However, in terms of the recognition accuracy of the model on the KAIST dataset, the detection accuracy of the model increased from 92.2% to 94.5% (as shown in Tab. 5), the model detection speed increased from 2.6 ms to 2.7 ms (as shown in Fig. 9). The detection accuracy of the model is significantly improved without significantly reducing the model speed. We believe that from the perspective of actual deployment, the comprehensive performance of our proposed model is better.

## 5. Conclusions and Discussions

In this paper, we propose a novel multimodal fusion framework for pedestrian detection based on the YOLOv5 framework. To improve the multi-scale pedestrian detection performance, we embed the MSA and the graph convolution modules to explore the contextual information. The MSA can exploit the global contextual information of pedestrian features. Moreover, the graph convolution module is embedded for local contextual information mining. At last, we adopt the joint cross-attention multimodal fusion mechanism to exploit the complementary relationship between different modalities. We can effectively mine intra-modality relationships and exploit multi-modal complementary simultaneously. We conduct extensive experiments on two multimodal pedestrian detection datasets. The experimental results confirm that our proposed model shows the highest performance and robustness compared to other multimodal fusion models in pedestrian detection tasks.

Future research work will mainly focus on the following two aspects: 1) Due to the increasingly complex environment of pedestrian detection, in addition to changes in scale, object occlusion, low resolution, etc., all of these issues will greatly affect the accuracy of pedestrian detection. A more efficient deep network structure needs to be designed to improve the model's ability to extract complex pedestrian features. 2) Although there are many multimodal datasets, not all modal data are easy to collect. The training of deep networks requires a large amount of data. Therefore, building large-scale multimodal datasets can effectively improve the robustness of models in complex scenarios.

## References

[1] GUPTA, A., ANPALAGAN, A., GUAN, L., et al. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, 2021, vol. 10, p. 1–20. DOI: 10.1016/j.array.2021.100057

[2] SAPONARA, S., ELHANASHI, A., GAGLIARDI, A. Real-time video fire/smoke detection based on CNN in antifire surveillance systems. *Journal of Real-Time Image Processing*, 2021, vol. 18, p. 889–900. DOI: 10.1007/s11554-020-01044-0

[3] KIM, K., KIM, S., SHCHUR, D. A UAS-based work zone safety monitoring system by integrating internal traffic control plan (ITCP) and automated object detection in game engine environment. *Automation in Construction*, 2021, vol. 128, p. 1–20. DOI: 10.1016/j.autcon.2021.103736

[4] GAO, F., WANG, C., LI, C. A combined object detection method with application to pedestrian detection. *IEEE Access*, 2020, vol. 8, p. 194457–194465. DOI: 10.1109/ACCESS.2020.3031005

[5] CHEN, L., LIN, S., LU, X., et al. Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021, vol. 22, no. 6, p. 3234–3246. DOI: 10.1109/TITS.2020.2993926

[6] ISLAM, M. M., KARIMODDINI, A. Pedestrian detection for autonomous cars: inference fusion of deep neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 2022, vol. 23, no. 12, p. 23358–23368. DOI: 10.1109/TITS.2022.3210186

[7] GIRSHICK, R., DONAHUE, J., DARRELL, T., et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Columbus (USA), 2014, p. 580–587. DOI: 10.1109/CVPR.2014.81

[8] ZHU, F., CHEN, X., GAO, X., et al. Constraint-weighted support vector ordinal regression to resist constraint noises. *Information Sciences*, 2023, vol. 649, p. 1–17. DOI: 10.1016/j.ins.2023.119644

[9] LI, J., LIANG, X., SHEN, S. M., et al. Scale-aware fast R-CNN for pedestrian detection. *IEEE Transactions on Multimedia*, 2017, vol. 20, no. 4, p. 985–996. DOI: 10.1109/tmm.2017.2759508

[10] YU, X., SI, Y., LI, L. Pedestrian detection based on improved Faster RCNN algorithm. In *2019 IEEE/CIC International Conference on Communications in China (ICCC)*. Chengdu (China), 2019, p. 346–351. DOI: 10.1109/ICCChina.2019.8855960

