



Article

Research on Multi-Modal Pedestrian Detection and Tracking Algorithm Based on Deep Learning

Rui Zhao ^{1,†} , Jutao Hao ^{2,†} and Huan Huo ^{1,*}

¹ Faculty of Engineering and IT, University of Technology Sydney, Ultimo 2007, Australia; rui.zhao-2@student.uts.edu.au

² School of Electric Information Engineering, Shanghai Dianji University, Shuihua Rd., Shanghai 201306, China; haojt@sdju.edu.cn

* Correspondence: huan.huo@uts.edu.au

† These authors contributed equally to this work.

Abstract: In the realm of intelligent transportation, pedestrian detection has witnessed significant advancements. However, it continues to grapple with challenging issues, notably the detection of pedestrians in complex lighting scenarios. Conventional visible light mode imaging is profoundly affected by varying lighting conditions. Under optimal daytime lighting, visibility is enhanced, leading to superior pedestrian detection outcomes. Conversely, under low-light conditions, visible light mode imaging falters due to the inadequate provision of pedestrian target information, resulting in a marked decline in detection efficacy. In this context, infrared light mode imaging emerges as a valuable supplement, bolstering pedestrian information provision. This paper delves into pedestrian detection and tracking algorithms within a multi-modal image framework grounded in deep learning methodologies. Leveraging the YOLOv4 algorithm as a foundation, augmented by a channel stack fusion module, a novel multi-modal pedestrian detection algorithm tailored for intelligent transportation is proposed. This algorithm capitalizes on the fusion of visible and infrared light mode image features to enhance pedestrian detection performance amidst complex road environments. Experimental findings demonstrate that compared to the Visible-YOLOv4 algorithm, renowned for its high performance, the proposed Double-YOLOv4-CSE algorithm exhibits a notable improvement, boasting a 5.0% accuracy rate enhancement and a 6.9% reduction in logarithmic average missing rate. This research's goal is to ensure that the algorithm can run smoothly even on a low configuration 1080 Ti GPU and to improve the algorithm's coverage at the application layer, making it affordable and practical for both urban and rural areas. This addresses the broader research problem within the scope of smart cities and remote ends with limited computational power.

Keywords: deep learning; multimodal; smart cities; future internet; pedestrian detection; feature fusion; YOLOv4 algorithm



Citation: Zhao, R.; Hao, J.; Huo, H. Research on Multi-Modal Pedestrian Detection and Tracking Algorithm Based on Deep Learning. *Future Internet* **2024**, *16*, 194. <https://doi.org/10.3390/fi16060194>

Academic Editors: Daniel Gutiérrez Reina and Ivan Serina

Received: 11 March 2024

Revised: 22 May 2024

Accepted: 29 May 2024

Published: 31 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In public transportation, pedestrians on the road as the most extensive group of participants have always been one of the most noteworthy objects in the field of safe traffic, and pedestrian detection on road traffic is also the most difficult target detection problem for intelligent transportation. The object detection task aims to find out the corresponding object category and coordinate position from the given input image or video. Because of its wide application, it has been given great importance both in academia and industry. However, in the real public road scenarios, many problems such as the different sizes of the target scale, the mutual occlusion between pedestrians and objects, and the serious interference of the lighting environment make target detection relatively complicated. Therefore, it is very difficult to develop an accurate and highly robust pedestrian detection system. Traditional pedestrian detection systems have not achieved satisfactory detection

results for road pedestrian detection tasks under complex conditions, and a big reason is that it is too one-sided to pay attention to the pedestrian information carried by visible light modal images and ignores the different performance and recording forms of the target under different information acquisition methods.

In the current situation of the development of the Internet of Things, more and more intelligent sensors continue to appear; these can provide a variety of modal information about pedestrians on the road and the surrounding environment. However, how to realize the combination of these multi-modal image features to enhance the task of pedestrian detection in special complex scenes has been widely discussed by researchers. R. Joseph (2016) proposed a YOLO target detection algorithm, which abandoned the mode of regional feature extraction + classification detection regression boundary box of the R-CNN series [1]. In Jin (2016), the detector uses extended ACFs (aggregated channel features) to detect pedestrian targets in complex scenes and then proposes an aggregated pedestrian detection method based on binocular stereo vision [2]. Lin (2017) introduced a top-down hierarchical pyramid structure on the basis of the original regional convolutional neural network, proposed the FPN (feature pyramid network) target detection algorithm, and realized the ability to construct high-level semantic feature information at different scales. It improves the ability of the network to capture the target [3]. Konig, D., (2017) proposed a new fusion region proposal network (RPN) for the pedestrian detection of multi-spectral video data.

Experiments verified the optimal convolution layer for the fusion of multi-spectral image information, but only the single feature layer fusion was considered. Hence, the detection of pedestrians with small targets is not accurate enough [4]. Qiu (2021) extracted infrared images processed by ICA with the SURF algorithm and fused them with the weighted fusion algorithm. In order to test the effectiveness of infrared cameras under the conditions of no occlusion, partial occlusion, and severe occlusion [5], Wang (2021) proposed the fusion of visual contrast mechanism and ROI to detect pedestrians in infrared images, but the effect was not good under complex dim lighting [6].

Through the summary of the above-mentioned literature on pedestrian detection using multi-mode fusion technology, it can be seen that the difficulty of road pedestrian target detection for intelligent transportation lies in the fact that no matter how good the algorithm model is, it cannot distinguish the pedestrian target in the image from the simple single-mode visible light image in an efficient and reliable way. Therefore, it is very necessary to introduce an infrared light mode image into the detection algorithm to improve the actual detection effect. Visible light mode images retain the color and texture information of objects better in the daytime, with good contrast and brightness, the pedestrian imaging effect is better, and it is easy to distinguish the background and foreground, so it is very suitable for normal road pedestrian detection tasks.

However, at night or even in the special case of uneven illumination or overexposure during the day, the imaging effect of visible light mode image will be very inferior, making the model unable to easily obtain any information about the target, which also results in the conventional pedestrian detection algorithm not being able to achieve the desired detection effect.

Considering the diverse economic conditions of various regions, including countries, cities, towns, and rural areas, the importance of accessible and user-friendly detection algorithms becomes evident. This study aims to develop a multi-modal pedestrian detection algorithm that performs efficiently on low-configuration GPUs, such as the 1080 Ti. Additionally, it seeks to extend the algorithm's applicability, ensuring affordability and practicality across both urban and rural settings. This research targets the broader issue within the realm of smart cities and remote areas with limited computational resources, striving to improve the accuracy and robustness of detection outcomes in varied environments.

