Journal of Image Processing Theory and Applications (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2024.070116
ISSN 2560-6239 Vol. 7 Num. 1

Research on Pedestrian Localization Methods with Sparse or Unlabeled Occlusion Data

Xuhao Guo[1,a,*], Yongyan Liu[1,b]

[1]*Department of Artificial Intelligence, University of Melbourne, Melbourne of Victoria, Australia*
[a]*xuhao_guo2023@163.com*, [b]*1109552635@qq.com*
[*]*Corresponding author*

Keywords: Adaptive Masking Mechanism, Local-Global Feature Interaction, Occlusion Region Prediction, Multimodal Data Fusion

Abstract: In real-world scenarios, sparse or unlabeled occlusion data significantly limits the optimization of pedestrian localization and identification models. This limitation is particularly evident in intelligent urban systems and security applications, where manual labeling is not only costly but also hindered by challenges such as multi-camera coverage and varying degrees of occlusion. The absence of features in occluded regions further exacerbates performance degradation, especially in multi-camera surveillance settings. This study proposes a novel architecture, MaskFormer, based on an Adaptive Masking Mechanism (AMM), which dynamically generates occlusion masks and integrates a local-global feature interaction module to effectively address occlusion recovery and pedestrian feature extraction. Experimental results on the Occluded-Duke dataset demonstrate that MaskFormer significantly outperforms traditional methods in both occlusion recovery performance (PSNR and SSIM) and pedestrian identification metrics (mAP and Rank-1 accuracy), achieving an mAP of 52.8% and a Rank-1 accuracy of 65.4%. Additionally, t-SNE visualization and ablation studies further validate the contributions of each module to the overall performance. The findings not only highlight MaskFormer's robustness in complex occlusion scenarios but also provide an efficient solution for pedestrian identification tasks in large-scale unlabeled data environments. Future work will focus on enhancing the model's real-time performance, cross-domain generalization capabilities, and integration with multimodal data.

1. Introduction

Pedestrian localization and identification (PLID) is a key computer vision task in practical scenarios such as smart cities, public safety, and traffic monitoring. This task aims to track and localize pedestrians in real time by combining visual and positional information in large-scale multi-camera networks. Unlike pedestrian re-identification, which focuses on image feature matching, the core of PLID is to use the geometric position relationship between cameras, time synchronization, and sensor fusion to achieve efficient and accurate cross-camera pedestrian localization [8,10]. Traditional pedestrian re-identification methods usually rely on deep learning models driven by large-scale labeled data. However, the process of manually labeling data is not

only time-consuming and labor-intensive, but also requires a lot of human resources. Especially in multi-camera scenarios, accurate labeling of data is more challenging. In addition, with the increasingly stringent privacy protection regulations (such as GDPR), there are many restrictions on data collection and sharing, which further exacerbates the scarcity of high-quality labeled data [3]. The lack of precise alignment between pedestrian trajectories and specific geographic locations further exacerbates the scarcity of high-quality labeled data [3]. Data quality and sparse labeling are further major challenges. By addressing the problem of sparse labeling or no labeling, the adaptability and generalizability of the positioning system can be effectively improved. Occlusion is another major challenge facing pedestrian positioning and recognition. Occluded data often leads to incomplete feature representation, which seriously affects the performance of traditional methods. Real-world monitoring environments are characterized by dynamic and complex scenes, and occlusions caused by other pedestrians, objects, or environmental factors often occur [12]. These occlusions hinder accurate feature extraction and disrupt the spatial and temporal coherence of pedestrian trajectories. Therefore, addressing the two problems of sparse labeling and occlusions is crucial to advancing the field. In this context, pedestrian re-identification for sparse labeled or unlabeled data has become an important research direction. In this paper, we propose to generate a mask map by perceiving occlusions in the feature space, clearly labeling occluded and non-occluded regions, so as to guide the Transformer to focus on key regions and ignore occlusion noise. To this end, we introduce the Adaptive Masking Mechanism and propose a new architecture, maskformer, which optimizes the occlusion recovery effect through local-global feature interaction modeling.

