





A Tourist Flow Monitoring and Management System for Scenic Areas Using Image Recognition

Jing Yao¹, Jiawei Wang², Yiran Wang², Fang Hong^{2*}

¹ China Iron & Steel Research Institute Group, Beijing 100081, China

² Faculty of Hospitality and Tourism Management, Macau University of Science and Technology, Taipa, China

Corresponding Author Email: fhong@must.edu.mo

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.420230>

ABSTRACT

Received: 10 September 2024

Revised: 8 February 2025

Accepted: 18 March 2025

Available online: 30 April 2025

Keywords:

image recognition, tourist attractions, tourist object detection, multi-object tracking, visitor flow monitoring, deep learning

With the rapid development of the tourism industry, the surge in visitor volumes has imposed higher demands on the management and operation of tourist attractions. Traditional manual counting methods and infrastructure-based monitoring systems have become insufficient to meet the modern requirements for visitor flow surveillance. Image recognition-based techniques for tourist detection and multi-object tracking have emerged as intelligent solutions capable of providing real-time dynamic data on visitor distribution and movement without interfering with tourists. These methods offer precise decision support for scenic area management. However, existing studies frequently suffer from limitations such as inadequate detection accuracy and tracking failures, compromising the reliability and precision of flow monitoring. Therefore, how to improve the accuracy of tourist object detection in complex environments and achieve stable multi-object tracking has become an urgent problem to be solved. Addressing these challenges, a tourist detection method based on an improved Single Shot MultiBox Detector (SSD) network was proposed, integrated with multi-object tracking techniques for comprehensive visitor flow monitoring. The enhanced SSD network enables more accurate detection of tourist objects under complex environmental conditions, while multi-object tracking ensures stable tracking and counting of individual visitors. Through this system, real-time dynamic variations in visitor flows can be monitored, providing critical data support for tourist safety management, resource allocation optimization, and service quality enhancement. Experimental results demonstrate that the proposed method achieves improvements in both accuracy and efficiency, highlighting its significant practical value and potential for broader application.

1. INTRODUCTION

With the rapid global development of the tourism industry, the annual increase in visitor flow has posed significant challenges to the management and services of tourist attractions [1, 2]. Traditional visitor flow monitoring methods [3-5], which primarily rely on manual counting or infrastructure-based monitoring systems, have been associated with high labor costs and susceptibility to human error, leading to deficiencies in both the accuracy and timeliness of the collected data. As a result, visitor flow monitoring systems based on image recognition technology have gradually emerged as a critical tool for intelligent management in modern tourist attractions. This technology not only enhances monitoring efficiency but also enables the real-time acquisition of dynamic information on visitor movement without disrupting the tourist experience, thereby facilitating scientific decision-making by management authorities.

The significance of visitor flow monitoring in tourist attractions lies in the application of advanced image recognition techniques [6, 7], where computer vision and deep learning methods are employed to accurately detect and track tourist objects [8-10], providing new solutions for visitor flow

surveillance and management. Such systems have been shown to substantially improve the level of intelligent management within scenic areas and contribute to visitor flow prediction, tourist safety management, and resource allocation optimization [11-13]. Through the implementation of these systems, real-time insights into visitor distribution and movement trajectories can be obtained, enabling the optimization of personnel guidance strategies and the layout of service facilities, thereby enhancing the overall visitor experience and improving operational efficiency.

Despite recent advancements in the field, existing studies on image recognition-based visitor flow monitoring methods have exhibited certain limitations. First, many approaches have demonstrated low detection accuracy in complex environments [14], particularly when detecting tourist objects within dense crowds or across varying scales. Second, current multi-object tracking techniques have been prone to tracking loss or mis-tracking under long-duration and high-density conditions [15, 16], thereby compromising the accuracy of visitor flow statistics. Furthermore, much of the existing research has remained reliant on traditional object detection networks [17], failing to fully leverage the advantages offered by deep learning algorithms, which has resulted in deficiencies

in both detection efficiency and precision. Therefore, the enhancement of tourist detection accuracy and the stabilization of multi-object tracking performance remain critical challenges in the field.

In response to these challenges, a tourist detection method based on an improved SSD network was proposed, integrated with multi-object tracking techniques to achieve precise monitoring and statistical analysis of visitor flow in tourist attractions. Specifically, the research focuses on two primary components: first, tourist detection based on the improved SSD network, where detection accuracy is significantly enhanced under complex environmental conditions; second, visitor flow monitoring through multi-object tracking and statistical analysis, enabling the efficient tracking of tourists and the dynamic recording of visitor distribution changes within scenic areas. The principal innovation of this study lies in the application of an improved deep learning framework, which optimizes existing tourist detection and tracking strategies and demonstrates substantial practical value and potential for widespread adoption. Through this system, real-time and accurate monitoring of visitor flow can be achieved, providing robust data support for management operations and advancing the process of intelligent management in the tourism industry.

2. TOURIST OBJECT DETECTION IN SCENIC AREAS BASED ON AN IMPROVED SSD NETWORK

In practical applications, tourists in scenic areas are often situated within complex background environments, characterized by numerous interfering factors such as cluttered scenes, variations in lighting, and the presence of buildings and vegetation. These conditions significantly affect the accuracy of object detection. Furthermore, tourists frequently appear at varying scales and densities, particularly in crowded regions of scenic areas, where objects tend to be small or overlapping. Traditional SSD algorithms have shown limited capability in detecting such small-scale objects. Due to the fixed convolutional kernel sizes and the inadequate feature extraction capabilities of shallow layers in the original SSD architecture, considerable challenges have been encountered in handling these complex conditions. To enhance detection accuracy under the environmental conditions of scenic areas, optimizations targeting these limitations of the SSD algorithm were proposed, aiming to strengthen feature extraction capabilities and improve small-object detection performance, thereby increasing the accuracy of tourist detection.

The proposed improvement strategy first involves the incorporation of an Inception module. By extracting features at multiple scales through parallel pathways, the Inception module significantly enhances the model's adaptability to multi-scale objects, particularly benefiting the detection of small tourist objects. The optimized Inception module was integrated into the SSD network by replacing the Conv4_3 and Conv7 convolutional layers, thereby improving the representational power of shallow features and expanding the receptive field, enabling the model to better handle detection tasks in complex backgrounds. Moreover, an exclusive loss term, denoted as *Lextude*, was introduced into the loss function to refine prior box localization. This modification results in more precise bounding boxes and accelerates the convergence speed of the model. In the dynamic and complex environments characteristic of tourist attractions, detection

models are required to maintain high precision, rapid response times, and strong robustness. The aforementioned optimizations allow the model to achieve more accurate tourist localization while effectively addressing challenges associated with background interference and small-object detection.

