International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 1, February, 2026, pp. 185-197
Journal homepage: http://iieta.org/journals/ts

A Framework for Tourist Behavior Recognition and Interest Region Generation in Digital Cultural Tourism Platforms

Fangzhou Fan

School of Economics and Management, Shangqiu Polytechnic, Shangqiu 476000, China

Corresponding Author Email: ffzz698@126.com

https://doi.org/10.18280/ts.430114

Received: 6 September 2025
Revised: 13 December 2025
Accepted: 15 January 2026
Available online: 28 February 2026

Keywords:
digital cultural tourism platforms, tourist behavior recognition, interest region generation, multi-task learning, spatiotemporal graph convolution, attention mechanism

ABSTRACT

The intelligent transformation of digital cultural tourism platforms has imposed increasing demands on accurate tourist behavior analysis and efficient identification of regions of interest. Existing image processing approaches typically treat tourist behavior recognition and interest region generation as independent tasks, resulting in suboptimal feature utilization and neglect of semantic correlations during interest region generation. Such limitations hinder their applicability in complex cultural tourism scenarios. To address these challenges, an end-to-end multi-task learning framework, termed Tourist Behavior-Aware Interest Region Network (TBA-IRNet), was proposed to achieve the joint optimization of tourist behavior recognition and interest region generation. In the framework, multi-scale spatiotemporal features are extracted through a shared backbone network, while an interactive attention mechanism is employed to enable bidirectional information exchange between the two tasks. A behavior-aware weighting strategy is introduced, in which behavior categories and dwell time are integrated to construct an interest intensity map. Furthermore, a bidirectional guidance mechanism combining spatiotemporal graphs and region-level attention is designed to enhance semantic modeling capability. The framework is fully differentiable, thereby supporting end-to-end joint training and optimization. Experimental results obtained on a self-constructed cultural tourism dataset demonstrate that superior performance is achieved in both tourist behavior recognition accuracy and interest region generation precision, compared with existing methods. The proposed framework advances the methodology of multi-task collaborative learning and provides a novel paradigm for cross-task inference in image processing. In addition, it offers a practical and efficient technical solution for intelligent management and service optimization in digital cultural tourism platforms.

1. INTRODUCTION

With the continuous advancement of digital transformation in the cultural tourism industry, massive volumes of video data generated by surveillance systems in scenic areas have provided rich data support for tourist behavior recognition and interest region generation [1-3]. Accurate recognition of tourist behaviors [4, 5] and automatic identification of interest regions [6, 7] are essential for enabling optimization of scenic area management, service enhancement, and resource allocation, and have become central requirements for the intelligent development of digital cultural tourism platforms. Image processing techniques have been widely applied to tourist analysis tasks in cultural tourism scenarios. However, in most existing approaches, tourist behavior recognition and interest region generation are treated as two independent tasks. The semantic correlations between these tasks have not been sufficiently explored, resulting in low feature utilization efficiency. Consequently, such methods struggle to adapt to the practical challenges of cultural tourism environments, which are characterized by high crowd density, diverse behavioral patterns, and complex backgrounds [8, 9]. These limitations have constrained the level of intelligence achievable in digital cultural tourism platforms.

Despite recent progress, several critical challenges remain unresolved in existing studies, which can be summarized into four main aspects. First, task decoupling remains prevalent. In most methods, tourist behavior recognition and interest region generation are processed independently [10], and the intrinsic semantic relationships between the two tasks are not effectively exploited. This leads to feature redundancy and information loss, thereby limiting overall model performance. Second, significant deficiencies exist in current interest region generation methods. Traditional approaches are primarily based on crowd density for region delineation [11, 12], while critical semantic factors such as behavior categories and dwell time are often neglected. As a result, the generated regions fail to accurately reflect tourists' true interest preferences and lack practical applicability. Third, limitations are observed in the application of graph modeling and attention mechanisms. Existing graph convolution-based methods for behavior recognition have not effectively incorporated region-level semantic information [13], and attention mechanisms are typically implemented in a unidirectional guidance manner

[14]. As a consequence, bidirectional collaborative enhancement between behavior recognition and interest region generation has not been achieved, making it difficult to accurately model tourist interactions and region associations in complex scenarios. Fourth, insufficient end-to-end differentiability remains a critical issue. In some existing models, key components involve non-differentiable operations [15, 16], preventing full-process joint optimization. This limitation reduces both training efficiency and the potential for performance improvement.

The objective of this study is to develop a unified image processing framework tailored to cultural tourism scenarios, in which the task boundary between tourist behavior recognition and interest region generation is eliminated. Through joint optimization of these two tasks, significant improvements are achieved in the accuracy of tourist behavior recognition as well as in the precision and efficiency of interest region generation. To this end, the main contributions are summarized. First, a multi-task collaborative image processing framework is proposed. By leveraging shared feature extraction and an interactive attention mechanism, bidirectional information flow between tourist behavior recognition and interest region generation is enabled, thereby improving feature utilization efficiency. Second, a behavior-aware interest region generation method is designed. Behavioral semantics and temporal information are integrated to overcome the limitations of conventional methods that rely solely on crowd density, thereby enhancing the semantic validity of the generated regions. Third, a bidirectional guidance mechanism integrating spatiotemporal graphs and region-level attention is constructed. This mechanism improves contextual modeling capability in tourist behavior recognition while enhancing the semantic accuracy of interest region generation, resulting in improved adaptability to complex cultural tourism environments. Fourth, a fully differentiable end-to-end training framework is established. Through the joint optimization of multi-task loss functions, overall model performance and generalization capability are ensured, enabling effective deployment in real-world cultural tourism scenarios.

The remainder of this study is organized below. In Section 2, the overall architecture of the proposed Tourist Behavior-Aware Interest Region Network (TBA-IRNet) framework is described in detail, including the technical design of its core modules. Section 3 presents comparative experiments and ablation studies conducted on a self-constructed cultural tourism dataset to validate the effectiveness and superiority of the proposed framework. In Section 4, the experimental results are analyzed in depth, and the strengths and limitations of the model are discussed, along with potential directions for future research. Finally, Section 5 summarizes the study and reiterates its primary contributions and practical significance.

2. PROPOSED METHOD

2.1 Overall framework architecture

The proposed TBA-IRNet adopts an encoder–decoder architecture, in which surveillance video streams from cultural tourism scenarios are taken as input. The overall design is centered on shared feature extraction, parallel multi-task inference, bidirectional attention interaction, and end-to-end optimization. The encoder is constructed based on a three-dimensional convolutional backbone network, through which multi-scale spatiotemporal features are extracted from consecutive video frames. These features provide a unified and highly expressive representation for subsequent dual-task inference, thereby mitigating feature redundancy and information fragmentation. The spatiotemporal feature maps generated by the encoder are simultaneously fed into two parallel task branches, which are responsible for tourist behavior recognition and interest region generation, respectively. These branches are not independently operated; instead, bidirectional information exchange is enabled through a collaborative attention mechanism. In this manner, tourist behavior recognition is enhanced by incorporating region-level semantic information to improve classification accuracy, while interest region generation is guided to focus on behaviorally active tourist groups. As a result, mutual reinforcement between the two tasks is achieved. The entire framework is designed to be fully differentiable, allowing joint training and optimization of the shared backbone and both task branches through a unified multi-task loss function. End-to-end inference is thus realized, spanning from raw video input to the outputs of behavior categories and interest regions. This design effectively improves overall model performance and inference efficiency, while ensuring robustness in complex cultural tourism environments characterized by high crowd density and diverse behavioral patterns.