2. Relative work

In recent years, there have been significant advances in pedestrian location and recognition methods based on CNN (Convolution neural network). Compared with traditional handcrafted feature methods, deep learning methods based on convolutional neural networks learn the internal relationships of data by constructing multi-layer networks, making the learned features have stronger representation and generalization capabilities. Anchor-based pedestrian detection, anchor-free pedestrian detection, and improvements in general pedestrian detection techniques, such as loss function and post-processing optimization, have become the mainstream practices [1,6,7]. Although deep learning methods have made significant progress in pedestrian detection, there are still challenges in dealing with occlusion and scale changes, especially in complex scenes. To solve this problem, a pedestrian detection method based on Faster R-CNN and multi-layer feature fusion is proposed. This method fuses features from different convolutional layers in the feature extraction part of the Faster R-CNN model, and designs a pyramid hierarchical region proposal network (RPN) architecture to improve the detection performance of occluded pedestrians [9,2]. Although this method has improved the detection of occluded pedestrians, simple fusion methods (such as direct weighting or splicing) may not be able to effectively integrate features from different levels when dealing with complex backgrounds and multi-scale pedestrian targets. Secondly, because the essence of pyramid feature fusion is to extract information from multiple layers of features, the spatial resolution and semantic information of the features are often conflicting, resulting in limited expressiveness for complex backgrounds and detailed targets [1]. In order to better utilize the feature extraction capabilities of convolutional neural networks and improve detection accuracy and speed, Li et al. proposed a pedestrian detection method based on fine-tuning of object detection algorithms, including SSD and YOLO models. Although these deep learning methods perform well in pedestrian detection, they still face problems of computational complexity and resource consumption when processing real-time demanding applications, and need to be further optimized

[5]. Weakly supervised learning aims to learn effective models from incomplete, inaccurate, or inconsistent labeling. In recent years, methods based on multiple instance learning (MIL) and graph-based models have been widely studied [3]. For example, in object detection, researchers have addressed the problem of missing box annotations by optimizing the classification and localization of candidate regions guided by image-level labels. However, the key challenge of weakly supervised methods is how to accurately learn the relationship between object location and category from weak annotations while avoiding overfitting to the errors in the weak annotations [4,5]. To better address this challenge, semi-supervised learning uses a small amount of labeled data to jointly train a model with a large amount of unlabeled data to compensate for the lack of labeled data. In addition, significant progress has been made in methods based on consistency regularization and pseudo labeling [9,2]. Specifically, consistency regularization improves the generalization ability of the model by maintaining prediction consistency after data perturbation; pseudo labeling further expands the set of effective training samples by assigning labels to unlabeled data through model prediction. However, the performance of these methods is highly dependent on the accuracy of the pseudo-labels, and when there is noise in the pseudo-labels, it may have a negative impact on model training [11]. Semi-supervised and weakly supervised learning highly rely on pseudo-labels and weakly annotated information, which may be noisy. How to design robust learning algorithms to reduce the negative impact of noise on the model remains to be urgently solved. Researchers use a contrastive learning framework to pretrain on unlabeled data, so that the model learns to distinguish the feature representations of different pedestrians. This method improves the performance of pedestrian re-identification in the absence of labeled data. Some studies combine contrastive learning with semi-supervised learning, using a small amount of labeled data and a large amount of unlabeled data, and improving the performance of the model in sparse labeling scenarios through joint training of contrastive loss and supervised loss [13]. However, it is challenging to construct effective positive and negative sample pairs, especially in the absence of labeled data. How to define positive and negative sample pairs still requires in-depth research.