2.1 Structural optimization of the SSD network

In the task of tourist detection within scenic areas, the complexity of the environment and the diversity of objects require that the network efficiently extract multi-scale features and accurately recognize tourist objects. Traditional convolutional neural networks (CNNs) have typically enhanced feature extraction capabilities by increasing network depth. However, in dynamic and complex environments such as those found in tourist attractions, excessively deep networks are prone to overfitting and result in increased computational complexity, along with higher consumption of storage and hardware resources. In practical scenarios, tourist objects generally exhibit significant variations in scale and density, particularly when visitors are sparsely distributed or concentrated within crowded regions. These conditions lead to considerable differences in object sizes. Therefore, the simultaneous processing of multi-scale feature information and the enhancement of network adaptability to small and varying-scale objects have been identified as key challenges in the optimization of object detection models. To address these challenges, optimization of the Inception module was undertaken to overcome the limitations of traditional convolutional networks in multi-scale object detection.

Figure 1 illustrates the architecture of the traditional Inception module. The specific optimization strategy is outlined below. First, to address the high computational cost associated with the 5×5 convolutional kernels in the Inception module, a replacement strategy was adopted, where each 5×5 convolutional operation was decomposed into two consecutive 3×3 convolutions. This replacement not only reduces the total number of parameters generated during model training but also decreases the number of feature maps produced at each network layer, thereby effectively shortening computation time. For real-time tourist detection scenarios in scenic areas, where computational efficiency is critical, such optimization significantly improves processing speed, enabling efficient real-time detection under constrained hardware resources.

Furthermore, batch normalization (BN) operations were introduced into the optimized Inception module. Specifically, BN was applied after each 1×1 and 3×3 convolutional operation. Through this approach, feature data across channels were normalized within each batch, stabilizing the mean and variance and obtaining the learnable hyperparameters. In real-world applications within scenic areas, where tourists exhibit diverse distributions and postures, the use of BN allows the network to better accommodate the variability of multi-dimensional data, resulting in more accurate and stable feature extraction. Figure 2 presents the architecture of the improved Inception module.

The improved SSD network model proposed for tourist detection in scenic areas enhances detection accuracy by optimizing the network structure and strengthening feature extraction capabilities. During the encoding stage, input video frames of 300×300 resolution was initially processed through multiple convolutional operations and feature extraction using the Visual Geometry Group (VGG) 16 backbone network. Subsequently, the local features were further refined within the

optimized Inception module. By incorporating parallel convolutions of different kernel sizes, the Inception module effectively captures multi-scale features of tourist objects, with particularly notable improvements observed in the detection of small-sized objects. Different scale features were concatenated (Concat operation) and normalized through BN, enabling the model to more effectively handle diversified backgrounds and tourist objects. This approach not only suppresses the influence of irrelevant background pixels but also improves adaptability to complex environments. Following the fusion and normalization of feature layers, a feature layer of size $38 \times 38 \times 512$ was generated, forming a robust foundation for subsequent object detection tasks.

During the decoding stage, the network was used to predict the location, category, and confidence scores of objects by analyzing prior boxes, with the final bounding boxes selected

through score ranking and Non-Maximum Suppression (NMS) algorithms. In the complex environments of tourist attractions, where objects are diverse and backgrounds are cluttered, the model effectively extracted key object information from multi-scale feature maps and performed precise regression and classification. Through the construction of a multi-scale feature pyramid and the integration of regression and classification tasks, the improved SSD network achieved more accurate recognition and localization of tourist objects. Even under conditions characterized by high crowd density and complex backgrounds, high detection accuracy was maintained. This architecture, by enhancing feature fusion and multi-scale feature processing capabilities, enabled the model to exhibit greater robustness and precision within the dynamic and complex environments typical of scenic areas.

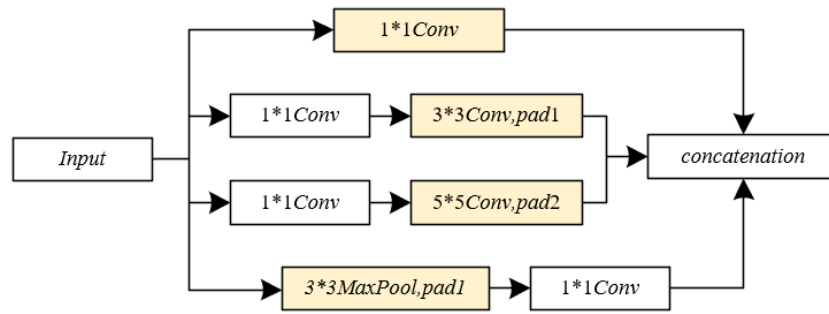


Figure 1. Architecture of the traditional Inception module

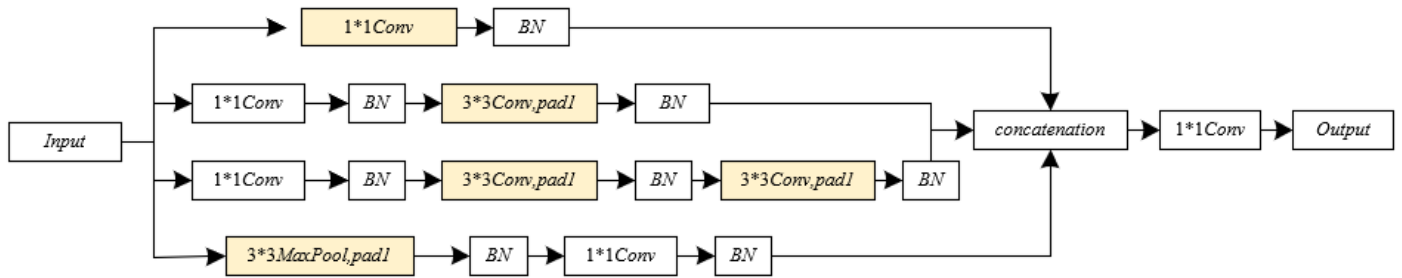


Figure 2. Architecture of the improved Inception module

2.2 Optimization of the SSD network loss function

The loss function of the improved SSD network was optimized to further enhance detection accuracy and robustness. In the original SSD network, the loss function primarily consisted of two components: localization and confidence errors. Although this structure proved effective under most conditions, it exhibited limitations in complex, dynamic environments such as those found in tourist attractions, where frequent false and missed detections occurred. To improve the precision of tourist object localization, an additional mutually exclusive loss term (M_{EX}) was introduced into the original loss function. The incorporation of this loss term effectively reduced the occurrence of false positives, particularly in regions characterized by high tourist density or complex backgrounds, by constraining the network during the prediction process to avoid incorrect overlaps and misclassifications. Further optimization of the loss function was achieved by reweighting the different loss components through a weighted summation approach. Given the diversity in scale, posture, and location of tourist objects, the traditional loss function often struggled to

ensure accurate detection across different object types. By integrating localization error, confidence error, and the newly added mutually exclusive loss into the weighted total loss, the model's stability and accuracy in complex environments were significantly enhanced. The loss function is expressed as:

$$M = \left(\frac{1}{V} (M_{CO}(a, z)) + \beta M_{LO}(a, m, h) \right) + \alpha M_{EX} \quad (1)$$

The newly introduced loss term M_{EX} was specifically designed to increase the distance between the detection boxes of tourist objects and the nearby ground truth boxes classified as non-tourist objects, thereby optimizing detection performance. The detailed calculation procedure is outlined as follows:

(a) A set of processed anchor boxes was obtained, and both the designated object set and the exclusion object set were constructed. Specifically, the set of all ground truth boxes was collected and denoted as $\{H\}$. Simultaneously, all prior boxes whose Intersection over Union (IoU) with each ground truth box H_v exceeds a predefined threshold were collected into a set denoted as $\{O\}$. For each ground truth box in the set $\{H\}$,

the prior box from $\{O\}$ with the highest IoU value was selected. After traversing all ground truth boxes, the selected prior boxes form a new set $\{OS\}$, which is defined as the designated object set. Subsequently, all elements of the designated object set $\{OS\}$ were removed from the overall prior box set $\{O\}$, resulting in an updated prior box set denoted as $\{O-OS\}$. Following this, for each ground truth box H_v , the prior box with the highest IoU value was selected from the updated set $\{O-OS\}$. After traversal of all ground truth boxes, the selected prior boxes form a new set denoted as O_{DE}^H , which is referred to as the exclusion object set. The designated object set $\{OS\}$ is expressed as:

$$\{OS\} = \operatorname{argmax} IOU(H, O) \quad (2)$$

The exclusion object set O_{DE}^H is expressed as:

$$O_{DE}^H = \operatorname{argmax} IOU\{H, (O - OS)\} \quad (3)$$

(b) New IoU values were computed. The overlapped IoU values between all prior boxes and the designated object set $\{OS\}$ and the exclusion object set O_{DE}^H were calculated to form a new loss term. When designated objects correspond to tourists in scenic areas, this loss term effectively suppresses excessive overlap between prior boxes and non-tourist objects, ensuring that prior boxes are more accurately aligned with tourist object regions rather than other object categories. This process increases the distance between the tourist object boxes and the surrounding non-tourist object boxes, thereby improving the localization precision of prior boxes and optimizing detection outcomes. Through this strategy, the improved loss function can better handle object detection tasks

in the complex backgrounds typical of tourist attractions, significantly enhancing detection accuracy for distant or small-scale tourist objects, while simultaneously reducing false and missed detections. A more accurate foundation is thus established for subsequent object tracking and behavior analysis. The specific calculation is expressed as:

$$M_{EX} = \frac{AREA(O_{DE}^H \cup \{OS\})}{AREA(\{O\})} \quad (4)$$

Finally, to optimize the regularization term within the improved loss function, an $L2$ norm was adopted to prevent model overfitting. The $L2$ norm constrains the hyperparameters in the loss function by calculating the square root of the sum of the squares of all weights, thus mitigating the risk of overfitting due to an excessive number of hyperparameters. This optimization strategy is particularly important for the complex and dynamic scenarios typical of tourist attractions, enhancing the generalization ability and stability of the model and ensuring consistently high detection performance across varying environments. The specific calculation method is defined as:

$$\|a\| = SQRT[a_1^2 + a_2^2 + \dots + a_v^2] \quad (5)$$

To avoid the square root operation, the above expression was further optimized as:

$$\|a\| = 1/2[a_1^2 + a_2^2 + \dots + a_v^2] \quad (6)$$

Figure 3 illustrates the architecture of the improved SSD network model.

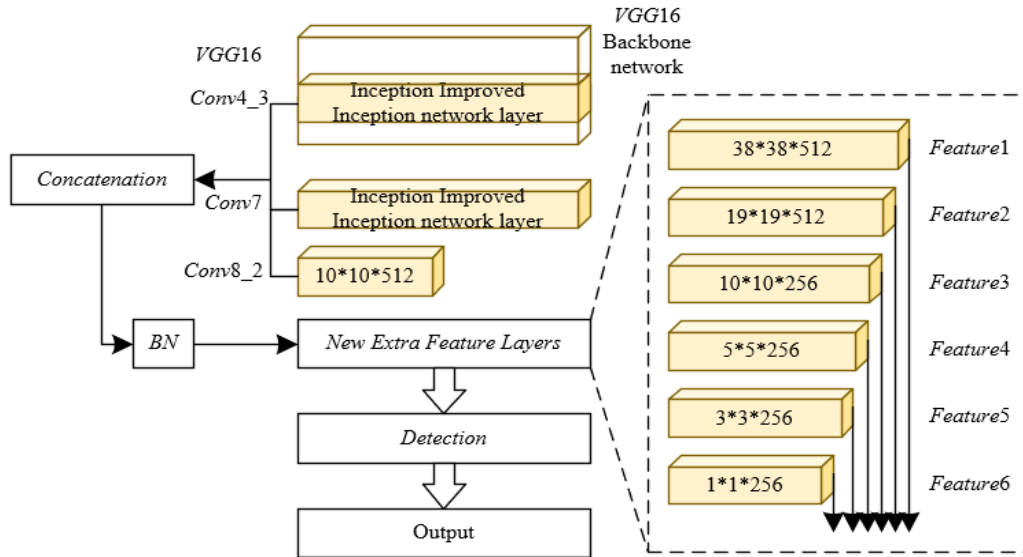


Figure 3. Architecture of the improved SSD network model

3. TOURIST TRACKING AND FLOW STATISTICS IN SCENIC AREAS

To achieve accurate visitor flow monitoring in scenic areas, a multi-object tracking and statistical method for tourists based on the Deep Learning-based Simple Online and Realtime Tracking (DeepSORT) algorithm was adopted, primarily due

to its excellent real-time performance and stability in handling complex dynamic environments. In tourist attractions characterized by dense crowds and complex object behaviors, real-time and accurate tracking of tourists is crucial. DeepSORT provides more precise solutions for issues such as tourist interactions, occlusions, and data association between consecutive frames. By incorporating deep learning-based

feature representations, DeepSORT significantly enhances the model's ability to distinguish between visually similar objects, thereby enabling more accurate identification and differentiation of individual tourists, particularly in crowded scenarios where object similarity is high.

3.1 SORT algorithm

The core concept of the SORT algorithm is based on Kalman filtering and the Hungarian algorithm, enabling the tracking of multiple tourists by predicting and updating object positions. Specifically, SORT first employs a Kalman filter to predict the current position of each tourist within the scenic area and estimate their location in the subsequent frame. The Hungarian algorithm is then utilized to match the predicted positions with the actual detected tourist locations by calculating the IoU between bounding boxes, thereby establishing an optimal object association. Through continuous iterative updates, the SORT algorithm can track each tourist's movement trajectory in real time, ensuring accurate localization in every frame of video imagery.

In complex environments typical of tourist attractions, tourist behavior often exhibits significant variability, with frequent occurrences of occlusions and interactions among objects. Although the SORT algorithm offers notable advantages in terms of real-time performance and computational efficiency, challenges such as frequent identity switches may arise when tourists experience occlusion or rapid position changes. Nevertheless, SORT remains effective in managing large-scale tourist data and provides stable support for subsequent multi-object tracking processes. By continuously updating tourist location information in real time, SORT enables efficient tracking across dynamic environments within scenic areas, including dense crowd zones and areas characterized by rapid visitor movement, thereby facilitating the monitoring of visitor flow patterns and dynamic changes in crowd density.