2.2 Detailed description of core modules

2.2.1 Multi-task collaborative interactive attention mechanism

To eliminate the task separation between tourist behavior recognition and interest region generation and to achieve synergistic enhancement between the two tasks, a multi-task collaborative interactive attention module is designed. The core innovation of this module lies in the construction of a bidirectional information flow mechanism between the two tasks. In contrast to conventional unidirectional attention guidance schemes, mutual feature enhancement and dynamic optimization are enabled, while full differentiability is preserved to support end-to-end joint training. The overall architecture of the module is illustrated in Figure 1. Based on the spatiotemporal feature maps extracted by the shared backbone network, dedicated interaction pathways are established between the tourist behavior recognition branch and the interest region generation branch. No additional redundant feature extraction modules are introduced, thereby improving feature utilization efficiency and effectively addressing the fundamental limitation of task-wise information fragmentation observed in existing methods.

The bidirectional information flow mechanism constitutes the core technical innovation of this module, and its implementation is described below. When behavioral semantic information is transmitted from the tourist behavior recognition branch to the interest region generation branch, the initial feature representations of the behavior recognition branch are first subjected to a linear projection. Through this transformation, high-dimensional features are mapped into attention weights that are dimensionally aligned with the feature maps of the interest region generation branch. The projection process can be formulated as:

$$F_{act\_proj}=W_{act}\cdot F_{act}+b_{act} \tag{1}$$

where, F_{act} denotes the initial feature representation of the tourist behavior recognition branch, W_{act} and b_{act} represent the learnable projection weight matrix and bias vector, respectively, and F_{act_proj} denotes the resulting attention weight map after projection. Subsequently, a Sigmoid activation function is applied to normalize the attention weights, yielding attention coefficients within the range [0, 1]. These coefficients are then multiplied element-wise with the feature maps of the interest region generation branch. In this manner, behavioral semantics are effectively injected into the region generation process, enabling the generated interest regions to focus on behaviorally active tourist groups and thereby improving the semantic accuracy of interest region generation.

Conversely, when contextual information is transmitted from the interest region generation branch to the tourist behavior recognition branch, candidate region features are first extracted from the interest region generation branch. These features are similarly transformed via a linear projection to obtain context-constrained vectors. The resulting representations are then integrated into the node feature aggregation process of the graph convolutional network within the behavior recognition branch, thereby enhancing the contextual coherence of node features. The projection and fusion process can be expressed as:

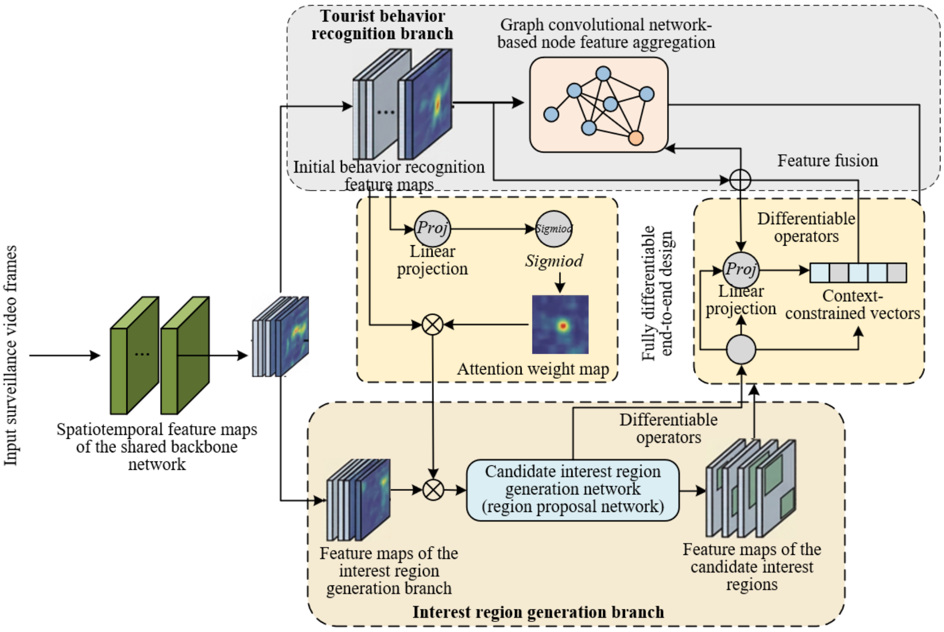

Tourist behavior recognition branch
Graph convolutional network-based node feature aggregation
Initial behavior recognition feature maps
Feature fusion
Input surveillance video frames
Spatiotemporal feature maps of the shared backbone network
Proj
Linear projection
Sigmiod
Sigmiod
Attention weight map
Fully differentiable end-to-end design
Differentiable operators
Proj
Linear projection
Context-constrained vectors
Differentiable operators
Candidate interest region generation network (region proposal network)
Feature maps of the interest region generation branch
Feature maps of the candidate interest regions
Interest region generation branch

Figure 1. Network architecture of the multi-task collaborative bidirectional interactive attention mechanism

$$F_{node\_att}=F_{node}+\alpha\cdot(W_{ir}\cdot F_{ir\_cand}+b_{ir}) \qquad (2)$$

where, F_{ir_cand} denotes the candidate interest region features, W_{ir} and b_{ir} represent the corresponding projection parameters, α is a weighting coefficient controlling the contribution of contextual fusion, F_{node} denotes the initial node features in the graph convolutional network, and F_{node_att} represents the node features after integration of region-level contextual information. This process improves the accuracy of tourist behavior classification.

The fully differentiable design of the module constitutes a critical innovation for enabling end-to-end joint optimization. To address the issue of training discontinuities caused by non-differentiable operations in conventional attention mechanisms, all components of the proposed module—including feature projection, attention weight computation, and feature fusion—are implemented using differentiable operators. Linear projections are realized through learnable parameter matrices. Attention weight normalization is performed using the Sigmoid differentiable activation function. Feature fusion is achieved through element-wise multiplication and linear combination, both of which are differentiable operations. As a result, gradient backpropagation is preserved throughout the entire interaction process. This design enables the shared backbone network, the tourist behavior recognition branch, the interest region generation branch, and the interactive attention module to be jointly trained under a unified loss function. Consequently, information loss during transmission is mitigated, while training stability and computational efficiency are improved. The synergistic benefits of the dual-task framework are thereby fully realized.

2.2.2 Interest region generation method with behavior-aware weighting

To address the fundamental limitation of conventional interest region generation methods, which rely solely on crowd density while neglecting behavioral semantics and temporal information [17, 18], an interest region generation method with behavior-aware weighting is proposed. The core innovation of this approach lies in the integration of tourist behavior categories, dwell time, and individual-scale features into the modeling of interest intensity, thereby enabling semantically meaningful and precise generation of interest regions. Meanwhile, full differentiability is preserved throughout the generation process, ensuring compatibility with end-to-end joint training. Specifically, the proposed method constructs an initial interest intensity map that accurately reflects tourists' interest preferences through three key components: behavior weight assignment, temporal decay factor design, and adaptive Gaussian kernel construction. These components provide accurate semantic guidance for subsequent region proposal generation. A schematic illustration of the interest region intensity map generation process based on behavior-aware weighting is presented in Figure 2.