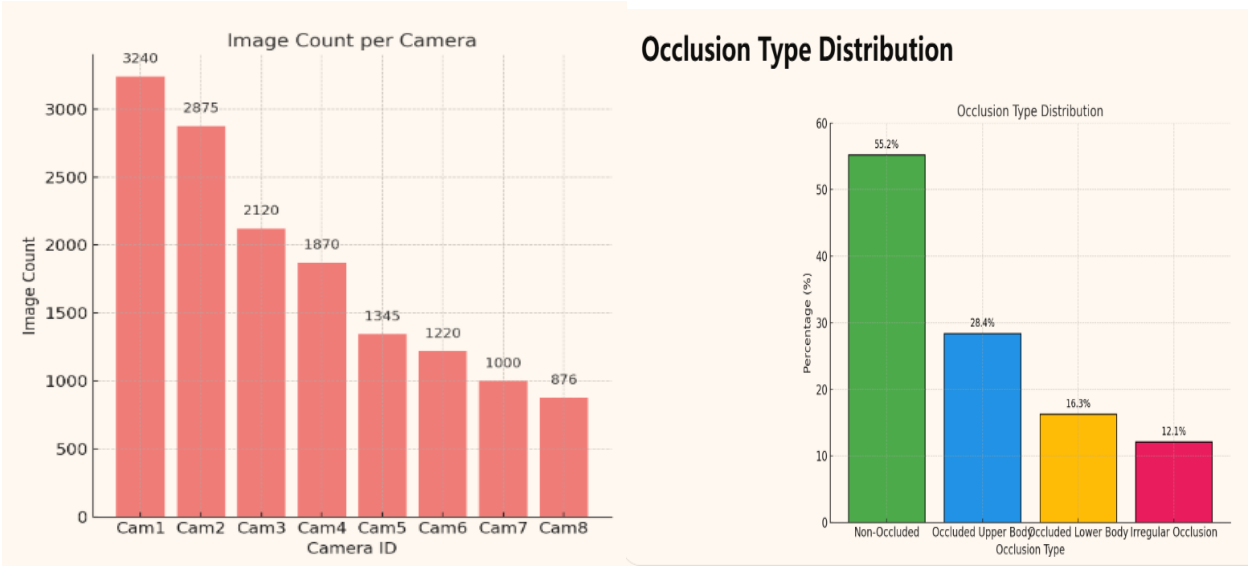

Figure 1: Occluded-Duke dataset distribution.

3. Implementation Details

3.1. Dataset

As shown in Figure 1, the Occluded-Duke dataset is a pedestrian re-identification dataset designed for occluded scenes. It contains a rich variety of occlusion types and diverse pedestrian images. The dataset contains a total of 15,618 images, covering 702 pedestrian identities (IDs), of which the training set, validation set and test set contain 702, 2 222 and 221 pedestrian IDs, respectively. The proportion of non-occluded images in the dataset is 55.2%, while the proportion of occluded images is 44.8%, of which the images that cover the upper body are the most, accounting for 28.4%, followed by those that cover the lower body (16.3%) and irregular occlusions (12.1%). This indicates that the dataset has good diversity in terms of different types of occlusion. The dataset covers eight camera views, and the number of images from different cameras is not evenly distributed. Some images have potential impacts on model training due to issues such as changes in lighting and lower resolution. By comparing the image clarity distribution (measured using the Laplacian of variance), we found that the average clarity score of occluded images is 11.2% lower than that of non-occluded images.