3.2 DeepSORT algorithm

The DeepSORT algorithm represents an improvement over the original SORT framework by incorporating a re-identification (ReID) network for feature extraction, thereby effectively preserving individualized feature information among different tourists. This enhancement significantly reduces the frequency of identity switches caused by object occlusion or interactions. In scenic areas characterized by high visitor mobility and dynamic object behaviors, the improvements introduced in DeepSORT are particularly critical. By utilizing deep learning-extracted appearance features and motion state information, DeepSORT substantially improves the accuracy of tourist tracking and identity preservation, effectively reducing object loss and misassociation incidents. The fundamental process of the DeepSORT algorithm is described below. First, the Mahalanobis distance is calculated to evaluate the motion state of each object, predicting the probable location of the object at the next time step. In the dynamic and complex environments of scenic areas, tourist motion trajectories are often influenced by factors such as rapid walking, sudden changes in direction, or temporary stops. The Mahalanobis distance utilizes motion information—including velocity and direction—combined with the Kalman filter's predicted motion state to estimate the object's subsequent position in the next frame. Let the center

point of a bounding box in image coordinates be denoted by (i,n) , the aspect ratio be denoted by b , and the height be denoted by g . The predicted motion state by the Kalman filter is represented by (i,s,b,g) , where these eight parameters of $(i,n,b,g,i,,b,g)$ describe the ground truth box and the predicted box. Assuming that the coordinates of the k -th detection box are denoted by f_k , and the predicted coordinates by the u -th tracker are denoted by b_u , with the covariance matrix between the detection coordinates and the mean predicted coordinates represented by T_u and the Mahalanobis distance between two anchor boxes by $f^{(1)}(u,k)$. The degree of motion association is computed using the following formula:

$$f^{(1)}(u,k) = (f_k - b_u)^T T_u^{-1} (f_k - b_u) \quad (7)$$

Since the Mahalanobis distance calculation is not entirely accurate, covariance and standard deviation were introduced to mitigate measurement uncertainty. This step was implemented by thresholding the Mahalanobis distance $f^{(1)}(u,k)$, as shown below. If the calculated Mahalanobis distance is less than a specified threshold $s^{(1)}$, the motion states of the two anchor boxes are considered successfully associated. In the context of tourist attractions, this step assists in accurately associating and predicting tourist trajectories even during brief occlusions, such as when individuals are temporarily blocked by other tourists or structures, thereby ensuring the continuity of tracking.

$$y_{u,k}^{(1)} = \prod [f^{(1)}(u,k) \leq s^{(1)}] \quad (8)$$

When the camera moves along with the object, the effectiveness of Mahalanobis distance may degrade. For example, when the camera follows a moving face or tracks an object, motion-only association becomes unreliable. In such cases, appearance feature association using cosine distance must be introduced. DeepSORT addresses this by extracting a 128-dimensional unit vector for each anchor box through a ReID network model. Cosine similarity is then used to compare the angle size of the feature vector e at each anchor box coordinate f_k , requiring the norm of the feature vector to be 1, i.e., $\|e^{(u)}\|=1$. In tourist attractions, this process is very important for dealing with the variations of tourists in viewpoint and illumination. This feature-matching method enables the ReID of temporarily occluded tourists across video frames, ensuring the continuity and reliability of multi-object tracking. Assuming the feature vector extracted from the detection box coordinate f_k is denoted as e_k , and the vector library used for storage is denoted as e_j , the cosine distance between two anchor box coordinates can be calculated as follows:

$$f^{(2)}(u,k) = \text{MIN} \{1 - e_k^T e_j^{(u)} \mid e_j^{(u)} \in R_u\} \quad (9)$$

After the cosine distance is computed, the DeepSORT algorithm compares it against a predefined threshold $s^{(2)}$. If the calculated cosine distance is less than the threshold, the appearance features of the two anchor boxes are considered successfully associated. Particularly in scenic areas, where tourists frequently disappear from the field of view due to occlusion by other individuals or objects, the use of cosine distance enables the algorithm to re-identify tourists even after multiple frames of occlusion through feature matching. The

following equation can be used as the indicator:

$$y_{u,k}^{(2)} = \prod [f^2(u,k) \leq s^{(2)}] \quad (10)$$

Subsequently, the matching condition between the current detection box and the Kalman filter-predicted object box is examined. If the association satisfies the specified threshold, a weighted calculation is performed. By integrating the weighted sum of two different measurement methods, DeepSORT comprehensively evaluates the degree of object association. In the context of tourist attractions, the trajectories of tourists between video frames may exhibit significant variations, and environmental factors may introduce partial measurement inaccuracies. Therefore, this weighted calculation dynamically adjusts the relative importance of the Mahalanobis distance and cosine distance, ensuring that the matching degree of objects is accurately assessed under varying conditions.

Finally, DeepSORT calculates the final matching degree by combining the Mahalanobis distance and cosine distance through the weighted calculation. Based on the predefined threshold range, a determination is made regarding the successful association of two anchor boxes. If the combined result $z_{(u,k)}$ falls within the set threshold range for both Mahalanobis and cosine distances, the two anchor boxes are considered successfully associated.

$$z_{u,k} = \eta f^{(1)}(u,k) + (1-\eta) f^{(2)}(u,k) \quad (11)$$

3.3 Training of the ReID model of the DeepSORT network

The proposed tourist multi-object tracking algorithm for visitor flow monitoring in scenic areas was designed as follows: the optimized SSD network is first applied to enhance the speed and accuracy of tourist object detection in scenic environments; subsequently, the ReID feature extractor is retrained using a tourist ReID dataset tailored for scenic areas to improve object data association between consecutive frames, ensuring suitability for the tourist tracking and matching process during the tracking stage.

The core objective of introducing the ReID network is to enhance the matching accuracy of tourists across different video frames or surveillance regions. The training of the ReID network must account for the unique behavioral patterns and appearance characteristics of tourists within scenic environments. For example, tourists often exhibit highly diverse clothing styles and engage in behaviors such as pausing, gathering, or weaving through crowds. Therefore, specialized training of the ReID network targeting tourist objects is essential to extract robust features under varying environmental conditions. The structural parameters of the ReID network differ from those of the original pedestrian ReID network used in DeepSORT. In the conventional DeepSORT framework, the ReID network is typically based on the Faster Region-based Convolutional Neural Network (Faster R-CNN) and Residual Network (ResNet) architectures for pedestrian ReID training. However, these structures are not fully adapted to the diverse and dynamic characteristics of tourists in scenic areas. The heterogeneity of tourist populations and the complexity of their activity patterns render conventional pedestrian-focused approaches insufficient. Accordingly, a customized ReID training approach was developed, incorporating feature extraction modules

specifically designed to optimize the ability to identify tourist objects. For example, the network structure was adjusted to capture the distinctive appearance characteristics of tourists within complex environments, enhancing the recognition of dynamic tourist behaviors, such as group activities or rapid movements commonly observed in scenic areas. Through these customized ReID training strategies, effective tourist tracking and identification within scenic areas can be achieved, thereby significantly improving object matching accuracy under high-density conditions and providing high-quality data support for visitor flow monitoring and management.