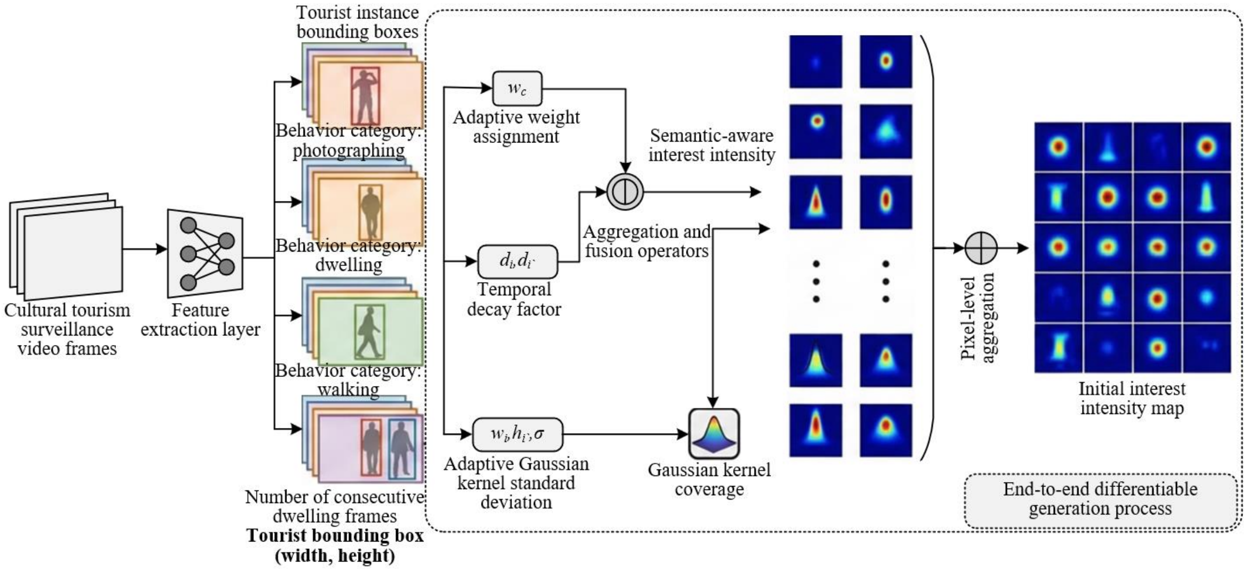

Tourist instance bounding boxes
Behavior category: photographing
Behavior category: dwelling
Behavior category: walking
Number of consecutive dwelling frames
Tourist bounding box (width, height)
Cultural tourism surveillance video frames
Feature extraction layer
w_c
Adaptive weight assignment
$d_i, d_{i'}$
Temporal decay factor
w_i, h_i, σ
Adaptive Gaussian kernel standard deviation
Semantic-aware interest intensity
Aggregation and fusion operators
Gaussian kernel coverage
Pixel-level aggregation
Initial interest intensity map
End-to-end differentiable generation process

Figure 2. Schematic diagram of the interest region intensity map generation process based on behavior-aware weighting

The behavior weight assignment strategy is designed to capture the semantic differences among tourist behaviors in cultural tourism scenarios. An adaptive weighting mechanism is introduced, in which weight levels are determined according to the indicative strength of each behavior with respect to interest regions. Specifically, photographing behavior is assigned the highest weight, followed by dwelling behavior, while walking behavior is assigned the lowest weight. The weight values are dynamically calibrated through five-fold cross-validation to ensure adaptability across different cultural tourism environments, thereby enhancing both rationality and generalization capability. A temporal decay factor is introduced to distinguish between short-term and long-term dwell behaviors, thereby preventing transient passersby from being incorrectly identified as core contributors to interest regions. The decay function is constructed based on the number of consecutive dwelling frames for each tourist, and is defined as:

$$d_i'=1+\log(1+d_i) \tag{3}$$

where, d_i denotes the number of consecutive dwelling frames of the i-th tourist prior to the current frame, and d_i' represents the normalized temporal decay factor. The logarithmic function effectively suppresses the imbalance caused by rapid increases in frame counts, allowing the contribution of long-term dwellers to be reasonably amplified while mitigating the influence of short-term visitors. Furthermore, an adaptive Gaussian kernel is designed to address the mismatch between fixed kernel sizes and varying tourist scales in conventional methods. The standard deviation of the Gaussian kernel is dynamically adjusted according to the size of each tourist's bounding box, and is computed as:

$$\sigma=0.3\times\max(w_i,h_i) \tag{4}$$

where, w_i and h_i denote the width and height of the i-th tourist bounding box, respectively. This design ensures that the spatial coverage of the Gaussian kernel is well aligned with the actual scale of each tourist, thereby avoiding spatial bias introduced by fixed kernel sizes and improving the spatial accuracy of the generated interest intensity map.

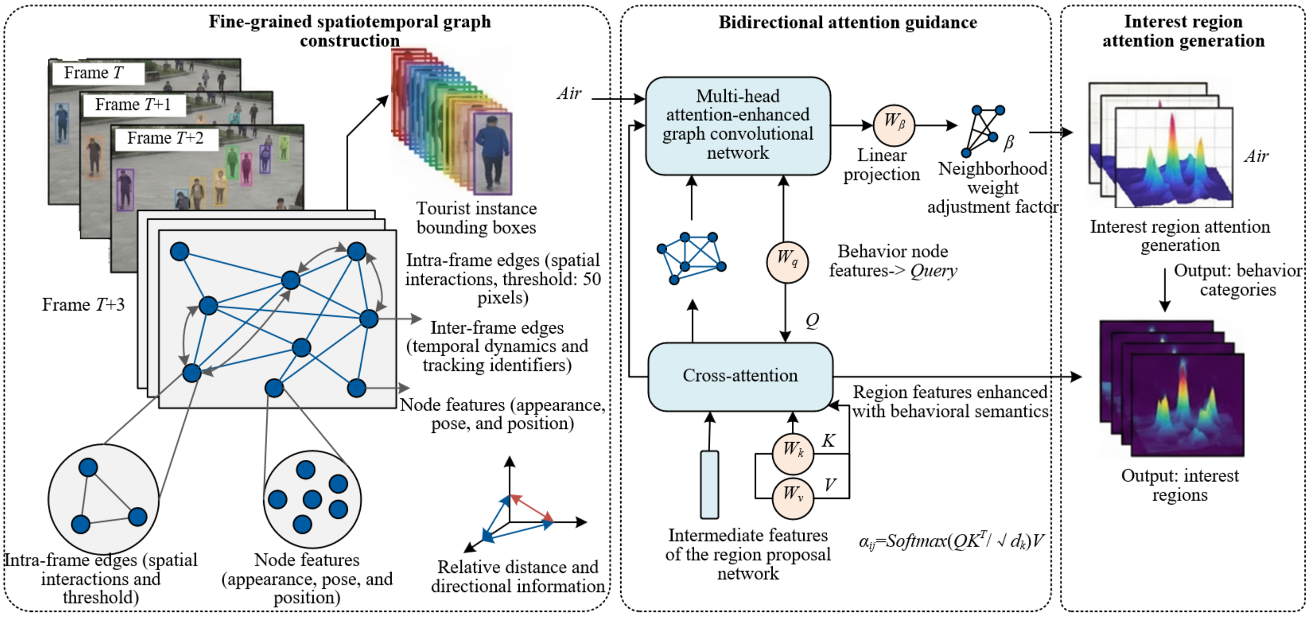
Based on the aforementioned design, the interest intensity map is generated by aggregating the weighted Gaussian kernels of all tourists. The weighted Gaussian kernel corresponding to an individual tourist is defined as:

$$H_i(x,y)=w_c\cdot d'_i\cdot \exp\left(-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma^2}\right) \tag{5}$$

where, w_c denotes the weight associated with the tourist behavior category, (x_i,y_i) represents the center coordinates of the tourist bounding box, and (x,y) denotes the pixel coordinates in the image. The initial interest intensity map for the entire image is obtained by summing the weighted Gaussian kernels of all tourists at the pixel level. The resulting pixel values are positively correlated with the degree of tourist interest at corresponding spatial locations, thereby enabling precise characterization of the interest distribution in the scene based on behavioral semantics. The entire generation process is implemented using differentiable operations, ensuring seamless integration with other components of the framework during joint training. The generated interest intensity map not only encodes crowd density information but also incorporates behavioral semantics and temporal characteristics. Consequently, the semantic accuracy and practical relevance of subsequent interest region proposals are significantly improved.