The research framework depicted in Figure 1 below elucidates the structure of this study.

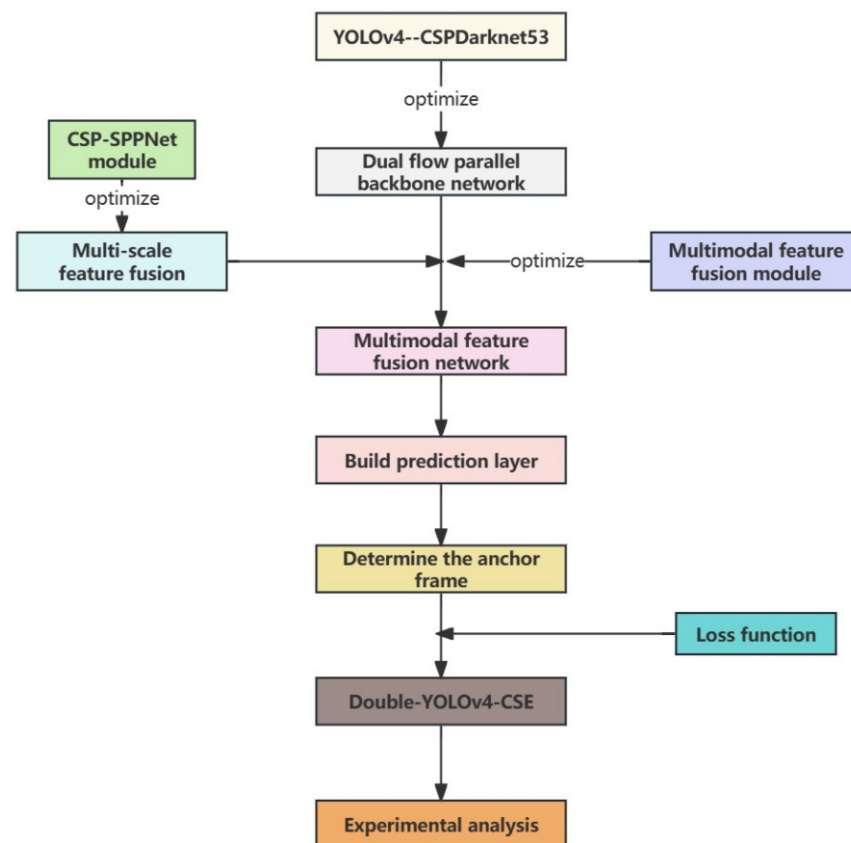


Figure 1. Research frame diagram.

The key contributions of this paper are presented as follows:

(1) A two-stream parallel feature extraction backbone network is devised based on YOLOv4, accompanied by the design of a channel stack fusion module to address the challenge of multimodal feature fusion.

(2) The process of multi-scale feature fusion layer prediction and enhancement is analyzed, culminating in the determination of the loss function and network calculation method during training to facilitate more effective learning.

(3) By employing a combination of qualitative and quantitative analysis methods, this paper verifies and analyzes the pedestrian detection performance of the algorithm through experimentation.

2. Build Multi-Modal Feature Fusion Network

2.1. Dual-Stream Parallel Backbone Network

In this paper, the CSPDarknet53 architecture is employed as the feature extraction backbone network in conjunction with the YOLOv4 algorithm. Through this setup, after three network blocks, each output characteristic of three different sizes are generated, thereby accomplishing multi-scale prediction. CSPDarknet53 introduces the CSPNet (cross-stage partial network) structure to address the drawbacks of the original Darknet53 network, such as high computational complexity and limited learning and feature extraction capabilities. By leveraging channel compression, CSPNet reduces the number of parameters in subsequent network modules by utilizing input feature matrices from two parallel branches. This optimization aids in minimizing video memory consumption, facilitating calculation, and enabling the propagation of segmented gradient information. The structure of CSPNet is depicted in Figure 2.

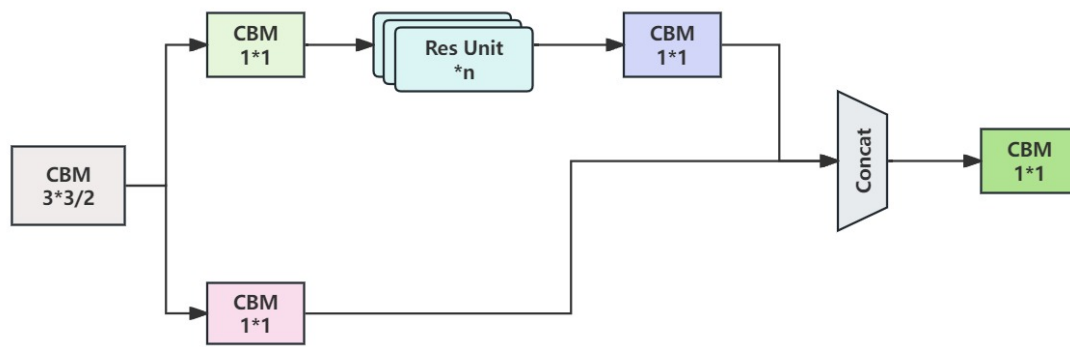


Figure 2. CSPNet structure.

Based on the above content, this chapter designs a dual-stream parallel mixed-mode feature extraction network based on CSPDarknet53 to realize the synchronous extraction of visible and infrared mode image features and facilitate the subsequent fusion of mixed mode features [7]. In the dual-flow parallel network, the branch of visible light mode feature extraction is CSPDarknet53-Visible, while the branch of infrared light mode feature extraction is CSPDarknet53-Infrared. Set the size of the input image of the network to $3 \times 512 \times 512$, then the backbone network will output the feature matrix of $256 \times 64 \times 64$, $512 \times 32 \times 32$ and $1024 \times 16 \times 16$, respectively, in the positions of CSPBlock3, CSPBlock4, and CSPBlock5. The outputs of the visible light mode are denoted as V_1 , V_2 , and V_3 , and the outputs of the infrared light mode are denoted as I_1 , I_2 , and I_3 .

The dual-flow parallel backbone extraction network architecture is shown in Figure 3 below.