The overall processing framework of the improved SSD algorithm combined with the DeepSORT-based tourist multi-object detection and tracking method for scenic areas is described below. The improved SSD algorithm is first employed to process input video frames, performing object detection and outputting tourist location information. Through structural enhancements, the SSD algorithm achieves improved detection precision, enabling more accurate identification of tourists under complex background conditions while assigning a unique label to each detected tourist. Given the high density and dynamic movement patterns of tourists in scenic environments, the improved SSD algorithm effectively identifies multiple tourists within a frame. Additionally, preprocessing steps such as image filtering and denoising are applied to enhance video quality, providing clearer input data for subsequent tracking operations. Subsequently, the DeepSORT algorithm is utilized to predict and track the motion states of tourists using a Kalman filter. The ReID module is integrated to extract appearance feature information for each tourist. By combining motion state predictions and appearance features, an association matrix is formed. Leveraging this association matrix, the DeepSORT algorithm performs precise object tracking across consecutive frames, addressing challenges related to occlusion and concealment in complex environments. In scenarios characterized by dense crowds or rapid tourist movement, the Hungarian algorithm is employed to perform matching within the association matrix, ensuring that each tourist's trajectory is accurately tracked and minimizing the risks of incorrect matching or object loss.

3.4 Visitor flow monitoring in scenic areas

Following the completion of tourist multi-object tracking and statistical analysis, visitor flow monitoring in scenic areas can be further enhanced through intelligent analysis and real-time data processing to achieve more precise management outcomes. After the completion of multi-object tracking and statistical aggregation, the core strategy for visitor flow monitoring is the establishment of a dynamic, region-based real-time monitoring and early warning system. Initially, tourist position and trajectory data derived from multi-object tracking must be utilized, in combination with the spatial attributes of scenic area maps, to partition key monitoring zones. Edge computing devices or cloud servers were employed to calculate tourist density in each zone in real time. For instance, coordinate information from tracking data was matched with geofencing technology to compute the instantaneous number of individuals within each sub-region, and a tiered warning mechanism was activated by setting dynamic thresholds. Simultaneously, trajectory features such as movement direction and speed were analyzed to predict changes in visitor distribution trends over the next 5 to 15

minutes. Dynamic information displays or scenic area mobile applications were utilized to disseminate guidance information, facilitating a transition from passive statistics to proactive intervention. Based on the analysis of visitor distribution within the scenic area, critical information such as congestion levels and dwell times across different zones can be accurately captured. This enables real-time monitoring of crowd density by scenic area managers, who can take timely measures to guide tourists and mitigate safety risks or declines in service quality caused by overcrowding. Through the use of real-time monitoring data, dynamic scheduling of visitor flow across different time periods and attractions can be implemented, leading to smoother tourist experiences and minimizing congestion during peak periods. Figure 4 presents the process flow for visitor flow monitoring in scenic areas.

In combination with tourist multi-object tracking and statistical data, more refined flow prediction and management can also be realized within scenic areas. Through long-term

accumulation and analysis of tracking data, predictive models for visitor flow can be established by incorporating factors such as weather conditions, holidays, and special events. These models can provide scientific support for planning and early warning mechanisms within scenic areas. For instance, forecasts regarding the number of visitors in specific time periods or regions can enable early adjustments to personnel deployment and transportation arrangements, thereby mitigating the pressure associated with peak visitor flows. Moreover, by leveraging artificial intelligence and machine learning technologies, scenic areas can optimize tourist route recommendation systems, dynamically adjusting tour paths based on real-time data to prevent crowding at particular attractions and improve overall operational efficiency. Through the implementation of these strategies, scenic areas are enabled to manage visitor flow with greater precision and efficiency, enhancing both the safety and satisfaction of tourists.

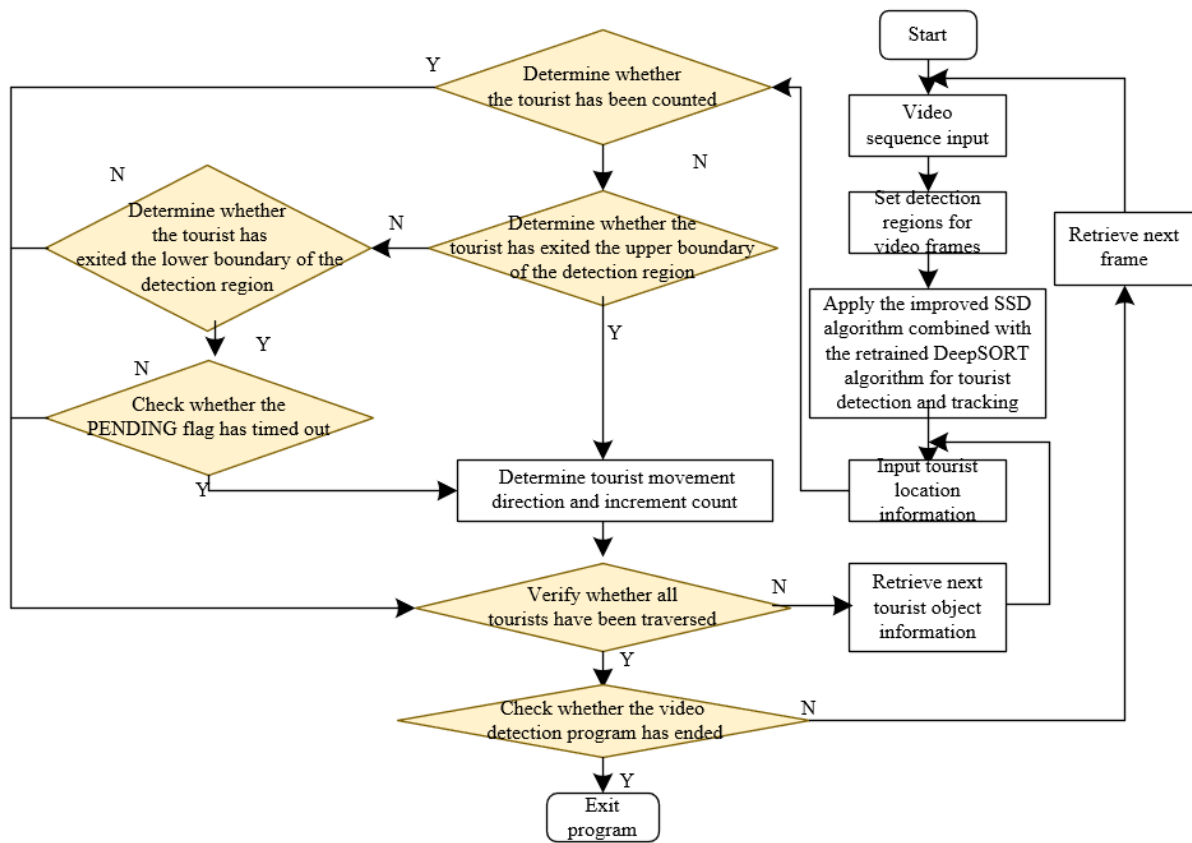


Figure 4. Visitor flow monitoring process in scenic areas

4. EXPERIMENTAL RESULTS AND ANALYSIS

In the experimental evaluation, a comparative analysis was conducted between the original SSD algorithm and the improved SSD algorithm in terms of object detection performance, with particular focus on the variation of loss values during the training process. As shown in Figure 5, a distinct difference can be observed between the two approaches. In the case of the original SSD algorithm, stabilization of the loss value occurred only after approximately 125 iterations, with an overall slower rate of decrease. This behavior indicates that the original algorithm exhibited a slower convergence speed when exposed to complex environments, potentially resulting in prolonged