2.2.3 Bidirectional guidance mechanism of spatiotemporal graphs and region-level attention

To accurately model the spatiotemporal interaction characteristics of tourists in cultural tourism scenarios, while achieving semantic alignment between tourist behavior recognition and interest region generation, a bidirectional guidance mechanism integrating spatiotemporal graphs and region-level attention is proposed. The core innovation of this mechanism lies in the construction of a fine-grained spatiotemporal graph model and a bidirectional attention interaction pathway. This design effectively addresses the limitations of conventional graph-based methods, in which region-level semantic information is insufficiently incorporated, and attention mechanisms are typically constrained to unidirectional guidance. As a result, mutual enhancement between the two tasks is achieved, leading to improved overall performance. Specifically, the refined spatiotemporal graphs constructed and a graph convolutional framework enhanced by multi-head attention designed provide a robust feature foundation for bidirectional guidance, thereby ensuring both the precision and effectiveness of the guidance process.



Fine-grained spatiotemporal graph construction
Frame *T*
Frame *T*+1
Frame *T*+2
Frame *T*+3
Tourist instance bounding boxes
Intra-frame edges (spatial interactions, threshold: 50 pixels)
Inter-frame edges (temporal dynamics and tracking identifiers)
Node features (appearance, pose, and position)
Intra-frame edges (spatial interactions and threshold)
Node features (appearance, pose, and position)
Relative distance and directional information
Air
Bidirectional attention guidance
Multi-head attention-enhanced graph convolutional network
W_β
Linear projection
β
Neighborhood weight adjustment factor
W_q
Behavior node features-> *Query*
Q
Cross-attention
Region features enhanced with behavioral semantics
W_k
K
W_v
V
Intermediate features of the region proposal network
α_{ij}=*Softmax*(QK^T/ √ d_k)*V*
Interest region attention generation
Air
Interest region attention generation
Output: behavior categories
Output: interest regions

Figure 3. Schematic diagram of spatiotemporal graph construction and bidirectional guidance via region-level attention

As illustrated in Figure 3, the proposed mechanism for spatiotemporal graph construction and bidirectional guidance via region-level attention is presented. The fine-grained construction of the spatiotemporal graphs is centered on the innovative design of both node and edge features, thereby overcoming the limitations of insufficient semantic expressiveness in conventional graph structures. Node features are constructed using a multi-dimensional feature concatenation strategy, in which tourist appearance features, normalized pose keypoint coordinates, and bounding box positional features are integrated to comprehensively capture individual attributes and spatial information. Among these, pose keypoint features are incorporated to facilitate the representation of subtle behavioral variations, thereby enhancing the semantic expressiveness of node representations for tourist behavior recognition. Edge features are designed to jointly capture spatial interactions and temporal dynamics. Intra-frame edges are established based on a Euclidean distance threshold between tourists, which is determined through statistical analysis of the dataset and set to 50 pixels. Specifically, an intra-frame edge is created when the Euclidean distance between two tourists is less than this threshold, enabling the modeling of spatial interaction relationships. Inter-frame edges are constructed by associating nodes of the same tourist across consecutive frames using tracking identifiers, thereby capturing temporal dynamics at the individual level. In addition, relative distance and directional information are incorporated into edge features to further enrich interaction semantics. On this basis, a multi-head attention mechanism is introduced into the spatial graph convolutional operation. Four attention heads are employed, each of which adaptively learns neighborhood weights under different interaction patterns. The attention weight is computed as:

$$\alpha_{ij}=\frac{\exp(\text{LeakyReLU}(W_q h_i^T W_k h_j))}{\sum_{k\in N(i)}\exp(\text{LeakyReLU}(W_q h_i^T W_k h_k))} \qquad (6)$$

where, α_{ij} denotes the attention weight assigned by node i to its neighboring node j; h_i and h_j represent the features of nodes i and j, respectively; W_q and W_k are learnable parameter matrices, and $N(i)$ denotes the set of neighboring nodes of node i. This design addresses the limitation of fixed neighborhood weights in conventional spatiotemporal graph convolution methods and enhances the model's capability to capture complex group behaviors, such as crowd aggregation and collective attention patterns.

The bidirectional guidance mechanism constitutes the core innovation of this module, through which mutual constraint and optimization between tourist behavior recognition and interest region generation are achieved. When region-level attention is used to guide tourist behavior recognition, the attention map generated by the interest region generation branch is first extracted and then transformed into a neighborhood weight adjustment factor for graph convolution through a linear projection. The projection process is defined as $\beta=W_\beta A_{ir}+b_\beta$, where A_{ir} denotes the interest region attention map, W_β and b_β represent projection parameters, and β denotes the adjustment factor. This factor is multiplied with the neighborhood weights in the graph convolutional operation, thereby enabling node feature aggregation to focus more effectively on tourist interactions within interest regions, which leads to improved accuracy in tourist behavior recognition. Conversely, when behavioral features are used to guide interest region generation, node features from the tourist behavior recognition branch are projected into the query vector Q, while intermediate features from the region proposal network are treated as key K and value V. Cross-attention is then employed to incorporate behavioral semantics into region generation. The cross-attention operation is defined as:

$$\text{Att}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (7)$$

where, d_k denotes the dimensionality of the key vectors.

Through this operation, behavioral semantics are effectively embedded into region-level features, thereby improving the confidence prediction of candidate regions and filtering out redundant regions associated with behaviorally inactive areas. As a result, the generated interest regions are ensured to be highly consistent with tourist behavior semantics. Through this bidirectional guidance mechanism, deep interaction between spatiotemporal graph features and region-level attention features is achieved. Consequently, tourist behavior recognition and interest region generation are jointly optimized, leading to substantial improvements in model adaptability in complex cultural tourism scenarios.

2.2.4 Fully differentiable end-to-end optimization framework

To enable coordinated training and performance enhancement across all modules, and to address key limitations of existing methods—such as partial non-differentiability, suboptimal loss function design, and limited generalization capability [19, 20]—a fully differentiable end-to-end optimization framework is proposed. The core innovation lies in the design of a multi-task joint loss function tailored to cultural tourism scenarios. In addition, a novel consistency loss and an improved density map loss are introduced, together with task-specific training strategies, to ensure full gradient backpropagation throughout the pipeline and to achieve joint optimization of overall model performance. Within this framework, the shared backbone network, dual-task branches, and all innovative modules are integrated into a unified training system. Training discontinuities are thereby eliminated, allowing the synergistic advantages of all components to be fully exploited.