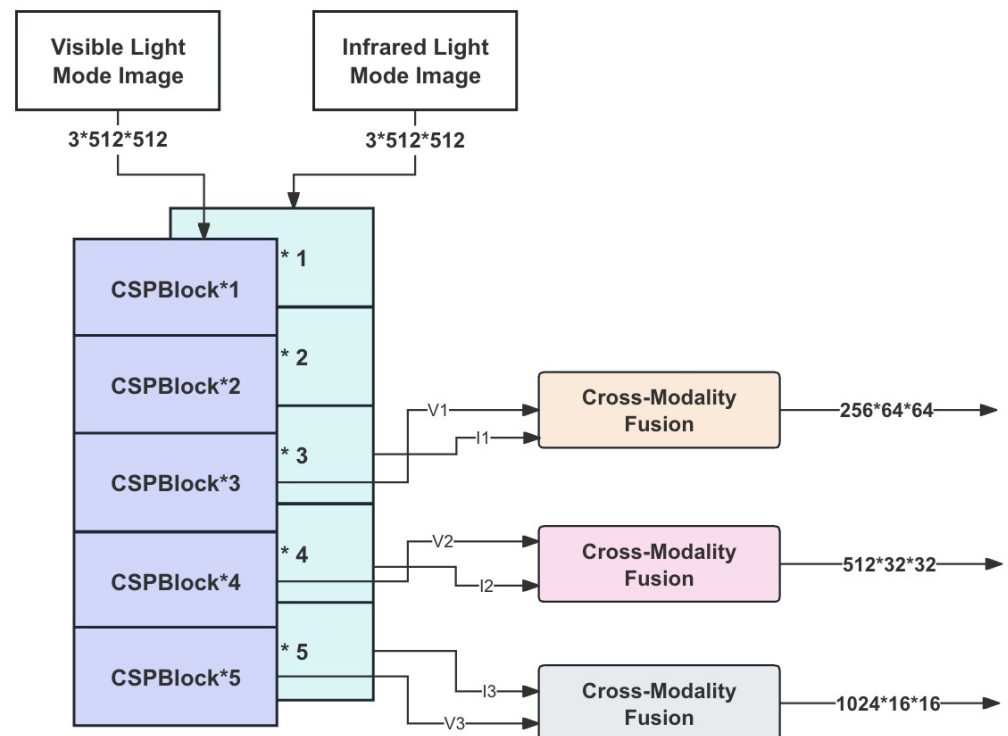


Figure 3. Dual-stream parallel backbone feature extraction network architecture.

2.2. Multi-Modal Feature Fusion Module

To facilitate the fusion of visible and infrared mode feature information, this paper introduces the channel stack fusion mode feature fusion scheme based on the aforementioned dual-stream parallel feature extraction backbone network. Ref. [8] In this scheme, the feature matrices from the two different modes are first stacked in depth, followed by the

utilization of a channel attention module to re-weight each stacked channel. Subsequently, a CBM (convolutional + batch normalization + mish) layer is employed to reduce the dimensionality of the fused feature matrix. It is worth noting that the CBM layer serves as a fundamental component in the YOLOv4 network architecture and is primarily responsible for feature extraction and transformation. The structure of the fusion module utilizing channel overlay is depicted in Figure 4 below.

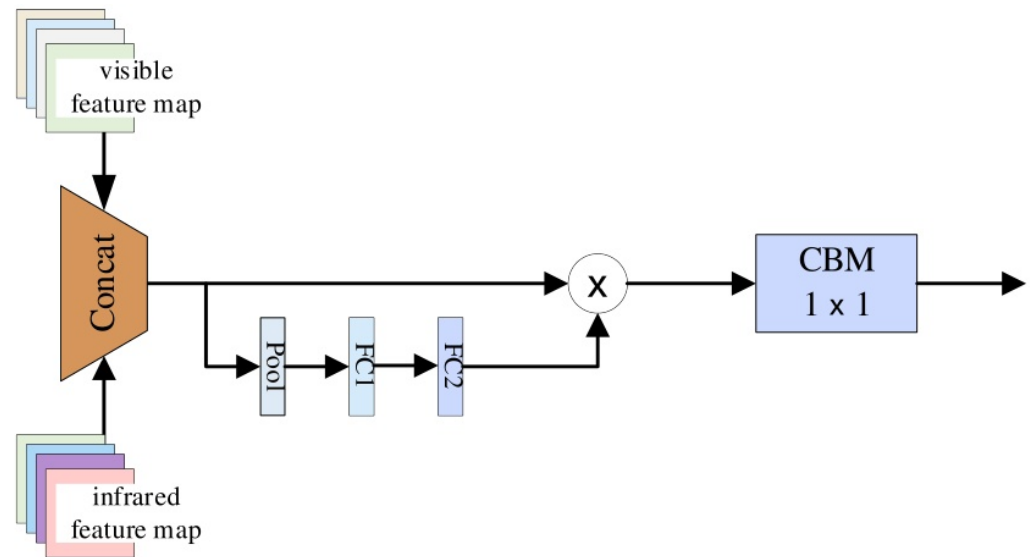


Figure 4. Channel stack fusion module.

2.3. Multi-Scale Feature Fusion Network

The feature pyramid network (FPN), built upon the original multi-scale feature extraction framework, introduces a bottom-up fusion branch to enhance the fusion of deep feature maps with shallow feature matrices. This integration enables the shallow detection branch to access more abstract feature semantic information, thereby significantly boosting the network's multi-scale target detection performance. Ref. [9] Although FPN indirectly integrates shallow semantic features through P3, further enhancements are required as the multi-scale feature fusion effect of FPN can still be improved. This is due to the elongated information flow route and the presence of multiple convolution operations within it.

In the YOLOv4 target detection algorithm proposed in this paper, a path aggregation network (PANet) is introduced to address these limitations. PANet facilitates faster information fusion by introducing bottom-up bypass connections between the FPN's P1 and P3 layers, thereby shortening the information path between low-level and high-level features. This enriches the network's feature representation capabilities and improves the accuracy of information storage. The PANet multi-scale feature fusion network structure, an enhancement built upon FPN, is depicted in Figure 5.

The hybrid mode multi-scale pedestrian detection network, proposed in this paper and based on YOLOv4, enhances the spatial pyramid pooling network (SPPNet) utilized in YOLOv4 by integrating it with the CSPNet structure. The SPP module employs multiple pooling layers to generate features of varying abstraction levels, culminating in the fusion and output of a fixed-size matrix. This significantly enhances the perception of deep-level feature maps [10]. In conjunction with the CSPNet, the SPP module in this paper amplifies the gradient propagation path, thereby augmenting the network's learning capabilities. This combination leverages the strengths of both SPPNet and CSPNet, enabling the retention of SPPNet's capacity to process multi-scale features while addressing gradient disappearance issues through CSPNet, consequently enhancing the network's learning capabilities.