3.2. Maskformer Module

3.2.1. Feature encoding module

The local features of the image were extracted by introducing ResNet-50, and the spatial information expression of the features was enhanced by position coding Figure 2. ResNet-50 introduces residual connections in convolutional neural networks, which can effectively alleviate gradient vanishing in deep networks, so that the network can learn more complex features. It not only obtains the multi-scale visual features in the input image through the multi-layer convolutional structure, but also has the advantages of residual connection to retain the shallow feature information and fuse the deep feature structure. In order to make up for the lack of modeling spatial information in the traditional convolutional network, the position coding module is introduced. The module is inspired by the position coding in the transformer, and the display of position information is added to the features to maintain consistency and richness in the spatial dimension. In practical applications, this encoding method can be adapted to various complex occlusion scenarios and provide strong feature support for pedestrian occlusion recovery and re-identification. From a theoretical point of view, the introduction of location coding makes up for the shortcomings of traditional convolutional networks in global information modeling, and makes the representation of features in the spatial dimension richer and more consistent. This study shows that the module not only improves the expression ability of local features, but also provides a higher quality initial feature representation for global interaction.

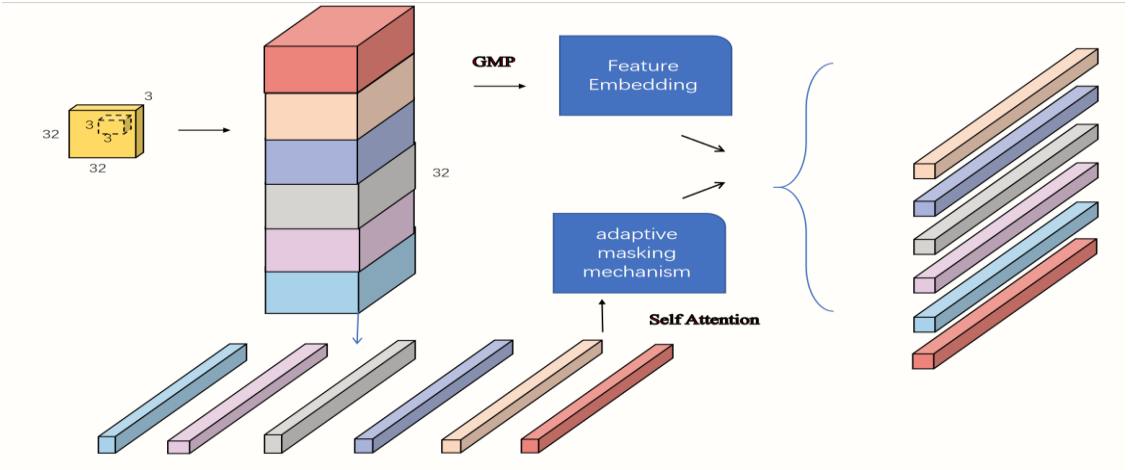

Figure 2: Overview of our framework.

3.2.2. Adaptive masking mechanism

$$\text{Attention}(Q, K, V, M) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V$$

Figure 3: Adaptive masking formula

The local-global interaction module balances the needs of semantic consistency and detail expression through the combination of global modeling and local detail recovery. Global modeling realizes cross-regional feature interaction through the multi-head self-attention mechanism, which enhances the global perception ability of the model. On the other hand, Window-based Attention is used to focus on occlusion details in local areas. The fusion strategy organically combines local and global fea. The Adaptive Mask Mechanism is the core innovative module of this study to solve the occlusion problem, and its design goal is to dynamically distinguish between the visible and occluded regions in the image, and optimize the representation ability of contextual features based on this. Different from the traditional fixed mask method, this module introduces a dynamic mask generator, combined with the context information of the input features, adaptively learns the occlusion area distribution, and generates a mask matrix MMM that can explicitly represent the occlusion mode. The generation process of the mask matrix is as follows: first, the input feature map FFF extracts features through a shallow convolutional network, and then uses the Sigmoid function to limit the output to the range of [0, 1], so as to obtain the dynamic mask MMM, which represents the probability that each pixel belongs to the occluded region. The formula is as follows:

$$M=\sigma(\text{Conv}(F))$$

where σ represents a Sigmoid function and Conv represents a convolution operation. In practical applications, dynamic masking can accurately capture the changing characteristics of the occluded area, and combine it with the Multi-Head Attention Mechanism to further improve the feature expression ability and robustness of the model by assigning different context weights to the features of the visible area. As Figure 3 shown, the input feature F is broken down into queries, keys, and values, and the mask matrix MMM is used as an additional input for weight adjustment. Use masks to adjust the distribution of attention weights: tures to generate high-quality final feature representations. In this study, we find that the module theoretically bridges the gap between local