The design of the multi-task loss function constitutes the central component of the end-to-end optimization framework. Through careful adaptation and refinement of individual loss terms, balanced learning and semantic alignment across tasks are achieved. Considering the characteristics of cultural tourism datasets, five-fold cross-validation is employed to optimize the weighting coefficients of each loss term. The final weights are determined as λ_1=0.2, λ_2=0.1, λ_3=0.3, λ_4=0.2, λ_5=0.1, and λ_6=0.1. This configuration prioritizes the core tasks of tourist behavior recognition and interest region generation, while maintaining an appropriate balance with auxiliary tasks such as detection and pose estimation. A consistency loss is introduced to enforce semantic alignment between tourist behavior recognition and interest region generation. By computing the Kullback–Leibler (KL) divergence between the graph attention weights of the behavior recognition branch and the attention maps of the interest region generation branch, the discrepancy between the feature distributions of the two tasks is quantified. The formulation is expressed as follows:

$$L_{cons}=\sum_{i=1}^{N} p_i \log\left(\frac{p_i}{q_i}\right) \tag{8}$$

where, p_i denotes the attention weight distribution from the tourist behavior recognition branch, and q_i denotes the attention weight distribution from the interest region generation branch. This loss encourages consistency between the feature distributions of the two tasks, thereby promoting bidirectional collaborative enhancement. An improved density map loss is further introduced to address the limitation of conventional density-based losses, which typically ignore semantic information. The ground-truth interest density map is constructed by combining manually annotated interest regions with trajectory clustering of tourists, followed by Gaussian smoothing. The L2 distance between the predicted interest intensity map and the ground-truth density map is then computed as:

$$L_{density}=\frac{1}{H\times W}\sum_{x=1}^{H}\sum_{y=1}^{W}(S(x,y)-\widehat{S}(x,y))^2 \tag{9}$$

where, H and W denote the height and width of the image, respectively, $S(x,y)$ denotes the ground-truth pixel value of the interest density map, and $\widehat{S}(x,y)$ denotes the predicted pixel value of the interest density map. By integrating all loss components, the overall loss function is formulated as:

$$L=\lambda_1 L_{det}+\lambda_2 L_{pose}+\lambda_3 L_{act}+\lambda_4 L_{region}+\lambda_5 L_{density}+\lambda_6 L_{cons} \tag{10}$$

Further optimization of the training strategy is performed to enhance model generalization and to better align with the characteristics of cultural tourism surveillance video scenarios. The backbone network is initialized using pretrained weights, and only the higher-level convolutional layers are fine-tuned. In this manner, the general feature extraction capability of lower layers is preserved, thereby mitigating overfitting and improving training efficiency. The data augmentation strategy is specifically adapted to the complexity of cultural tourism environments. Techniques, including random cropping, horizontal flipping, color jittering, and temporal frame shuffling, are employed to increase data diversity. These operations effectively alleviate the limitations imposed by the relatively small scale of the self-constructed dataset and improve the model's robustness to variations in illumination, viewpoint, and crowd density. The entire training process is conducted under a unified optimization objective, in which all modules are designed to be fully differentiable. Consequently, gradients can be propagated seamlessly throughout the network, enabling coordinated training of the shared backbone network and dual-task branches. This design significantly enhances overall model performance and generalization capability, thereby ensuring suitability for real-world deployment in complex cultural tourism scenarios.

2.3 Inference pipeline

The inference pipeline of TBA-IRNet is characterized by its fully differentiable end-to-end design, enabling efficient and coordinated inference from surveillance video stream input to the outputs of tourist behavior categories and interest regions. No intermediate manual intervention or feature transformation is required, thereby significantly improving both inference efficiency and accuracy. The inference process begins with the input of continuous frames from cultural tourism surveillance video streams. Multi-scale spatiotemporal features are first extracted from the input frames by the shared backbone network, producing a spatiotemporal feature map with strong semantic representation capability. The feature map provides a unified feature foundation for both task branches, thereby avoiding feature redundancy and repeated extraction, and establishing the basis for efficient inference.

Based on the extracted spatiotemporal features, tourist detection and pose estimation are first performed. Structured tourist instances are then generated, including bounding boxes, pose keypoints, appearance features, and unique tracking identifiers for each individual. These structured representations provide fine-grained individual-level information for subsequent dual-task inference. The pipeline then proceeds to the parallel inference stage of the two tasks. In the tourist behavior recognition branch, a fine-grained spatiotemporal graph is constructed based on the structured tourist instances. Behavior classification is performed using a multi-head attention-enhanced spatiotemporal graph convolutional network, yielding behavior categories for each tourist. In the interest region generation branch, an initial interest intensity map is generated based on behavior-aware weighting, dwell time, and adaptive Gaussian kernels. This map is combined with the spatiotemporal feature map and processed through a differentiable region proposal network to produce candidate interest regions. Throughout this process, real-time bidirectional information exchange between the two branches is achieved via the collaborative attention mechanism, without requiring additional feature mapping steps. Finally, optimal interest regions are selected through non-maximum suppression, and the results are output together with the corresponding tourist behavior recognition results. The entire pipeline is implemented using differentiable operations, thereby eliminating inference discontinuities caused by non-differentiable components in conventional approaches. As a result, smooth feature propagation across modules is ensured, inference latency is significantly reduced, and accuracy is maintained through the bidirectional guidance mechanism. This design enables a synergistic improvement in both efficiency and precision, making the framework well-suited for real-time processing requirements in complex cultural tourism surveillance scenarios.

3. EXPERIMENTS AND ANALYSIS

3.1 Experimental setup

To comprehensively evaluate the effectiveness, superiority, and robustness of the proposed TBA-IRNet framework, a series of comparative experiments, ablation studies, and parameter sensitivity analyses were conducted, taking into account the practical requirements of cultural tourism scenarios. All experiments were performed on a self-constructed cultural tourism dataset, and the reported results were obtained by averaging over three independent runs to ensure reliability and statistical significance.

Table 1. Statistics of the self-constructed cultural tourism dataset

Scenic Area Type	Number of Videos	Total Frames ($\times 10^4$)	Tourist Instances ($\times 10^4$)	Number of Behavior Categories	Number of Interest Region Annotations	Training Set Ratio	Validation Set Ratio	Test Set Ratio
Natural landscapes	350	18	7.2	5	1,260	80%	10%	10%
Cultural heritage sites	320	17	6.8	5	1,180	80%	10%	10%
Theme parks	330	15	6.0	5	1,320	80%	10%	10%
Total	1,000	50	20.0	5	3,760	80%	10%	10%

Given that existing public datasets lack scenario-specific characteristics of cultural tourism environments and therefore cannot adequately support the tasks of tourist behavior recognition and interest region generation, a dedicated dataset covering multiple types of cultural tourism scenarios was constructed. The detailed statistical information of the dataset is summarized in Table 1. The dataset consisted of surveillance videos collected from three representative categories of cultural tourism environments, including natural landscapes, cultural heritage sites, and theme parks. In total, 1,000 video sequences and 500,000 image frames were included, comprising approximately 200,000 annotated tourist instances. The annotations encompassed tourist bounding boxes, 17 pose keypoints, five core behavior categories (walking, dwelling, photographing, conversing, and grouping), and 3,760 interest regions. The dataset was partitioned into training, validation, and test sets with a ratio of 8:1:1. The annotation process strictly followed established protocols in the field of image processing and was conducted by three professional annotators. Annotation consistency was verified using the Cohen's Kappa coefficient, with a value exceeding 0.85. The dataset is characterized by high crowd density, complex backgrounds, and diverse behavioral patterns, thereby effectively simulating real-world cultural tourism scenarios. These characteristics address the limitations of existing public datasets and provide a solid data foundation for validating the effectiveness of the proposed framework.

The experimental hardware environment consisted of eight NVIDIA V100 graphics processing units. The software environment was implemented using the PyTorch 1.12 framework, with Compute Unified Device Architecture (CUDA) 11.6 employed for computational acceleration, and Ubuntu 20.04 Long-Term Support (LTS) used as the operating system. The training parameters were determined through multiple rounds of tuning and cross-validation. The AdamW optimizer was adopted, with an initial learning rate of 1*e*-4 and a weight decay coefficient of 1*e*-5, which effectively mitigates overfitting. The model was trained for a total of 50 epochs, with a batch size of 32. A cosine annealing strategy was applied to dynamically adjust the learning rate, allowing it to decrease progressively during training and thereby improving convergence performance. The backbone network was initialized with pretrained weights, and only the higher-level convolutional layers were fine-tuned, preserving the general feature extraction capability of lower layers. The data augmentation strategy was specifically tailored to the characteristics of cultural tourism surveillance videos. Techniques, including random cropping, horizontal flipping, color jittering, and temporal frame shuffling, were employed to increase data diversity, thereby alleviating overfitting caused by the limited scale of the self-constructed dataset.