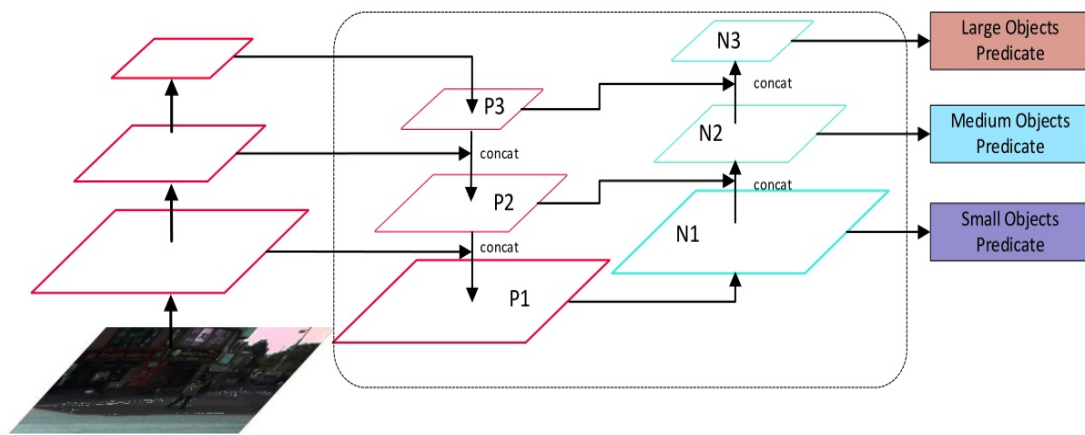


Figure 5. Improved PANet multi-scale feature fusion network based on FPN.

The hybrid mode multi-scale pedestrian detection network presented herein is adept at tackling challenges inherent in pedestrian detection tasks, such as scale changes, variations in pose, and complex background scenarios. Consequently, it elevates the accuracy and robustness of detection outcomes. Figure 6 illustrates the CSP-SPPNet structure, showcasing the integration of CSPNet and SPPNet components.

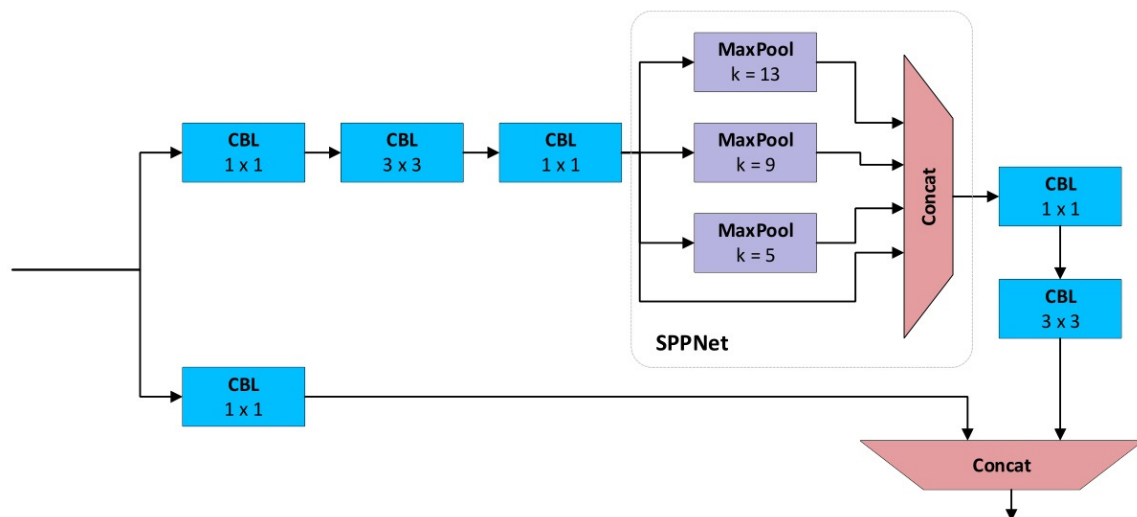


Figure 6. CSP-SPPNet module (CBL—the smallest component within the YOLOv3 network architecture, comprises three components: convolution (Conv), batch normalization (BN), and the leaky ReLU activation function).

3. Prediction Layer

3.1. Prediction Layer Design

In this paper, the YOLOv4 algorithm serves as the backbone of the proposed pedestrian detection framework. It partitions the input feature map into a grid of $S \times S$ at different scales, with each grid responsible for generating B bounding boxes. These bounding boxes are constrained by anchor boxes, which encode the geometric characteristics of the predicted targets, thus expediting the convergence rate of the prediction module and enhancing detection accuracy.

Consequently, the primary role of the prediction layer is to regress B bounding box parameters for each grid on the feature map. These parameters include the coordinates (t_{yx}) of the center point relative to the upper-left corner, as well as the width (t_w) and height (t_h) parameters of the bounding box. Additionally, the prediction layer is tasked with

outputting confidence scores (conf) indicating the likelihood of a bounding box containing the target, along with scores for various target classes $[c_1, c_2, \dots, c_n]$.

Therefore, the prediction layer generates vectors of size $B \times (4 + 1 + \text{num_classes})$ for each grid, resulting in an $S \times S \times B \times (4 + 1 + \text{num_classes})$ matrix for the entire input feature map. The process of the prediction layer is illustrated in Figure 7 below.

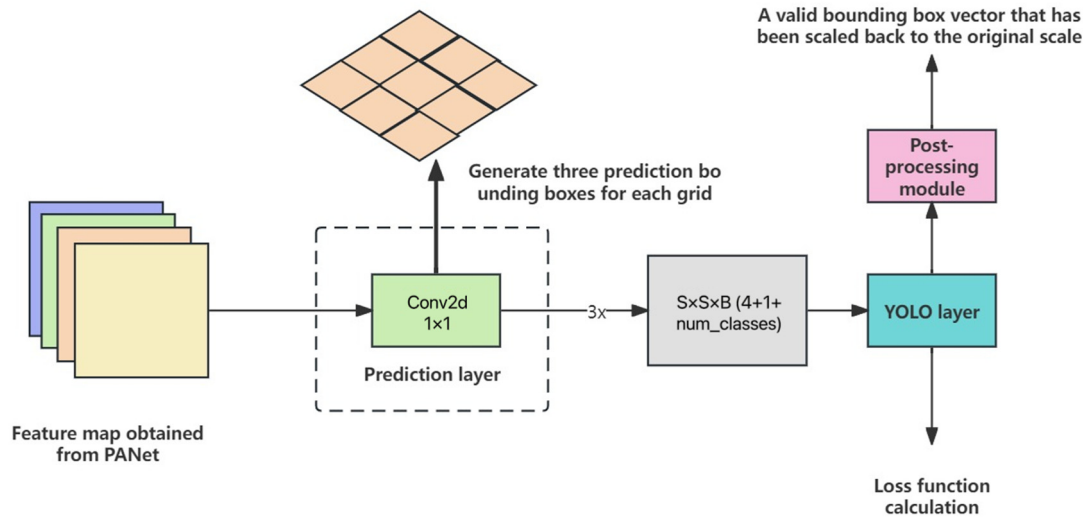


Figure 7. Processing flow of prediction layer.