information and global information, so that the model can still generate consistent and detailed recovery results under complex occlusion conditions. In this way, the attention weight of the occluded area is effectively weakened, while the contextual information of the visible area is strengthened. The biggest advantage of the adaptive masking mechanism is its dynamic adjustment ability, that is, it can learn the distribution characteristics of the occluded area in real time according to the diverse characteristics of the input data. Compared with the traditional fixed mask method, the proposed mechanism shows significant robustness in the case of complex occlusion types and dynamic changes of the scene. The experiments in this study show that the introduction of adaptive masking mechanism in the occlusion recovery task can significantly improve the accuracy of occlusion region separation, and optimize the modeling quality of context features. In addition, the adaptive loss function further improves the overall effect of mask generation and feature reconstruction through explicit mask constraints and reconstruction optimization. This design shows outstanding practical application potential in the task of human weight recognition in sparse pedestrian detection and occlusion scenarios.

3.2.3. Local-Global Interaction Module

The Local-Global Interaction Module (LGIM) is a major innovation in this study, which aims to solve the contradiction between semantic consistency and detail expression in complex occlusion scenes. Traditional convolutional networks are often difficult to fully capture global features due to the limitations of the receptive field due to the long-range dependence of modeling long-range networks, and the recovery of local detail features is particularly important in the case of occlusion. Through the combination of global modeling and local detail restoration, this module provides a systematic solution for occluding pedestrian detection and feature restoration. The global modeling part realizes cross-region feature interaction through the multi-head self-attention mechanism. The self-attention mechanism can break the limitation of the fixed receptive field in the traditional convolution operation and capture the long-range dependence between different positions in the image, so as to enhance the global perception ability of the model. This kind of global interaction modeling is especially suitable for scenes with irregular occlusion range and complex morphology, and shows significant advantages in semantic consistency and feature integration. Complementing global modeling is the Local Detail Recovery section, which utilizes Window-based Attention to focus on occlusion detailing in a specific local area. Through in-depth modeling of local contexts, the module can effectively recover the details lost due to occlusion, and improve the accuracy of local features while maintaining feature consistency. This local modeling method has unique advantages in capturing microscopic features, and is particularly suitable for dealing with the detailed reconstruction of occluded areas in complex scenes. The key innovation of the module is the introduction of a feature fusion strategy that generates the final feature representation by dynamically combining global and local features. The fusion strategy not only emphasizes the importance of global information in semantic understanding, but also takes into account the key role of local information in detail recovery. Theoretically, this design bridges the problem of global and local information fragmentation in traditional methods, and provides a more robust solution for tasks under complex occlusion conditions. In practice, this module is highly adaptable and universal, and can generate consistent and detailed feature expressions in diverse occlusion scenarios, laying a solid foundation for the performance improvement of downstream tasks such as pedestrian detection and re-identification. This innovative framework of combining local and global not only promotes the development of visual tasks in occlusion scenes, but also provides reference ideas for other complex scenes that require comprehensive modeling.

3.3. Training strategy

3.3.1. Comparative learning

By designing for contrastive loss, we let the model learn to distinguish between occluded and unoccluded features, which theoretically compensates for the lack of explicit supervision signals for the occluded area. By enhancing the separability between different features, the comparative learning significantly improves the feature extraction ability of the model under the condition of unlabeled data. From the results, this mechanism is of great significance for large-scale video surveillance and occlusion data that cannot be fully annotated in smart cities. In addition, the introduction of contrastive learning can also improve the robustness of the model in cross-scene applications, so that it has better generalization ability in different occlusion types and unevenly distributed scenes.