3.2 Comparative experiments

The comparative experiments were divided into two categories: single-task comparisons and multi-task comparisons. Representative state-of-the-art single-task and multi-task methods in the field of image processing were selected as baselines. The superiority of the proposed TBA-IRNet framework was validated through quantitative evaluation across multiple performance metrics.

For single-task comparisons, three mainstream tourist behavior recognition approaches—convolutional neural network-based, spatial–temporal graph convolutional network, and Transformer-based methods—were selected. The experimental results are summarized in Table 2. As indicated by the results, the proposed method consistently outperforms all comparison methods across all evaluation metrics. Specifically, an overall accuracy of 91.2%, an overall mean Average Precision (mAP) of 89.6%, and an overall F1-score of 90.4% are achieved. Compared with the best-performing baseline (Transformer-based method), improvements of 4.3%, 4.9%, and 4.6% are observed in accuracy, mAP, and F1-score, respectively, demonstrating substantial performance gains. In terms of category-specific behavior recognition, superior performance is observed across all behavior categories. Notably, the recognition accuracy for grouping behavior exhibits the most significant improvement, reaching an increase of 6.5%. This enhancement can be attributed to the proposed bidirectional guidance mechanism integrating spatiotemporal graphs and region-level attention, which enables effective modeling of group interactions in densely populated scenarios and improves the recognition accuracy of complex behaviors. In terms of inference efficiency, the proposed method achieves a processing speed of 35 frames per second, outperforming ST-GCN (32 frames per second) and Transformer-based methods (29 frames per second), while remaining slightly below the CNN-based method (38 frames per second). This result indicates a favorable balance between recognition accuracy and inference efficiency, addressing the longstanding challenge in conventional approaches of trading off performance for computational efficiency.

Table 2. Comparison of tourist behavior recognition performance with existing single-task methods

Method Type	Overall Accuracy (%)	Overall Mean Average Precision (%)	Overall F1 (%)	Photographing Accuracy (%)	Dwelling Accuracy (%)	Walking Accuracy (%)	Conversing Accuracy (%)	Grouping Accuracy (%)	Frames Per Second
Convolutional neural network-based method	82.3	79.5	80.8	80.1	83.5	85.7	78.2	76.4	38
Spatial–temporal graph convolutional network	85.6	83.2	84.3	84.2	86.8	87.9	82.5	81.1	32
Transformer-based method	86.9	84.7	85.8	85.7	87.6	88.3	83.8	82.4	29
Proposed method	91.2	89.6	90.4	90.5	92.3	93.1	89.7	88.9	35

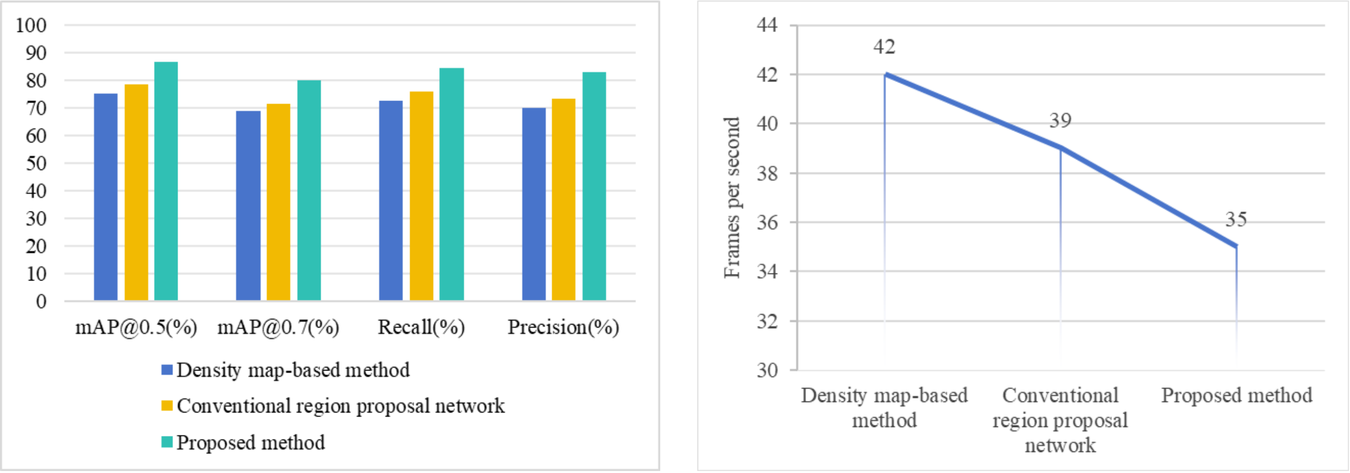


Figure 4. Comparison of interest region generation performance with existing single-task methods

For interest region generation, two representative single-task methods—density map-based and conventional region proposal network approaches—were selected for comparison. The experimental results are illustrated in Figure 4. As observed, the proposed method significantly outperforms both baseline methods across all key evaluation metrics. Specifically, mAP@0.5 and mAP@0.7 reach 86.7% and 79.8%, respectively, representing improvements of 8.1% and 8.3% over the conventional region proposal network method, and 11.4% and 10.9% over the density map-based method. In addition, recall and precision achieve 84.2% and 82.9%, respectively, both of which are substantially higher than those of the comparison methods. Notably, recall is improved by 8.4% relative to the conventional region proposal network method and by 11.6% relative to the density map-based method. These results strongly validate the effectiveness of the proposed behavior-aware weighting strategy and adaptive Gaussian kernel design. By incorporating behavioral semantics and temporal information, the fundamental limitation of conventional methods—namely, reliance solely on crowd density while neglecting actual tourist interest—has been effectively addressed, resulting in the generation of semantically meaningful interest regions. In terms of inference efficiency, a processing speed of 35 frames per second is achieved. Although this value is slightly lower than that of the density map-based method (42 frames per second) and the conventional region proposal network method (39 frames per second), it remains sufficient to satisfy real-time processing requirements in cultural tourism scenarios, thereby demonstrating the practical applicability and rationality of the proposed framework.

For multi-task comparisons, two representative joint learning approaches in the image processing domain—behavior recognition combined with object detection, and density estimation combined with region generation—were selected to evaluate the effectiveness of the proposed multi-task collaborative mechanism. The experimental results are presented in Figure 5. Existing multi-task methods generally suffer from insufficient task coupling, resulting in the inability to simultaneously optimize both tasks. Specifically, the behavior recognition + object detection method emphasizes behavior classification accuracy but neglects semantic consistency in interest region generation, yielding an interest region mAP@0.5 of only 79.3%. Conversely, the density estimation + region generation method focuses on region generation while failing to incorporate behavioral semantics, leading to a relatively low behavior recognition accuracy of 83.2%. In contrast, the proposed method achieves bidirectional collaborative enhancement between tasks through the multi-task interactive attention mechanism and the fully differentiable end-to-end optimization framework. As a result, a behavior recognition accuracy of 91.2% and an interest region mAP@0.5 of 86.7% are obtained, both of which significantly outperform the comparison methods. Specifically, behavior recognition accuracy is improved by 3.7% compared with the behavior recognition + object detection method, while interest region mAP@0.5 is improved by 5.2% compared with the density estimation + region generation method. Furthermore, an overall inference speed of 35 frames per second is maintained, which exceeds that of both comparison methods. These results further demonstrate the dual advantages of the proposed framework in terms of multi-task collaborative performance and computational efficiency, effectively overcoming the limitations of conventional multi-task approaches, where improvements in one task are often achieved at the expense of the other.