To obtain the relative center position and width and height values of the bounding box in the entire input image, the following formula is also needed to achieve the correct mapping [11]. The process is completed by the YOLO layer after the prediction layer. When using the Sigmoid activation function in the YOLO layer to the regression bounding box, the values of t_x and t_y can be kept between $[0, 1]$, which can effectively ensure that the target center is in the center of the grid unit performing the prediction and prevent the occurrence of deviation.

In addition, the confidence and category score in the prediction vector corresponding to the output of each grid anchor box are also noteworthy parameters when detecting the output of the YOLOv4 algorithm based on this paper. For confidence, on the one hand, it indicates whether there is a prospect target in the current prediction box [12]. On the other hand, it also reflects the overlap degree between the predicted target boundary box and the marked real boundary box, that is, IoU. Therefore, the expected calculation expression of confidence is as follows:

$$C_i^j = P_r(object) \times IoU_{pred}^{truth} \quad (1)$$

In the formula, C_i^j is the JTH predicted boundary box confidence for the i -th grid, $P_r(object)$ predicts the probability of the existence of an object in the bounding box, and IoU_{pred}^{truth} is the intersection ratio between the predicted boundary box and the marked actual boundary box.

3.2. Determine the Anchor Frame

To obtain anchor frame data that effectively represents the dataset, traditional methods like faster R-CNN often rely on the empirical judgment of researchers. In YOLOv2 and its subsequent iterations, researchers commonly utilize the k-means algorithm to determine suitable anchor frame data by clustering the centers of real bounding boxes from annotated information in the dataset. However, the conventional k-means algorithm, which relies on Euclidean distance, may not be ideal for clustering scenarios involving anchor frames. Hence, it becomes imperative to employ a suitable distance metric.

The residual cross-merge area ratio (RCMAR) offers a robust measure of similarity between the shapes and sizes of two bounding boxes. A smaller output value signifies a higher similarity between two bounding boxes, effectively addressing the issue.

Therefore, this article adopts RCMAR as the measurement method, with the expression as follows:

$$\text{Dist}(w_1, h_1, w_2, h_2) = 1 - \frac{\min(w_1, w_2) \cdot \min(h_1, h_2)}{w_1 \cdot h_1 + w_2 \cdot h_2 - \min(w_1, w_2) \cdot \min(h_1, h_2)} \quad (2)$$

Based on the clustering results obtained by k-means using the residual intersection area ratio measurement method, this paper introduces genetic algorithms to further optimize and obtain anchor frame data that best represents the dataset. This optimization aims to expedite the convergence process of the network and enhance the algorithm's robustness [13].

In this paper, the fitness function calculation method involves dividing the width and height of each bounding box in the dataset by the width and height of the anchor box combination. Then, the Maximum Matching Box Ratio (MMBR) is employed to compute the matching score between the current annotated bounding box and anchor box. Finally, the matching scores exceeding a specified threshold (default value is 0.25) are aggregated to calculate the average score, which reflects the overall fitness of the anchor box. The fitness function calculation formula and MMBR calculation formula proposed in this paper are as follows:

$$f(wh, anchors, thr) = \frac{1}{N} \sum_{i=1}^N \delta(\text{MMBR}(w_i, h_i, anchors), thr) \quad (3)$$

$$\delta(x, thr) = \begin{cases} x, & \text{if } x > thr \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

$$\text{MMBR}(w, h, anchors) = \max\left\{\min\left[\gamma\left(\frac{w}{\text{anchor}_{w_1}}\right), \gamma\left(\frac{h}{\text{anchor}_{h_1}}\right)\right], \dots, \min\left[\gamma\left(\frac{w}{\text{anchor}_{w_n}}\right), \gamma\left(\frac{h}{\text{anchor}_{h_n}}\right)\right]\right\} \quad (5)$$

In the formula, h, w is the height and width information of all annotated boundary frames of the training set, anchors are the anchoring frame combination obtained by the genetic algorithm, n is the total number of annotated boundary frames, thr is the threshold value calculated for MMBR, w_i is the width of the i -th annotated boundary frame, h_i is the height of the i -th annotated boundary frame, anchor w_i is the width of the i -th prior frame in the anchor frame combination, and anchor h_i is the height of the i -th prior frame in the anchor frame combination.

4. Loss Function

The proposed detection algorithms involved in the loss function mainly includes three aspects: boundary box loss, loss of confidence, and classification loss. The confidence level and classification loss calculation formula is:

$$L_{\text{conf}(o,c)} = -\frac{\sum_i (o_i \ln(\hat{c}_i) + (1 - o_i) \ln(1 - \hat{c}_i))}{N} \quad (6)$$

$$L_{\text{cls}}(O, C) = -\frac{\sum_{i \in \text{pos}} \sum_{j \in \text{cla}} (O_{ij} \ln(\hat{C}_{ij}) + (1 - O_{ij}) \ln(1 - \hat{C}_{ij}))}{N_{\text{pos}}} \quad (7)$$

For the calculation of positioning loss, CIoU (Complete IoU) Loss is used in this paper instead of MSE [14]. The advantage of this kind of loss function is that it not only takes into account the influence of the overlapping area between the predicted boundary frames on

the loss calculation but also takes into account the influence of the center point distance and aspect ratio. The expression of the function is as follows:

$$CIoULoss = IoU - (\frac{\rho^2(b, b^{gt})}{c^2} + \alpha v) \quad (8)$$

$$v = \frac{4}{\pi^2} (\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h})^2 \quad (9)$$

$$\alpha = \frac{v}{(1 - IoU) + v} \quad (10)$$

In the formula, $\rho(b, b^{gt})$ is the center distance between the predicted boundary box and the real boundary box; c is the minimum diagonal distance between the smallest external frame between the two boundary boxes; α and v are the balance coefficient and the aspect ratio coefficient, respectively; w^{gt} is the width of the marked boundary box; w is the width of the predicted boundary box; h^{gt} is the height of the marked boundary box; and h is the height of the predicted boundary box.