training times and increased consumption of computational resources. By contrast, the improved SSD algorithm demonstrated significantly enhanced performance during training. Specifically, within the first 25 iterations, the loss value rapidly decreased to approximately 2.0, and stabilization was achieved by around the 90th iteration. This phenomenon suggests that the improved algorithm possesses a stronger adaptation capability to complex scenes, enabling faster convergence and substantially reducing training time, while also lowering hardware resource demands. Based on the experimental findings, it can be concluded that the improved SSD algorithm, through structural optimization or adjustments in the training strategy, achieves more efficient object detection performance, particularly within complex tourist

scenic area environments. The rapid decline of the loss value signifies that the improved algorithm is capable of completing effective learning within a shorter period, thereby enhancing system real-time performance and processing efficiency. This outcome is critically important for visitor flow monitoring systems in scenic areas, where real-time responsiveness and resource utilization efficiency constitute key operational factors. By reducing training time and accelerating model convergence, the improved algorithm not only enhances detection accuracy but also provides more reliable data support for multi-object tracking and visitor flow monitoring, thereby contributing to the realization of efficient and precise scenic area management and tourist traffic control.

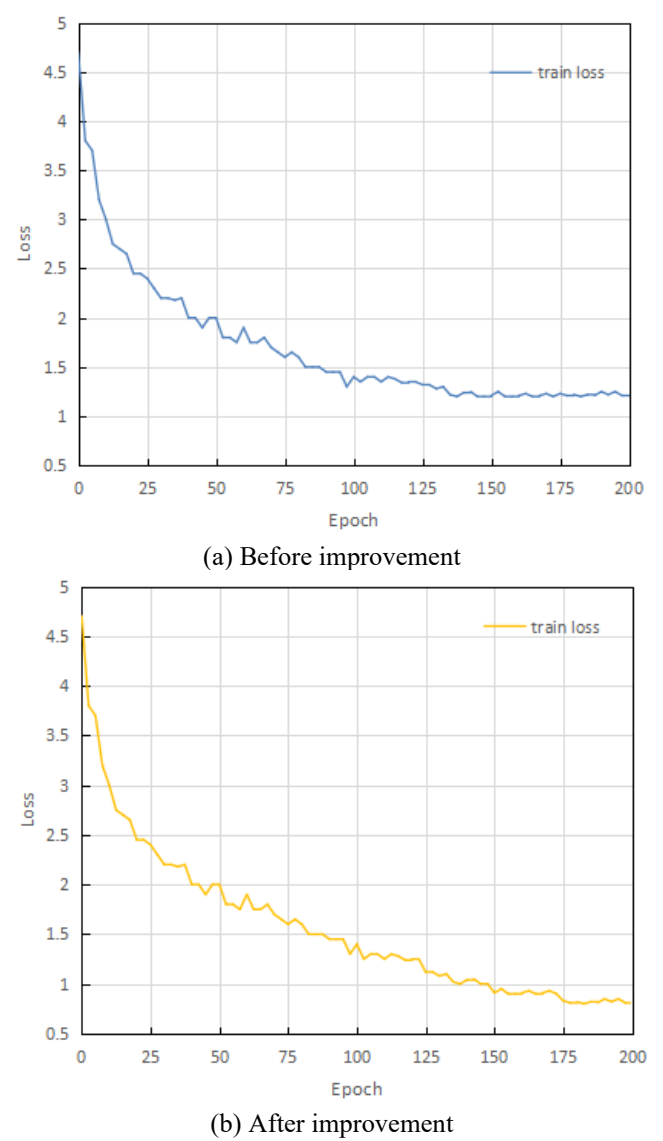


Figure 5. Loss curves of the SSD algorithm before and after improvement

As shown in the experimental results presented in Table 1, the performance of different models in tourist object detection within scenic areas varies significantly. Faster R-CNN achieved a mean Average Precision (mAP) of 71.5% and a frame rate of 17 frames per second (FPS), demonstrating relatively high detection accuracy but limited real-time performance. The original SSD model obtained a comparable mAP of 71.6% while achieving a higher frame rate of 24 FPS, indicating an improvement in processing speed. You Only Look Once version 3 (YOLOv3) achieved a slightly higher

mAP of 72.8% and a frame rate of 26 FPS, demonstrating a favorable balance between detection accuracy and real-time performance. The improved SSD model exhibited the best detection accuracy among all models, achieving an mAP of 73.9%. Although its frame rate was 22 FPS, slightly lower than that of YOLOv3, it still demonstrated relatively high processing speed along with superior accuracy, thereby validating the performance enhancements achieved through model improvement. Based on the experimental results, it can be concluded that the improved SSD algorithm outperforms other models in terms of detection accuracy, achieving the highest mAP value of 73.9%, while maintaining a competitive processing speed of 22 FPS. This finding suggests that although the improved SSD model is marginally less real-time compared to YOLOv3 (26 FPS), its superior accuracy makes it better suited for visitor flow monitoring systems in scenic areas, where high detection precision is critical. Furthermore, compared to traditional models such as Faster R-CNN and the original SSD, the improved SSD model not only enhances detection accuracy but also reduces training and inference times, demonstrating more efficient resource utilization. These advantages enable the improved SSD model to effectively support real-time monitoring and management requirements within scenic areas.

Table 1. Tourist object detection performance of different models in scenic areas

Algorithm Model	mAP (%)	FPS
Faster R-CNN	71.5	17
Original SSD	71.6	24
YOLOv3	72.8	26
Improved SSD	73.9	22

Figure 6 presents the detection results of tourist objects within scenic areas. As observed in Figure 6, the detection method based on the improved SSD network accurately identifies tourist objects in various complex scenes within the scenic area. On the left side, in a scene featuring scenic buildings and mountainous backgrounds, and on the right side, at the entrance of the scenic area partially obscured by trees, tourists are clearly enclosed within red bounding boxes. This demonstrates the enhanced detection precision of the improved network under complex environmental conditions, aligning with the research objectives outlined in this study. Even in the presence of background interference factors such as mountains and dense vegetation, tourist objects can still be effectively captured, thereby verifying the high reliability of the proposed detection method in practical scenic area scenarios.