3.3 Ablation studies

Ablation experiments were conducted to validate the effectiveness of the proposed core modules. Four key components—namely, the collaborative attention module, the behavior-aware weighting mechanism, the bidirectional guidance mechanism, and the consistency loss—were individually removed from the full framework to construct four ablation variants. These variants were then compared with the complete model to quantify the contribution of each component. The experimental results are presented in Figure 6.

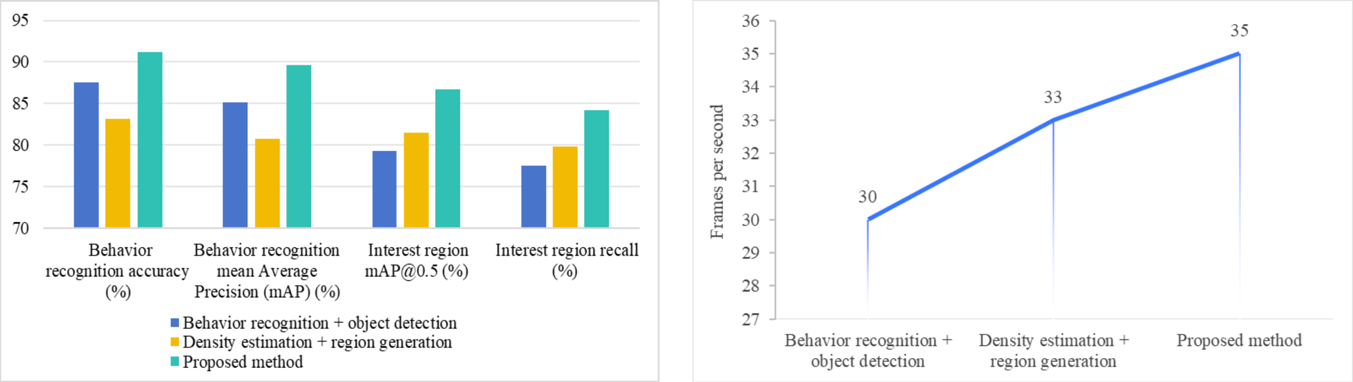
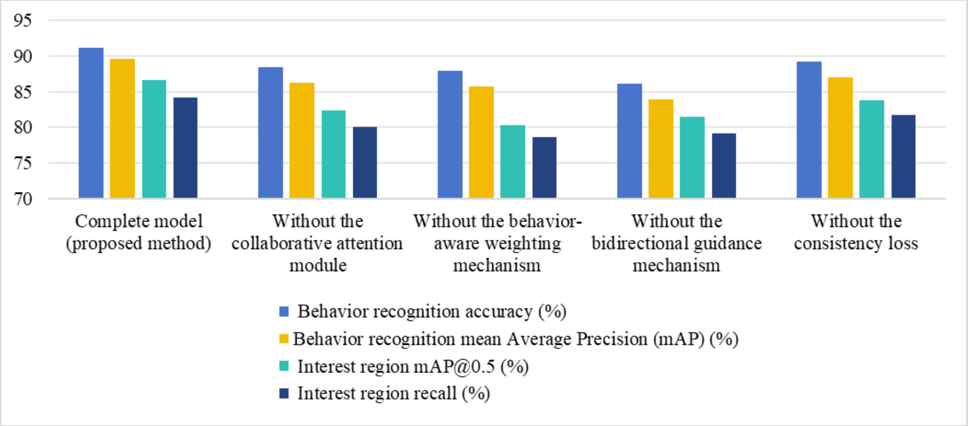


Figure 5. Comparison with existing multi-task methods



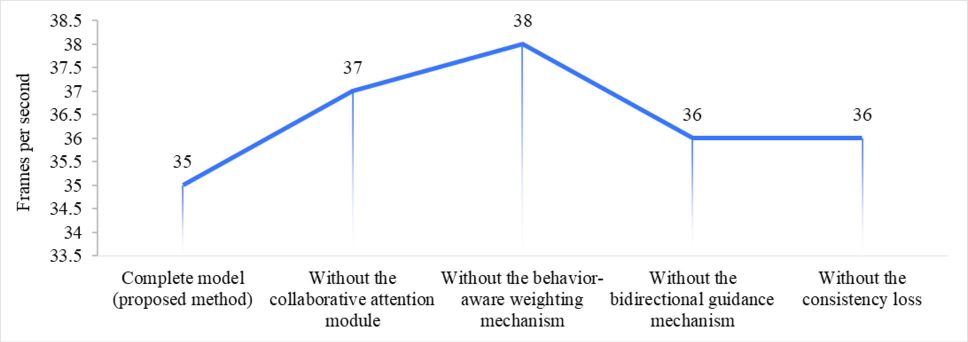
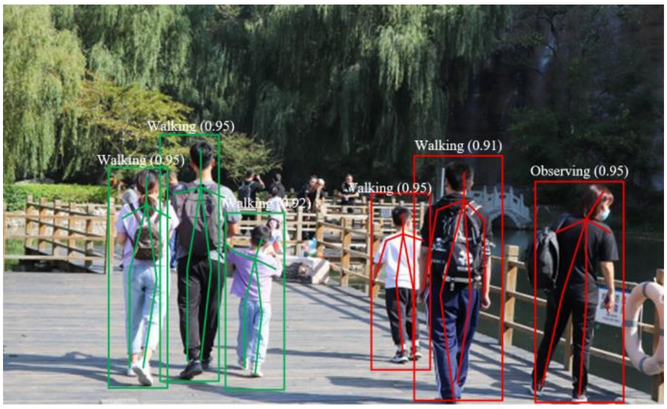


Figure 6. Results of ablation studies

As indicated by the results (Figure 6), the removal of any core module leads to noticeable performance degradation in both tourist behavior recognition and interest region generation, demonstrating that each module plays a critical role in improving overall performance and that their contributions are complementary. Specifically, when the collaborative attention module is removed, behavior recognition accuracy and interest region mAP@0.5 decrease by 2.7% and 4.3%, respectively. This observation indicates that the bidirectional information flow enabled by this module effectively facilitates cross-task interaction and mitigates information fragmentation. When the behavior-aware weighting mechanism is removed, interest region mAP@0.5 decreases by 6.4%, and behavior recognition accuracy decreases by 3.3%. These results highlight the importance of incorporating behavioral semantics and temporal information, confirming that the proposed weighting strategy effectively addresses the limitation of conventional methods that ignore semantic cues. When the bidirectional guidance mechanism is removed, behavior recognition accuracy and interest region mAP@0.5 decrease by 5.1% and 5.2%, respectively. This finding demonstrates that the integration of spatiotemporal graphs and region-level attention significantly enhances semantic modeling and improves both behavior recognition and interest region localization in complex scenarios. When the consistency loss is removed, performance degradation of approximately 2% is observed in both tasks. This result suggests that the consistency loss effectively enforces semantic alignment between the two tasks, promoting convergence toward similar feature distributions and further strengthening multi-task collaboration. Overall, the complete model achieves the best performance across all evaluation metrics, indicating that the proposed modules operate synergistically and provide complementary benefits, thereby substantially improving the overall capability of the framework.