5. Experimental Analysis

5.1. Experimental Design

To verify the detection performance of the algorithm proposed in this paper, the actual effect of pedestrian detection based on the mixed modes of visible light and infrared light is tested and compared with the traditional single-mode detection algorithm based on visible light only in order to prove the effectiveness of the mixed mode detection algorithm proposed in this paper [15]. Therefore, this chapter evaluates the actual performance of the algorithm by quantitative analysis and qualitative analysis. Among them, quantitative analysis mainly obtains the most objective result through the above evaluation indicators, while qualitative analysis artificially makes subjective judgments by observing the actual pedestrian detection effect under different road scenes.

5.2. Experimental Configuration

The experiment's code was implemented using the PyTorch deep learning framework. Before training, the network model loads pre-trained weight parameters from the COCO dataset, this transfer learning strategy greatly reduced the training time [16].

In the actual experiments, the training iteration epochs are set to 50, with a training image batch size of 8. The Adam optimization algorithm is employed, with an initial learning rate set to 0.0013. Additionally, the cosine vector annealing strategy is utilized for learning rate decay, with a weight decay coefficient of 0.0005. The first three rounds of training serve as a warm-up period.

Multi-scale training is adopted to enhance the network's robustness during training. During verification, images are scaled to 512 by 512 pixels. The experimental equipment parameters are detailed in Table 1 below.

Table 1. Experimental equipment parameters.

CPU Configuration	Model Number	Xeon Silver 4210R
GPU Configuration	Internal memory	192 GB
	Model number	NVIDIA GeForce GTX 1080Ti
	Core frequency	1582 MHz
	Video memory frequency	11,000 MHz
	Video memory capacity	11 GB
	Video memory width	352 bit
	CUDA Core	3584

During the experiment, the input image size was set to 512×512 . The k-means clustering algorithm and genetic algorithm were utilized to obtain nine types of anchor frame data of different scales based on the KAIST dataset, as illustrated in Table 2 below [17].

Table 2. Related data of nine kinds of anchor frames used in the experiment.

Prediction Branch	Output Feature Map Size	Predict Prior Box Size	Output the Predicted Bounding Box Number
Large target prediction layer	16×16	$(17 \times 35); (18 \times 45); (24 \times 44)$	$3 \times 16 \times 16$
Medium target prediction layer	32×32	$(22 \times 57); (28 \times 63); (33 \times 77)$	$3 \times 32 \times 32$
Small target prediction layer	64×64	$(39 \times 95); (47 \times 126); (63 \times 170)$	$3 \times 64 \times 64$

The KAIST dataset is a large dataset used for pedestrian detection tasks. It comprises 95,328 images, each containing a visible image and a corresponding long-wave infrared image, totaling 103,128 intensive annotations. It serves as a widely used benchmark in the field of multi-spectral pedestrian detection, providing abundant resources for the research and development of pedestrian detection algorithms.

5.3. Analysis of Results

In this experiment, the proposed multi-modal pedestrian detection algorithm based on YOLOv4 using the channel stack fusion module is called Double-YOLOv4-CSE, and the common pedestrian detection algorithm based on single visible light mode is called visible-YOLOv4. All the above-mentioned detection algorithm models were trained, verified, and tested on the cleaned KAIST dataset. Through the analysis and comparison of the samples predicted by the experimental model, it was found that the detection method using the mixed mode feature fusion method proposed in this chapter had better detection effects, as shown in Figure 8 below.

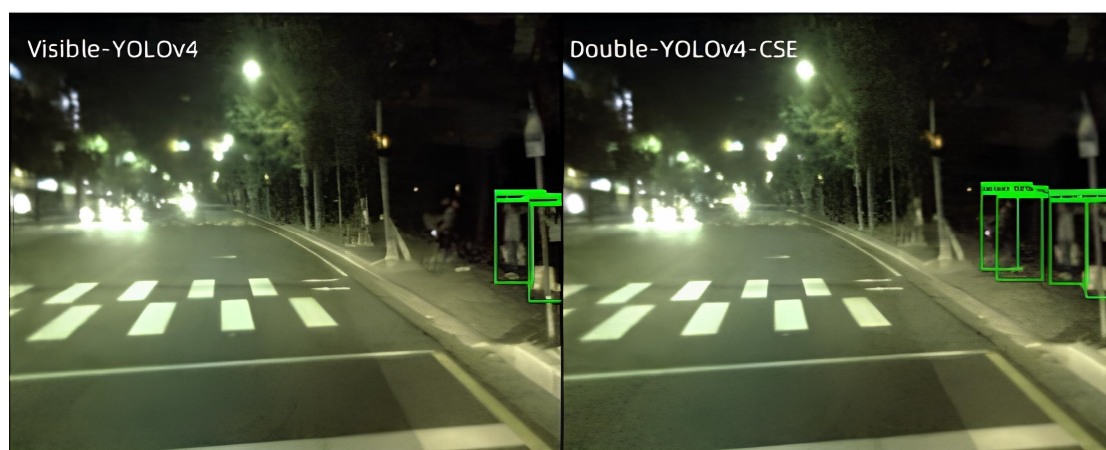


Figure 8. Comparison of traditional single-mode pedestrian detection algorithm and multi-mode pedestrian detection algorithm.

From this, it can be seen that the Double-YOLOv4-CSE proposed in this paper can more accurately detect dark pedestrians on the roadside in low-light scenes such as night and can effectively improve the ability to capture dark pedestrians.

At the same time, to further verify the effect of introducing multi-modal and corresponding feature fusion methods on improving pedestrian detection performance in the whole scene. This algorithm was compared with the logarithmic average miss rate and average accuracy obtained by the current mainstream and mature multi-modal pedestrian detection algorithms ACF + T + THOG [18], Fusion-RPN [4], GFD-SSD [19], and AR-CNN [20] on the KAIST dataset. The AR-CNN algorithm serves as a mechanism for amalgamating RGB images and infrared images. It leverages the global feature maps

derived from both modalities to identify candidate boxes. By expanding the coverage area of candidate boxes identified by the region proposal network (RPN), feature maps corresponding to individual modalities are extracted. Employing the infrared mode as the reference, the algorithm predicts the relative offset of the RGB mode. Subsequently, alignment of the two feature maps is performed based on this offset.

The default IoU threshold was 0.5, and the experimental results are shown in Table 3 below.

Table 3. Comparison of detection performance of pedestrian detection algorithms.