[11] ZHAI, S., SHANG, D., WANG, S., et al. DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion. *IEEE Access*, 2020, vol. 8, p. 24344–24357. DOI: 10.1109/ACCESS.2020.2971026

[12] JIANG, P., ERGU, D., LIU, F., et al. A review of Yolo algorithm developments. *Procedia Computer Science*, 2022, vol. 199, p. 1066–1073. DOI: 10.1016/j.procs.2022.01.135

[13] YU, W., XIANG, Z., JIANG, S., et al. YOLOv5-based dense small target detection algorithm for aerial images using DIOU-NMS. *Radioengineering*, 2024, vol. 33, no. 1, p. 12–22. DOI: 10.13164/re.2024.0012

[14] GUPTA, K., ASTHANA, A. Reducing the side-effects of oscillations in training of quantized YOLO networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Tucson (USA), 2024, p. 2452–2461. DOI: 10.1109/WACV57701.2024.00244

[15] LAN, W., DANG, J., WANG, Y., et al. Pedestrian detection based on YOLO network model. In *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*. Changchun (China), 2018, p. 1547–1551. DOI: 10.1109/ICMA.2018.8484698

[16] JIANG, C., REN, H., YE, X., et al. Object detection from UAV thermal infrared images and videos using YOLO models. *International Journal of Applied Earth Observation and Geoinformation*, 2022, vol. 112, p. 1–17. DOI: 10.1016/j.jag.2022.102912

[17] LI, Z., WANG, H., XUE, H., et al. Infrared image recognition method for pedestrian and vehicles based on improved YOLO. *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*, 2021, vol. 11763, p. 1422–1427. DOI: 10.1117/12.2587231

[18] HSU, W. Y., LIN, W. Y. Ratio-and-scale-aware YOLO for pedestrian detection. *IEEE Transactions on Image Processing*, 2020, vol. 30, p. 934–947. DOI: 10.1109/TIP.2020.3039574

[19] HSU, W. Y., LIN, W. Y. Adaptive fusion of multi-scale YOLO for pedestrian detection. *IEEE Access*, 2021, vol. 9, p. 110063–110073. DOI: 10.1109/ACCESS.2021.3102600

[20] HOU, Z., SUN, Y., GUO, H., et al. M-YOLO: An object detector based on global context information for infrared images. *Journal of Real-Time Image Processing*, 2022, vol. 19, no. 6, p. 1009–1022. DOI: 10.1007/s11554-022-01242-y

[21] XUE, P., CHEN, H., LI, Y., et al. Multiscale pedestrian detection with global local attention and multiscale receptive field context. *IET Computer Vision*, 2023, vol. 17, no. 1, p. 13–25. DOI: 10.1049/cvi2.12125

[22] XUE, Y., JU, Z., LI, Y., et al. MAF-YOLO: Multi-modal attention fusion based YOLO for pedestrian detection. *Infrared Physics & Technology*, 2021, vol. 118, p. 1–14. DOI: 10.1016/j.infrared.2021.103906

[23] DASGUPTA, K., DAS, A., DAS, S., et al. Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2022, vol. 23, no. 9, p. 15940–15950. DOI: 10.1109/TITS.2022.3146575

[24] KOLLURI, J., DAS, R. Intelligent multimodal pedestrian detection using hybrid metaheuristic optimization with deep learning model. *Image and Vision Computing*, 2023, vol. 131, p. 1–11. DOI: 10.1016/j.imavis.2023.104628

[25] LEE, W. Y., JOVANOR, L., PHILIPS, W. Cross-modality attention and multimodal fusion transformer for pedestrian detection. In *European Conference on Computer Vision*. Tel Aviv (Israel), 2022, p. 608–623. DOI: 10.1007/978-3-031-25072-9_41

[26] LI, G., LAI, W., QU, X. Pedestrian detection based on light perception fusion of visible and thermal images. *Optics & Laser Technology*, 2022, vol. 156, p. 1–14. DOI: 10.1016/j.optlastec.2022.108466

[27] TANG, L., MA, S., MA, X., et al. Research on image matching of improved sift algorithm based on stability factor and feature descriptor simplification. *Applied Sciences*, 2022, vol. 12, no. 17, p. 1–19. DOI: 10.3390/app12178448