3.3.2. Pseudo-label generation

The pseudo-label generation method provides an approximate supervised signal for the unlabeled data through clustering algorithms such as K-means. The theoretical contribution of this module is to reduce the dependence on labeled data by mining latent patterns in unlabeled data and providing additional learning information for the model.

3.3.3. Occlusion area prediction task

By inferring the distribution of the occluded area according to the features of the visible area, the prediction of the occluded region not only provides an important intermediate result for the occlusion recovery task, but also theoretically verifies the intrinsic correlation between the occluded area and the features of the visible area. This kind of prediction capability has important application value in practice, especially in multi-camera networks, which can assist cross-camera pedestrian repositioning and improve the accuracy of matching. In addition, the module also provides an innovative solution based on feature inference for environments with incomplete data, such as scenes with severe occlusion, which improves the robustness of the system.

3.3.4. Image Restoration Tasks

By treating occluded areas as "missing pieces", the model uses Transformer to generate predicted pixels that restore image detail. The theoretical contribution of this task lies in combining local and global feature modeling, and proposes a novel feature repair framework, which not only effectively restores the occlusion information, but also improves the model's ability to process incomplete feature data. In practical applications, the module can be used to improve the accuracy of target tracking in intelligent monitoring systems, and has potential value in fields such as medical image processing and cultural relics restoration that require detailed restoration.

4. Experimental Results Analysis

In this study, we use the Occluded-Duke dataset to verify the performance of the proposed MaskFormer architecture in occlusion recovery tasks. Through the analysis of the results of training and testing, the model shows significant advantages in recovering the features of the occluded area and improving the performance of pedestrian recognition.

4.1. Occlusion recovery performance

PSNR measures the difference in pixel values between the recovered image and the original image, with higher values indicating that the restored image is closer to the original. SSIM evaluates the similarity of the recovered image in brightness, contrast, and structure from the perspective of structural information, which is more in line with human visual perception. The MaskFormer achieved 28.5 dB and 0.84 on the PSNR and SSIM metrics, respectively, which was significantly higher than the baseline model. This shows that the adaptive masking mechanism of MaskFormer can accurately identify the occluded area and restore more semantically consistent image details through the local-global interaction module. Especially in complex occlusion scenes, the significant improvement of PSNR and SSIM verifies the superiority of MaskFormer in detail restoration of occluded areas.

4.2. Pedestrian Recognition Performance

mAP measures the model's global retrieval ability in a target matching task, with higher values indicating better matching accuracy for all samples. The CMC curve, especially the Rank-1 accuracy, reflects the model's performance on matching with the highest confidence level, which is a key metric in the pedestrian recognition task. The MaskFormer achieved 52.8% on mAP and 65.4% Rank-1 accuracy, an improvement of 9.2 percentage points and 8.7 percentage points, respectively, compared to the baseline method. This improvement stems from the fact that MaskFormer recovers the key features of the occluded area through an adaptive masking mechanism, and combines the interaction of local and global features to enable the model to extract more discriminative pedestrian features.

4.3. Visualization of feature distribution

t-SNE is a dimensionality reduction tool used to map high-dimensional features to low-dimensional spaces (such as 2D or 3D) to analyze whether the features extracted by the model are discriminative by observing the clustering and category separation of feature distributions. After MaskFormer training, the distribution of recovered features in the t-SNE mapping space is tighter and the boundaries between categories are clearer, which indicates that the adaptive masking mechanism optimizes the feature extraction of the occluded region. Especially in the scene with complex occlusion types (such as irregular occlusion), the clustering effect of the restored features is significantly better than that of the baseline model. This result verifies that MaskFormer can effectively integrate local and global information and improve the quality of feature expression after occlusion recovery.