Algorithm	Average Accuracy %	Logarithmic Average Miss Rate %
Double-YOLOv4-CSE	89.7	23.1
Visible-YOLOv4	84.7	30.0
ACF + T + THOG	53.2	54.8
Fusion-RPN	75.3	29.7
GFD-SSD	-	28.0
AR-CNN	83.1	25.0

From the data presented in the table, it is evident that the proposed Double-YOLOv4-CSE algorithm exhibits superior detection performance compared to other multimodal pedestrian detection algorithms. Compared with AR-CNN algorithm, the proposed model has certain advantages in average accuracy and log-average omission rate. Even when compared with the excellent Visible-YOLOv4 algorithm, the proposed Double-YOLOv4-CSE algorithm achieves a 5.0% improvement in accuracy and a 6.9% reduction in the logarithmic average missing rate. These results strongly indicate the outstanding performance of the Double-YOLOv4-CSE algorithm in pedestrian detection and tracking tasks.

The core concept of YOLO algorithms is to treat object detection as a regression problem, achieving fast and accurate detection by directly predicting bounding boxes and object categories. This fundamental principle remains consistent across different versions of the YOLO algorithm. Consequently, strategies such as multimodal information fusion and contextual information enhancement employed in the Double-YOLOv4-CSE algorithm can also be implemented in other versions of the YOLO algorithm.

Therefore, the methods developed in this research are also suitable as embeddable techniques. The approach is considered as an extendable module that can be applied to various state-of-the-art (SOTA) methods. As shown in Table 4 above, integrating the multimodal approach into advanced algorithms like YOLOv6 and YOLOv8 [21] results in improved accuracy, demonstrating the potential of this method to enhance existing advanced algorithms.

Table 4. Comparison of detection performance of other YOLO methods.

Method	Double-YOLOv4-CSE	Average Accuracy %
Visible-YOLOv4	-	84.7
Visible-YOLOv4	✓	89.7
YOLOv6	-	90.2
YOLOv6	✓	90.6
YOLOv8	-	91.1
YOLOv8	✓	91.3

Furthermore, several pedestrian detection methods currently recognized as state-of-the-art (SOTA) were selected and compared with the proposed Double-YOLOv4-CSE method:

With the iterative upgrades of YOLO versions, including YOLOv6 to YOLOv8, resource consumption and computational requirements increase. Despite this, YOLOv4 offers a commendable balance between performance and cost-effectiveness, requiring only hardware facilities such as the NVIDIA GeForce GTX 1080 Ti for implementation.

Given the varying economic strengths of different countries, cities, towns, and rural areas, the accessibility and ease of use of detection algorithms are particularly important. The proposed Double-YOLOv4-CSE algorithm offers the widest coverage and is better suited for smart cities with diverse economic development situations. Additionally, its deployment is comparatively straightforward.

Overall, whether integrating the proposed method as an embeddable module into YOLOv6 or YOLOv8 or comparing it directly with methods such as Faster R-CNN [22], RetinaNet [23], CenterNet [24], or HRNet [25] (as shown in Table 5 above), the experimental results are consistently favorable. This indicates the versatility and effectiveness of the proposed method within various YOLO frameworks as well as among other state-of-the-art (SOTA) techniques.

Table 5. Comparison of detection performance of SOTA methods.

Method Name	Core Algorithm	Dataset	Average Accuracy %
Double-YOLOv4-CSE	YOLOv4 + Channel Stack Fusion	KAIST	89.7
Faster R-CNN	Region-based Convolutional Neural Networks	KAIST	79.3
Retina Net	Focal Loss for Dense Object Detection	KAIST	76.5
Center Net	Objects as Points	KAIST	78.1
HRNet	High-Resolution Network	KAIST	71.2

6. Conclusions

In summary, this paper has proposed a novel multi-modal pedestrian detection algorithm based on YOLOv4, leveraging the complementary features between infrared mode images and visible light mode. The key contributions include designing a dual-stream parallel feature extraction backbone network, analyzing improvements in multi-scale feature fusion and prediction layer processing, determining the loss function calculation method, and introducing a fusion module using channel stacking.

Experimental analysis has validated the effectiveness of the proposed algorithm, particularly in low-light scenarios such as nighttime, showcasing superior performance compared to existing pedestrian detection algorithms. However, it is worth noting that the feature information inherent in visible light mode images differs significantly from that in infrared mode images in terms of form, quantity, and distribution.

Moving forward, future research will delve deeper into this topic, particularly exploring the utilization of more targeted network architectures for feature extraction in infrared mode images. This endeavor aims to further enhance the performance and robustness of multi-modal pedestrian detection algorithms.

7. Future Directions

Looking ahead, this research can further enhance the accuracy and efficiency of multi-modal pedestrian detection and tracking algorithms, especially in dynamic and changing traffic environments. Additionally, research can extend to vehicle detection and other ITS components, like traffic flow analysis and accident prevention, for more comprehensive traffic management and safety. By integrating a wider array of data sources and utilizing advanced deep learning models, such as graph convolutional networks (GCNs) and spatiotemporal networks, future studies are poised to make significant advancements in all ITS aspects, supporting safer and more efficient urban transportation systems, marking a significant step forward in the development of smart urban environments and the pioneering technologies shaping the future of the Internet.

This expansion not only broadens the application range of current technologies but also injects new vitality and direction into the evolving landscape of the future internet and smart city infrastructure, fundamentally augmenting urban safety measures, transportation efficacy, and environmental sustainability. Moreover, the prospective exploration into the synergistic amalgamation of pedestrian detection technologies with other pivotal smart city

innovations—ranging from autonomously operating vehicles to intelligent street lighting systems and sophisticated traffic flow management mechanisms—holds the promise of markedly accelerating smart urban development, thereby positioning these integrative technological advancements at the forefront of shaping the future internet paradigm.

Author Contributions: Writing—original draft, R.Z.; data curation, R.Z. and J.H.; conceptualization, R.Z.; formal analysis, R.Z. and J.H.; methodology, R.Z. and J.H.; software, R.Z. and J.H.; supervision, H.H. and R.Z.; validation, R.Z. and J.H.; visualization, H.H.; writing—review and editing, H.H.; project Administration, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The KAIST dataset is open and available at: <https://github.com/SoonminHwang/rgbt-ped-detection/tree/master>, accessed on 1 April 2024.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

Acronyms utilized within the text are expanded and spelled out in full below for clarity and comprehension.