[28] RANGANATHA, S., GOWRAMMA, Y. Eigen and HOG features based algorithm for human face tracking in different background challenging video sequences. *International Journal of Image, Graphics and Signal Processing*, 2022, vol. 10, no. 4, p. 70–83. DOI: 10.5815/ijigsp.2022.04.06

[29] ZHAI, Y., LIU, H. One class SVM model based on neural tangent kernel for anomaly detection task on small-scale data. *Journal of Intelligent & Fuzzy Systems*, 2022, vol. 43, no. 3, p. 2731–2746. DOI: 10.3233/JIFS-213088

[30] WEINBERG, G. Interference control in sliding window detection processes using a Bayesian approach. *Digital Signal Processing*, 2020, vol. 99, p. 1–12. DOI: 10.1016/j.dsp.2020.102658

[31] WANG, J., WU, R., YU, X. An algorithm of object detection based on regression learning for remote sensing images. *Journal of Physics: Conference Series*, 2021, vol. 1903, no. 1, p. 1–7. DOI: 10.1088/1742-6596/1903/1/012039

[32] ZHANG, L., LIN, L., LIANG, X., et al. Is faster R-CNN doing well for pedestrian detection? In *Computer Vision-ECCV 2016: 14th European Conference*. Amsterdam (The Netherlands), 2016, p. 443–457. DOI: 10.1007/978-3-319-46475-6_28

[33] LIU, W., ANGUELOV, D., ERHAN, D., et al. SSD: Single shot multibox detector. *Computer Vision-ECCV*. Amsterdam (The Netherlands), 2016, p. 21–37. DOI: 10.1007/978-3-319-46448-0_2

[34] JIANG, X., GAO, T., ZHU, Z., et al. Real-time face mask detection method based on YOLOv3. *Electronics*, 2021, vol. 10, no. 7, p. 1–17. DOI: 10.3390/electronics10070837

[35] BOCHKOVSKIY, A., WANG, C., LIAO, H. YOLOv4: Optimal speed and accuracy of object detection. *arXiv*, 2020, p. 1–17. DOI: 10.48550/arXiv.2004.10934

[36] GAI, Y., HE, W., ZHOU, Z. Pedestrian target tracking based on DeepSORT with YOLOv5. In *2nd International Conference on Computer Engineering and Intelligent Control (ICCEIC)*. Chongqing (China), 2021, p. 1–5. DOI: 10.1109/ICCEIC54227.2021.00008

[37] SUKKAR, M., KUMAR, D., SINDHA, J. Real-time pedestrians detection by YOLOv5. In *12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. Kharagpur (India), 2021, p. 1–6. DOI: 10.1109/ICCCNT51525.2021.9579808

[38] LV, H., YAN, H., LIU, K., et al. YOLOv5-AC: Attention mechanism-based lightweight YOLOv5 for track pedestrian detection. *Sensors*, 2022, vol. 22, no. 15, p. 1–25. DOI: 10.3390/s22155903

[39] KROTOSKY, S. J., TRIVEDI, M. M. On color-, infrared-, and multimodal-stereo approaches to pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 2007, vol. 8, no. 4, p. 619–629. DOI: 10.1109/TITS.2007.908722

[40] CHEN, Y. T., SHI, J., YE, Z., et al. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*. Cham (Switzerland), 2022, p. 139–158. DOI: 10.1007/978-3-031-20077-9_9

[41] CAO, Y., BIN, J., HAMARI, J., et al. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver (Canada), 2023, p. 403–411. DOI: 10.1109/CVPRW59228.2023.00046

[42] DRADRACH, A., KONERT, J., RUMINSKI, J. Multimodal camera for pedestrian detection with deep learning models. In *2023 IEEE International Conference on Industrial Technology (ICIT)*. Orlando (USA), 2023, p. 1–6. DOI: 10.1109/ICIT58465.2023.10143046

[43] WANG, C., LIU, Y., CHANG, F., et al. Pedestrian detection based on YOLOv3 multimodal data fusion. *Systems Science & Control Engineering*, 2022, vol. 10, no. 1, p. 832–845. DOI: 10.1080/21642583.2022.2129507