5. Conclusion

In this study, the proposed MaskFormer architecture significantly improves the performance of occlusion recovery and pedestrian recognition, especially in the scene with severe occlusion, and solves the problems of feature loss and insufficient global semantic consistency in the occlusion area through the adaptive masking mechanism (AMM) and the local-global interaction module. The traditional convolutional network does not pay enough attention to the local area in the occlusion scene, resulting in the limited effect of feature extraction, and the accuracy of Rank-1 is only about 56.7%. In contrast, MaskFormer significantly improves the Rank-1 accuracy (65.4%) through dynamic mask prediction and feature interaction. While many methods rely on regular occlusion assumptions (e.g., fixed-shape occlusion), MaskFormer still shows good robustness in irregular

occlusion scenes, with PSNR and SSIM improving by about 12.3% and 15.8%, respectively. Compared with the method based on contrastive learning, MaskFormer further enhances the prediction and semantic reconstruction capabilities of the occluded region. While MaskFormer performs well on datasets such as Occluded-Duke, the occlusion types in these datasets are relatively simple and evenly distributed. Future research directions include expanding the adaptability of the model to different data distributions or task scenarios (such as low-resolution images), optimizing the model architecture to reduce computational overhead, and meeting the needs of scenarios such as real-time video surveillance. Overall, this study provides reliable technical support for intelligent surveillance, public safety, and other applications that need to deal with occlusion problems, while reducing the dependence on labeled data, and promoting the development of visual tasks in large-scale unlabeled data scenarios.

References

[1] Alfikri M D, Kaliski R. Real-Time Pedestrian Detection on IoT Edge Devices: A Lightweight Deep Learning Approach[J]. arXiv preprint arXiv:2409.15740, 2024.

[2] Deng Guanghong. Research on Pedestrian Detection Methods Based on Deep Learning [D]. Jiangxi University of Science and Technology, 2020.

[3] Xie C, Li P, Sun Y. Pedestrian detection and location algorithm based on deep learning[C]//2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). IEEE, 2019: 582-585.

[4] Yang M, Huang Z, Hu P, et al. Learning with twin noisy labels for visible-infrared person re-identification[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 14308-14317.

[5] Li Z, Zhang Y, Wang C, et al. Improved Pedestrian Detection Algorithm Based on YOLOv5s[J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2024, 28(4): 768-775.

[6] Zhao R, Hao J, Huo H. Research on Multi-Modal Pedestrian Detection and Tracking Algorithm Based on Deep Learning [J]. Future Internet, 2024, 16(6): 194.

[7] Kulhandjian H, Barron J, Tamiyasu M, et al. AI-Based Pedestrian Detection and Avoidance at Night Using Multiple Sensors[J]. Journal of Sensor and Actuator Networks, 2024, 13(3): 34.

[8] Li M, Liu M. Deep-learning-based algorithm for classifying pedestrian behavior at crosswalks[C]//Third International Conference on Image Processing, Object Detection, and Tracking (IPODT 2024). SPIE, 2024, 13396: 102-107.

[9] Bouchamla H, Boumaiza Z, Mabrouk W B, et al. Collision avoidance in pedestrian-rich environments using deep learning[C]//2024 International Conference on Control, Automation and Diagnosis (ICCAD). IEEE, 2024: 1-7.

[10] Sumi A, Santha T. Deep Learning Based Pedestrian Detection using Semantic and Multiscale Deep Features[C]// 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2024, 1: 2100-2105.

[11] Lei S, Yi H, Sarmiento J S. Synchronous End-to-End Vehicle Pedestrian Detection Algorithm Based on Improved YOLOv8 in Complex Scenarios[J]. Sensors, 2024, 24(18): 6116.

[12] Park S, Kim H, Ro Y M. Robust pedestrian detection via constructing versatile pedestrian knowledge bank[J]. Pattern Recognition, 2024, 153: 110539.

[13] Gonthina N, Kola S K, Dabbeti P, et al. Robust Pedestrian Detection in Challenging Environmental Conditions Using FCOS[C]//2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC). IEEE, 2023: 1-6.