YOLOv4	You Only Look Once version 4
R-CNN	Region-based Convolutional Neural Networks
ACF	Aggregated Channel Features
FPN	Feature Pyramid Network
RPN	Region Proposal Network
ROI	Region of Interest
CSPNet	Cross-Stage Partial Network
CBM	Convolutional + Batch Normalization + Mish
CBL	Convolutional + Batch Normalization +LeakyRelu
PANet	Path Aggregation Network
SPPNet	Spatial Pyramid Pooling Network
IoU	Intersection over Union
RCMAR	Residual Cross-Merge Area Ratio
MMBR	Maximum Minimum Boundary Ratio
CIoU	Complete IoU Loss
ACF + T + THOG	Aggregated Channel Features Truncated Haar features with Oriented Gradients
GFD-SSD	Gated fusion double SSD for multispectral pedestrian detection
	Single Shot Multi-Box Detector
AR-CNN	Region Feature Alignment
	Convolutional Neural Networks
Fast R-CNN	Faster Region-Convolutional Neural Network

References

1. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
2. Jin, Z.; Zhao, M.; Zhang, R.; Wei, J. Optimized aggregated channel features pedestrian detection algorithm based on binocular vision. *Tianjin Daxue Xuebao (Ziran Kexue Yu Gongcheng Jishu Ban)/J. Tianjin Univ. Sci. Technol.* **2016**, *49*, 1225–1230.
3. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
4. Konig, D.; Adam, M.; Jarvers, C.; Layher, G.; Neumann, H.; Teutsch, M. Fully convolutional region proposal networks for multispectral person detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 49–56.
5. Gen, Q.; Li, W.; Kai, C. Research on high pressure vessel detection technology based on infrared image fusion algorithms. *Chin. Meas. Test* **2021**, *47*, 7.

6. Yüping, W.; Yi, Z. Pedestrian detection in infrared images using ROI fusion and human visual mechanism. *Chin. Meas. Test* **2021**, *47*, 87–93.
7. Jain, D.K.; Zhao, X.; González-Almagro, G.; Gan, C.; Kotecha, K. Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes. *Inf. Fusion* **2023**, *95*, 401–414. [[CrossRef](#)]
8. Wang, C.; Liu, Y.-S.; Chang, F.-X.; Lu, M. Pedestrian detection based on YOLOv3 multimodal data fusion. *Syst. Sci. Control Eng.* **2022**, *10*, 832–845. [[CrossRef](#)]
9. Li, Z.; An, Z.; Cheng, W.; Zhou, J.; Zheng, F.; Hu, B. MHA: A multimodal hierarchical attention model for depression detection in social media. *Health Inf. Sci. Syst.* **2023**, *11*, 6. [[CrossRef](#)] [[PubMed](#)]
10. Huo, F.; Liu, Z.; Guo, J.; Xu, W.; Guo, S. UTDNet: A unified triplet decoder network for multimodal salient object detection. *Neural Netw.* **2024**, *170*, 521–534. [[CrossRef](#)]
11. Reyes, M.F.; Xie, Y.; Yuan, X.; d'Angelo, P.; Kurz, F.; Cerra, D.; Tian, J. A 2D/3D multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *205*, 74–97. [[CrossRef](#)]
12. Li, Y.; Gu, W.; Yue, H.; Lei, G.; Guo, W.; Wen, Y.; Tang, H.; Luo, X.; Tu, W.; Ye, J. Real-time detection of laryngopharyngeal cancer using an artificial intelligence-assisted system with multimodal data. *J. Transl. Med.* **2023**, *21*, 698. [[CrossRef](#)] [[PubMed](#)]
13. Rokham, H.; Falakshahi, H.; Calhoun, V.D. A Deep Learning Approach for Psychosis Spectrum Label Noise Detection from Multimodal Neuroimaging Data. In Proceedings of the 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, 24–27 July 2023; pp. 1–4.
14. Rahman, M.A.; Brown, D.J.; Mahmud, M.; Harris, M.; Shopland, N.; Heym, N.; Sumich, A.; Turabee, Z.B.; Standen, B.; Downes, D. Enhancing biofeedback-driven self-guided virtual reality exposure therapy through arousal detection from multimodal data using machine learning. *Brain Inform.* **2023**, *10*, 14. [[CrossRef](#)]
15. Li, J.; Xu, Z.; Xu, L. Vehicle and pedestrian detection method based on improved YOLOv4-tiny. *Optoelectron. Lett.* **2023**, *19*, 623–628. [[CrossRef](#)]
16. He, H.; Li, Z.; Tian, G.; Chen, H.; Xie, L.; Lu, S.; Su, H. Towards accurate dense pedestrian detection via occlusion-prediction aware label assignment and hierarchical-NMS. *Pattern Recognit. Lett.* **2023**, *174*, 78–84. [[CrossRef](#)]
17. Choi, Y.; Kim, N.; Hwang, S.; Park, K.; Yoon, J.S.; An, K.; Kweon, I.S. KAIST multi-spectral day/night data set for autonomous and assisted driving. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 934–948. [[CrossRef](#)]
18. Wagner, J.; Fischer, V.; Herman, M.; Behnke, S. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. In Proceedings of the ESANN 2016 Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 27–29 April 2016; Volume 587, pp. 509–514.
19. Zheng, Y.; Izzat, I.H.; Ziaee, S. GFD-SSD: Gated fusion double SSD for multispectral pedestrian detection. *arXiv* **2019**, arXiv:1903.06999.
20. Engel, J.; Hoffman, M.; Roberts, A. Latent constraints: Learning to generate conditionally from unconditional generative models. *arXiv* **2017**, arXiv:1711.05772.
21. Hussain, M. YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection. *Machines* **2023**, *11*, 677. [[CrossRef](#)]
22. Schulz, D.; Perez, C.A. Two-Stage Pedestrian Detection Model Using a New Classification Head for Domain Generalization. *Sensors* **2023**, *23*, 9380. [[CrossRef](#)] [[PubMed](#)]
23. Liu, S.; Cai, T.; Tang, X.; Zhang, Y.; Wang, C. Visual Recognition of Traffic Signs in Natural Scenes Based on Improved RetinaNet. *Entropy* **2022**, *24*, 112. [[CrossRef](#)] [[PubMed](#)]
24. Gong, L.; Huang, X.; Chen, J.; Xiao, M.; Chao, Y. Reparameterized dilated architecture: A wider field of view for pedestrian detection. *Appl. Intell.* **2024**, *54*, 1525–1544. [[CrossRef](#)]
25. Li, R.; Yan, A.; Yang, S.; He, D.; Zeng, X.; Liu, H. Human Pose Estimation Based on Efficient and Lightweight High-Resolution Network (EL-HRNet). *Sensors* **2024**, *24*, 396. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.