[44] CAO, Z., YANG, H., ZHAO, J., et al. Attention fusion for one-stage multispectral pedestrian detection. *Sensors*, 2021, vol. 21, no. 12, p. 1–17. DOI: 10.3390/s21124184

[45] DAS, A., DAS, S., SISTU, G., et al. Revisiting modality imbalance in multimodal pedestrian detection. In *2023 IEEE International Conference on Image Processing (ICIP)*. Kuala Lumpur (Malaysia), 2023, p. 1755–1759. DOI: 10.1109/ICIP49359.2023.10222711

[46] LI, Q., ZHANG, C., HU, Q., et al. Confidence-aware fusion using Dempster-Shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 2022, vol. 25, p. 3420–3431. DOI: 10.1109/TMM.2022.3160589

[47] WANG, Q., CHI, Y., SHEN, T., et al. Improving RGB-infrared pedestrian detection by reducing cross-modality redundancy. In *2022 IEEE International Conference on Image Processing (ICIP)*. Bordeaux (France), 2022, p. 526–530. DOI: 10.1109/ICIP46576.2022.9897682

[48] HAN, K., XIAO, A., WU, E., et al. Transformer in transformer. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS)*. 2021 p. 15908–15919. DOI: 10.5555/3540261.3541478

[49] CARION, N., MASSA, F., SYNNAEVE, G., et al. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Glasgow (UK), 2020, p. 213–229. DOI: 10.1007/978-3-030-58452-8_13

[50] ZHU, X., SU, W., LU, L., et al. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv*, 2020, p. 1–16. DOI: 10.48550/arXiv.2010.04159

[51] ROH, B., SHIN, J. W., SHIN, W., et al. Sparse DETR: Efficient end-to-end object detection with learnable sparsity. *arXiv*, 2021, p. 1–23. DOI: 10.48550/arXiv.2111.14330

[52] WANG, Y., ZHANG, X., YANG, T., et al. Anchor DETR: Query design for transformer-based object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver (Canada), 2022, p. 2567–2575. DOI: 10.48550/arXiv.2109.07107

[53] DAI, X., CHEN, Y., YANG, J., et al. Dynamic DETR: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal (Canada), 2021, p. 2988–2997. DOI: 10.1109/ICCV48922.2021.00298

[54] MENG, D., CHEN, X., FAN, Z., et al. Conditional DETR for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal (Canada), 2021, p. 3651–3660. DOI: 10.1109/ICCV48922.2021.00363

[55] ZHU, X., LYU, S., WANG, X., et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal (Canada), 2021, p. 2778–2788. DOI: 10.1109/ICCVW54120.2021.00312

[56] SANKAR, A., WU, Y., GOU, L., et al. Dysat: Deep neural representation learning on dynamic graphs via self-attention networks. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. Houston (USA), 2020, p. 519–527. DOI: 10.1145/3336191.3371845

[57] PRAVEEN, R. G., DE MELO, W. C., ULLAH, N., et al. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans (USA), 2022, p. 2486–2495. DOI: 10.1109/CVPRW56347.2022.00278

[58] CHOI, Y., KIM, N., HWANG, S., et al. KAIST multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 2018, vol. 19, no. 3, p. 934–948. DOI: 10.1109/TITS.2018.2791533

[59] ZHOU, K., CHEN, L., CAO, X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Computer Visionâ€“ECCV*. Glasgow (UK), 2020, p. 787–803. DOI: 10.1007/978-3-030-58523-5_46

[60] LIU, J., ZHANG, S., WANG, S., et al. Multispectral deep neural networks for pedestrian detection. *arXiv*, 2016, p. 1–13. DOI: 10.48550/arXiv.1611.02644

[61] GUAN, D., CAO, Y., YANG, J., et al. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 2019, vol. 50, p. 148–157. DOI: 10.1016/j.inffus.2018.11.017

[62] LI, C., SONG, D., TONG, R., et al. Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv*, 2018, p. 1–12. DOI: 10.48550/arXiv.1808.04818