Figure 6. Tourist object detection results in scenic areas

A significant difference in statistical accuracy between

scenarios without and with DeepSORT-based tracking can be observed based on the experimental data presented in Tables 2 and 3. In Table 2, when DeepSORT was not employed, the detection errors were as follows: 5.35% for Video 1, 1.67% for Video 2, 2.38% for Video 3, 3.52% for Video 4, and 6.81% for Video 5. These results indicate that although the improved SSD algorithm enhanced detection accuracy, the absence of an effective multi-object tracking method resulted in relatively larger errors, particularly for Video 5, where the error reached as high as 6.81%. In contrast, after the application of DeepSORT, the errors shown in Table 3 were significantly reduced: 1.19% for Video 1, 1.11% for Video 2, 0% for Video 3, 2.35% for Video 4, and 2.27% for Video 5. After using DeepSORT for multi-object tracking, the consistency between the detection results and the actual number of tourists significantly improved. Notably, in Video 3, the detected number of tourists exactly matched the actual number, yielding an error rate of 0%. Based on the experimental results, it can be concluded that the DeepSORT multi-object tracking technique significantly improves the accuracy of tourist flow statistics in scenic areas. By employing DeepSORT, the system is capable of more accurately tracking the motion trajectories of individual tourists, thereby reducing instances of redundant counting and missed detections, which collectively contribute to a substantial reduction in overall statistical error.

Table 2. Tourist flow statistics in scenic areas without using DeepSORT

Test Video	Actual Number of Tourists	Detected Number of Tourists	Error (%)
Video 1	168	159	5.35
Video 2	179	176	1.67
Video 3	42	41	2.38
Video 4	85	82	3.52
Video 5	44	41	6.81

Table 3. Tourist flow statistics in scenic areas using DeepSORT

Test Video	Actual Number of Tourists	Detected Number of Tourists	Error (%)
Video 1	168	166	1.19
Video 2	179	181	1.11
Video 3	42	42	0
Video 4	85	87	2.35
Video 5	44	43	2.27

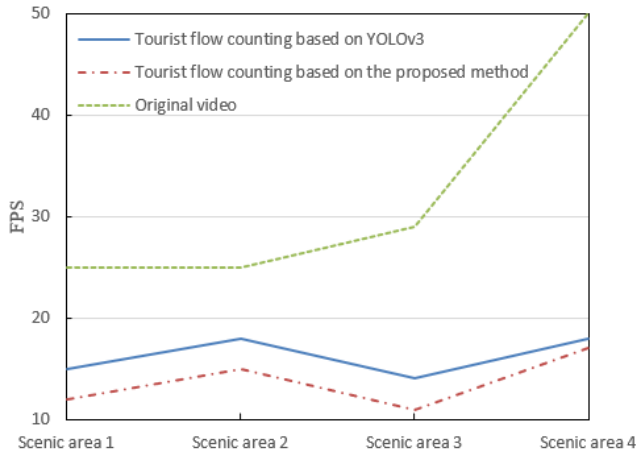
As indicated by the data presented in Table 4, notable differences exist between the proposed method and YOLOv3 in terms of video tourist flow statistical processing speed. For the video from Scenic Area 1 (1.1 seconds, 24 FPS), the proposed method required a total processing time of 159.265 seconds, with a per-frame processing time of 0.068 seconds, achieving 14 FPS. In comparison, YOLOv3 required a total processing time of 215.236 seconds, a per-frame processing time of 0.082 seconds, and achieved 11 FPS. For the video from Scenic Area 2 (514 seconds, 24 FPS), the proposed method achieved a total processing time of 717.235 seconds, a per-frame time of 0.055 seconds, and 17 FPS, whereas YOLOv3 recorded a total processing time of 812.325 seconds, a per-frame time of 0.062 seconds, and 15 FPS. For another video from Scenic Area 1 (61 seconds, 31 FPS), the proposed

method yielded a total processing time of 128.325 seconds, a per-frame time of 0.072 seconds, and 13 FPS. YOLOv3 achieved a total processing time of 159.235 seconds, a per-frame time of 0.087 seconds, and 12 FPS. For the final video from Scenic Area 1 (479 seconds, 51 FPS), the proposed method required a total processing time of 1124.235 seconds, a per-frame time of 0.051 seconds, and 18 FPS, while YOLOv3 required 1268.235 seconds, a per-frame time of 0.054 seconds, and 17 FPS. Based on the experimental results, it can be concluded that the proposed method demonstrates higher processing efficiency compared to YOLOv3 in most scenarios, particularly regarding video processing time and FPS. In the processing of videos from Scenic Areas 1 and 2, the proposed method consistently achieved shorter per-frame processing times and higher frame rates relative to YOLOv3. Although YOLOv3 exhibited relatively high processing speed in certain scenarios, especially for longer video durations, the overall results indicate that the proposed method is capable of completing equivalent video processing tasks in a shorter period while maintaining a higher FPS. This demonstrates a more optimized processing speed. These findings suggest that, although the total processing time of the proposed method may be slightly longer in specific instances, its advantages in resource utilization efficiency and processing speed confer stronger real-time performance and practical applicability for visitor flow monitoring in scenic areas. In particular, the ability to handle video streams requiring higher frame rates more effectively satisfies real-world operational demands.

In the conducted experiments, although both the improved SSD-based method and the YOLOv3-based method demonstrated effectiveness in tourist flow counting within scenic areas, neither was able to achieve real-time processing performance, particularly in terms of video processing speed. As shown in Figure 7, the red and orange lines represent the statistical processing speeds based on YOLOv3 and the proposed method, respectively, while the green line represents the original video's FPS. A considerable gap between the statistical processing speeds and the original FPS is evident across different videos. Particularly for the video from Scenic Area 4, the processing speed achieved only approximately one-third of the original frame rate, whereas for the other three videos, processing speeds reached roughly half of the original frame rate. These results indicate that although the proposed method outperformed YOLOv3 slightly in some scenarios, both methods failed to achieve real-time counting capabilities, leading to a deficiency in real-time tourist flow monitoring under certain conditions. Potential contributing factors include limitations imposed by the experimental environment as well as possible improvements yet to be realized in model training precision. Despite the inability to achieve complete real-time performance, it must be emphasized that counting accuracy remains a critical requirement for tourist flow monitoring in scenic areas. Therefore, the proposed method retains significant research value. Under conditions where strict real-time requirements are not imposed, the proposed method offers superior applicability during the detection phase, as it enables more accurate tourist counting. Particularly in complex scenic environments, the ability to provide accurate flow statistics outweighs the need for strict real-time performance. Consequently, although further optimization may be necessary to enhance processing speed, the proposed method remains a more reliable choice for practical tourist flow monitoring applications.