[63] ZHANG, L., ZHU, X., CHEN, X., et al. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul (South Korea), 2019, p. 5127–5137. DOI: 10.1109/ICCV.2019.00523

[64] DOLLAR, P., APPEL, R., BELONGIE, S., et al. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, vol. 36, no. 8, p. 1532–1545. DOI: 10.1109/TPAMI.2014.2300479

[65] KONIG, D., ADAM, M., JARVERS, C., et al. Fully convolutional region proposal networks for multispectral person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Honolulu (USA), 2017, p. 49–56. DOI: 10.1109/CVPRW.2017.36

[66] LI, C., SONG, D., TONG, R., et al. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognition*, 2019, vol. 85, p. 161–171. DOI: 10.1016/j.patcog.2018.08.005

[67] CHOI, H., KIM, S., PARK, K., et al. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In *23rd International Conference on Pattern Recognition (ICPR)*. Cancun (Mexico), 2016, p. 621–626. DOI: 10.1109/ICPR.2016.7899703

[68] PARK, K., KIM, S., SOHN, K. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognition*, 2018, vol. 80, p. 143–155. DOI: 10.1016/j.patcog.2018.03.007

[69] PRAVEEN, R. G., GRANGER, E., CARDINAL, P. Cross attentional audio-visual fusion for dimensional emotion recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. Jodhpur (India), 2021, p. 1–8. DOI: 10.1109/FG52635.2021.9667055

[70] SCHONEVELD, L., OTHMANI, A., ABDELKAWY, H. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, 2021, vol. 146, p. 1–7. DOI: 10.1016/j.patrec.2021.03.007

## About the Authors . . .

**Yuan SHU** (corresponding author) was born in China, in March 1981. He received the B.S. degrees in Information Engineering from Department of Radio Engineering at Southeast University, China in 2003, and his M.S. degree in Computer Technology from School of Computer Science, Wuhan University, China in 2007. He is currently a Associate Professor with the Artificial Intelligence Technology Applications, Nanjing Vocational Institute of Railway Technology. His research interests mainly lie in data analysis and image processing for rail transit applications.

**Youren WANG** received the B.S. and M.S. degrees in Signal Circuits and Systems from Department of Radio Engineering at Southeast University, China, and his Ph.D. degree in Department of Automatic Control, Nanjing University of Aeronautics and Astronautics, China in 1996. As a Senior Visiting Scholar Studied in University of Manchester, UK in 2008. He is currently a Professor in the Department of Testing Engineering at the School of Automation, Nanjing University of Aeronautics and Astronautics, engaged in teaching and research in the fields of detection technology and computer measurement and control systems, electronics and information processing technology. His research interests include computer measurement and control technology and intelligent systems, sensor technology and signal processing, aviation comprehensive testing and online detection and health forecasting.

**Min ZHANG** was born in 1990 in Shan Dong province. In 2017, she received her Ph.D degree from Hohai University of Physical Oceanography. Her main research interests are artificial intelligence and high-performance computing. She has been working at the School of Intelligence Engineering, Nanjing Vocational Institute of Railway Technology since 2022.

**Jie YANG** was born in China in 1974. He received a Master's degree in Electronic Science and Technology from the University of Science and Technology of China in 2007. He is currently a professor in Nanjing Vocational Institute of Railway Technical. His research interests are digital signal processing and pattern recognition.

**Yi WANG** was born in Jiangsu, China in 1991, is currently pursuing a doctoral degree in the field of intelligent control. He obtained his master's degree from Nanjing University of Technology in 2016 and is currently employed at the School of Intelligent Engineering, Nanjing Railway Vocational and Technical College.

**Jun WANG** was born in Anhui Province, China. he graduated from South China University of Technology and attained a postgraduate degree and a Master of Engineering in 2009. His main research area is circuit design and reliability. Currently, he serves as an Associate Professor at the School of Intelligent Engineering of Nanjing Vocational Institute of Railway Technology.

**Yunbin ZHANG** was born in China in October 1992. He received B.S. degree in Electrical automation from Southwest Jiaotong University. He is currently a deputy director of Jinan Electrical Service Section of China Railway Jinan Bureau Group Co., LTD. He specializes in railway signaling.