Table 4. Statistical processing speed of tourist flow in videos for different algorithms

Video Information	Scenic Area 1	Scenic Area 2	Scenic Area 1	Scenic Area 1
Video duration (s)	1.1	514	61	479
Video FPS	24	24	31	51
Total frames	1485	11256	1789	23512
Total tracking time (s) using the proposed method	159.265	717.235	128.325	1124.235
Processing time per frame (s) using the proposed method	0.068	0.055	0.072	0.051
FPS using the proposed method	14	17	13	18
Total tracking time (s) using YOLOv3	215.236	812.325	159.235	1268.235
Processing time per frame (s) using YOLOv3	0.082	0.062	0.087	0.054
FPS using YOLOv3	11	15	12	17

**Figure 7.** Tourist flow statistical processing speed for different algorithms in scenic area videos

5. CONCLUSION

A tourist flow monitoring and management system based on image recognition technology was proposed, with the primary objective of achieving accurate detection and statistical monitoring of tourists within scenic areas through an improved SSD network combined with multi-object tracking techniques. The research comprised two major components: first, tourist object detection based on the improved SSD network, which optimized conventional detection methods to better address the dynamic variations of tourists in complex environments; second, the integration of multi-object tracking technology to enable real-time tracking of tourists and dynamic monitoring of tourist flow within scenic areas. The improved SSD network enhanced detection precision, maintaining high accuracy even under conditions of dense tourist presence and complex backgrounds, while the multi-object tracking module ensured efficient tracking and statistical counting of tourist flow.

From an overall perspective, the system proposed in this study demonstrates substantial practical application value. Accurate monitoring of tourist flow is enabled, assisting scenic area management in enhancing visitor safety measures, optimizing resource allocation, and improving operational efficiency. Furthermore, by integrating image recognition and tracking technologies, a novel and forward-looking solution is provided for the application of computer vision techniques in scenic area management. The findings also highlight the broader applicability of the proposed technology in other public environments such as shopping malls and transportation hubs, indicating significant potential for widespread adoption.

However, despite the encouraging research outcomes, certain limitations of the system remain. First, although the improved SSD network significantly enhances detection precision, instances of incomplete or incorrect detection may still occur under extreme conditions, particularly in highly congested environments or against complex backgrounds. Second, the employed multi-object tracking technology demands considerable computational resources, especially in scenarios involving high tourist density, resulting in substantial computational overhead. Further optimization of the algorithms is therefore required to reduce computational demands. Additionally, the implementation of the system raises critical concerns regarding data privacy and security. Ensuring the protection of tourists' personal information remains one of the major challenges that must be addressed in future work. To overcome these limitations, future research could focus on further optimizing object detection and tracking algorithms to enhance system robustness and efficiency in high-density tourist scenarios. In particular, emerging technologies in deep learning, such as Transformer architectures and self-supervised learning, may provide promising avenues for advancing existing algorithms. Moreover, with continuous technological progress, future systems could be expanded to broader applications, including large commercial complexes, sporting events, and transportation hubs. Regarding data privacy protection, research efforts should emphasize the development of mechanisms that balance the efficient operation of the system with the safeguarding of personal information. Techniques such as encryption methods and differential privacy approaches are recommended to ensure the security and confidentiality of tourist data.

REFERENCES

- [1] AlKahtani, S.J.H., Xia, J.C., Veenendaaland, B., Caulfield, C., Hughes, M. (2015). Building a conceptual framework for determining individual differences of accessibility to tourist attractions. *Tourism Management Perspectives*, 16: 28-42. <https://doi.org/10.1016/j.tmp.2015.05.002>
- [2] Jamshidi, M.J., Barakpour, N. (2023). Analyzing synergistic effects of tourist attractions on sustainable development of neighborhoods with emphasis on urban smart growth principles. *International Journal of Sustainable Development & World Ecology*, 30(8): 949-963. <https://doi.org/10.1080/13504509.2023.2231885>
- [3] Rogowski, M. (2020). Monitoring system of tourist traffic (MSTT) for tourists monitoring in mid-mountain national park, SW Poland. *Journal of Mountain Science*, 17(8): 2035-2047. <https://doi.org/10.1007/s11629-019->

- 5965-y
- [4] Katircioglu, S., Cizreliogullari, M.N., Katircioglu, S. (2019). Estimating the role of climate changes on international tourist flows: Evidence from Mediterranean Island States. *Environmental Science and Pollution Research*, 26: 14393-14399. <https://doi.org/10.1007/s11356-019-04750-w>
 - [5] Ziger-Korn, N. (2022). In search of the management model of the nature reserve as a sustainable tourist destination. *Amazonia Investiga*, 11(60): 145-149. <https://doi.org/10.34069/AI/2022.60.12.15>
 - [6] Skowronek, E., Brzezińska-Wójcik, T., Stasiak, A. (2023). How to effectively build the image of an emerging destination. *Quaestiones Geographicae*, 42(4): 113-156. <https://doi.org/10.14746/quageo-2023-0034>
 - [7] Backes, A. R., Martinez, A.S., Bruno, O.M. (2011). Texture analysis using graphs generated by deterministic partially self-avoiding walks. *Pattern Recognition*, 44(8): 1684-1689. <https://doi.org/10.1016/j.patcog.2011.01.018>
 - [8] Song, Z., Lu, J. (2022). Early warning and management method of abnormal performance of tourist scenic spots assisted by image recognition technology. *Discrete Dynamics in Nature and Society*, 2022(1): 6217530. <https://doi.org/10.1155/2022/6217530>
 - [9] Huang, W., Zhu, S., Yao, X. (2021). Destination image recognition and emotion analysis: Evidence from user-generated content of online travel communities. *The Computer Journal*, 64(3): 296-304. <https://doi.org/10.1093/comjnl/bxaa064>
 - [10] Otoo, F.E., Kim, S., Choi, Y. (2021). Developing a multidimensional measurement scale for diaspora tourists' motivation. *Journal of Travel Research*, 60(2): 417-433. <https://doi.org/10.1177/0047287519899990>
 - [11] Qin, S., Man, J., Wang, X., Li, C., Dong, H., Ge, X. (2019). Applying big data analytics to monitor tourist flow for the scenic area operation management. *Discrete Dynamics in Nature and Society*, 2019(1): 8239047. <https://doi.org/10.1155/2019/8239047>
 - [12] Gao, Z., Zhang, J., Xu, Z., Zhang, X., et al. (2020). Method of predicting passenger flow in scenic areas considering multisource traffic data. *Sensors & Materials*, 32(11): 3907-3921. <https://doi.org/10.18494/SAM.2020.2970>
 - [13] Lu, H., Zhang, J., Xu, Z., Shi, R., Wang, J., Xu, S. (2021). Prediction of tourist flow based on multi-source traffic data in scenic spot. *Transactions in GIS*, 25(2): 1082-1103. <https://doi.org/10.1111/tgis.12724>
 - [14] Zachevitskiy, A.A., Krichigin, A.V., Mavrychev, E.A. (2013). On statistical properties of the detection range of a moving target. *Journal of Communications Technology and Electronics*, 58: 128-134. <https://doi.org/10.1134/S1064226913010105>
 - [15] Yoon, K., Song, Y.M., Jeon, M. (2018). Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views. *IET Image Processing*, 12(7): 1175-1184. <https://doi.org/10.1049/iet-ipr.2017.1244>
 - [16] Liang, M., Kim, D.Y., Kai, X. (2015). Multi-Bernoulli filter for target tracking with multi-static Doppler only measurement. *Signal Processing*, 108: 102-110. <https://doi.org/10.1016/j.sigpro.2014.09.013>
 - [17] Kwon, S., Park, T. (2020). Channel-based network for fast object detection of 3D LiDAR. *Electronics*, 9(7): 1122. <https://doi.org/10.3390/electronics9071122>