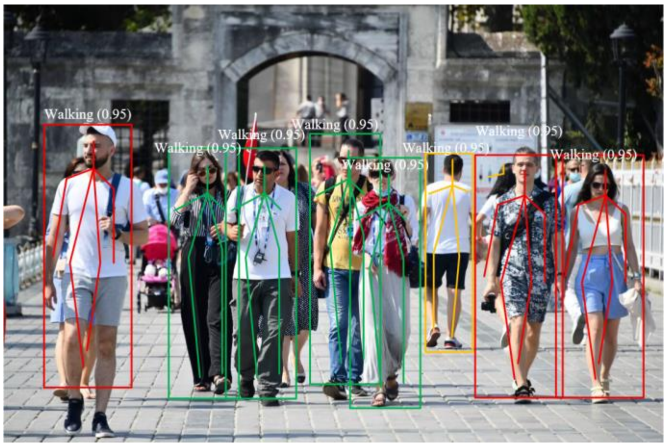


Figure 7. Visualization results of tourist behavior recognition

To validate the collaborative inference performance of the proposed framework in tourist detection, pose estimation, and behavior recognition under real-world cultural tourism scenarios, a visualization experiment was conducted using surveillance frames from a representative scenic area featuring traditional Chinese architecture. As illustrated in Figure 7, accurate detection of individual tourists is achieved in densely populated environments with complex backgrounds, without redundant detections or missed instances. In addition, pose estimation results include more than five visible keypoints per individual, along with complete skeletal connections, enabling detailed characterization of human motion. These outputs provide fine-grained spatiotemporal features that support robust tourist behavior recognition. Based on pose features and spatiotemporal graph modeling, accurate classification of multiple tourist behavior categories is achieved, with classification confidence scores exceeding 0.85. In regions characterized by group aggregation, behavior labels are assigned without category confusion, demonstrating the effectiveness of the proposed bidirectional attention guidance mechanism in modeling complex interaction patterns.

3.4 Parameter sensitivity analysis

To validate the rationality and robustness of key parameter settings in the proposed model, four critical parameters—behavior weight w_c, distance threshold θ_s, loss weight λ_3, and Gaussian kernel σ—were systematically varied. The impact of these parameters on model performance, including behavior recognition accuracy and interest region mAP@0.5, was analyzed. The experimental results are summarized in Table 3.

The results demonstrate that the selected parameter configurations exhibit strong rationality and robustness, and that their influence on model performance follows clear trends. For behavior weights, optimal performance is achieved when the weights are set as photographing = 0.8, dwelling = 0.5, and walking = 0.2. Excessively high weights result in overemphasis of specific behavioral semantics, while excessively low weights fail to effectively distinguish the contribution of different behaviors to interest region generation; both cases lead to performance degradation. For the distance threshold, the optimal value is observed at =50 pixels, which enables accurate identification of neighboring nodes and effective modeling of spatial interactions among tourists. Larger thresholds tend to introduce irrelevant nodes, whereas smaller thresholds may exclude meaningful interactions, both of which negatively affect behavior recognition accuracy. For loss weight, optimal performance is achieved at λ_3=0.3, which ensures a balanced learning priority between tourist behavior recognition and other tasks, thereby facilitating effective multi-task optimization. Values that are too high or too low lead to insufficient learning in one or more tasks, ultimately degrading overall performance. For the Gaussian kernel, the optimal configuration is achieved when σ=0.3×$max(w_i,h_i)$, which allows accurate alignment between the kernel scale and the size of each tourist instance, generating an interest intensity map with the highest spatial precision. Larger or smaller values of σ result in bias in the delineation of interest regions, thereby reducing the accuracy of interest region generation. Overall, the model exhibits a high degree of robustness with respect to variations in key parameters. The selected parameter settings are well adapted to the complex characteristics of cultural tourism scenarios, ensuring stable and reliable performance across diverse environments.

Table 3. Results of parameter sensitivity analysis

Parameter Type	Parameter Setting	Behavior Recognition Accuracy (%)	Interest Region mAP@0.5 (%)
Behavior weight	Photographing = 0.6, dwelling = 0.4, walking = 0.1	88.7	83.2
	Photographing = 0.8, dwelling = 0.5, walking = 0.2	91.2	86.7
	Photographing = 0.9, dwelling = 0.6, walking = 0.3	90.5	85.9
Distance threshold	40 pixels	89.8	84.5
	50 pixels	91.2	86.7
	60 pixels	90.1	85.3
Loss weight	0.2	89.5	85.1
	0.3	91.2	86.7
	0.4	90.3	86.1
Gaussian kernel	0.2×$max(w_i,h_i)$	88.9	83.7
	0.3×$max(w_i,h_i)$	91.2	86.7
	0.4×$max(w_i,h_i)$	90.4	85.8

3.5 Experimental conclusions

Based on the comprehensive results of comparative experiments, ablation studies, parameter sensitivity analysis, and visualization experiments, several conclusions can be drawn. The proposed TBA-IRNet framework demonstrates superior performance in both tourist behavior recognition and interest region generation tasks within cultural tourism scenarios. Across all key evaluation metrics, significant improvements are achieved compared with existing single-task and multi-task methods, indicating that a favorable balance between accuracy and computational efficiency has been successfully established. Each of the proposed core modules is shown to play a critical role in performance enhancement. The multi-task collaborative interactive attention mechanism effectively eliminates task boundaries and enables bidirectional information flow. The behavior-aware weighting mechanism incorporates behavioral semantics and temporal information, thereby addressing the limitation of conventional interest region generation methods that neglect semantic cues. The bidirectional guidance mechanism integrating spatiotemporal graphs and region-level attention enhances semantic modeling capability in complex scenarios. Furthermore, the fully differentiable end-to-end optimization framework ensures coordinated training and robust generalization performance. In addition, the proposed model exhibits strong real-time performance and robustness. The selected parameter configurations are demonstrated to be reasonable and well-suited to the complexity of cultural tourism environments. Overall, the proposed approach provides an efficient and practical technical solution for intelligent management and service optimization in digital cultural tourism platforms. Moreover, it offers a novel perspective and methodological reference for multi-task collaborative inference in the broader field of image processing.

4. DISCUSSION

A detailed analysis of the experimental results indicates that the significant performance improvements achieved by the proposed TBA-IRNet framework can be primarily attributed to four key innovations, which introduce theoretical advancements at the image processing level while being specifically tailored to the characteristics of cultural tourism scenarios. These innovations fundamentally distinguish the proposed approach from existing methods. First, the multi-task collaborative interactive attention mechanism effectively eliminates the task separation between tourist behavior recognition and interest region generation. Through bidirectional information flow, mutual feature enhancement is achieved, in contrast to conventional approaches that rely on unidirectional attention guidance or lack inter-task interaction altogether. This design significantly improves feature utilization efficiency and constitutes a critical factor enabling simultaneous performance gains in both tasks. Second, the behavior-aware weighting mechanism integrates behavioral semantics and temporal information into interest region generation. By moving beyond traditional density-based paradigms, interest regions are generated in a manner that more accurately reflects tourists' actual interests. This approach directly addresses the fundamental limitation of insufficient semantic correlation in existing methods. Third, the bidirectional guidance mechanism combining spatiotemporal graphs and region-level attention enhances semantic modeling capability in complex environments. Through refined graph structure design and attention-based feature fusion, both crowd interactions and temporal dynamics are accurately captured, thereby enabling effective adaptation to cultural tourism scenarios characterized by high crowd density and diverse behavioral patterns. Fourth, the fully differentiable end-to-end optimization framework ensures coordinated training across all modules. By eliminating non-differentiable components that typically introduce training discontinuities in conventional methods, improved convergence accuracy and generalization capability are achieved.

The unique characteristics of cultural tourism scenarios impose specific requirements on model design. High crowd density may lead to feature ambiguity, complex backgrounds can introduce substantial noise, and diverse behavioral patterns necessitate precise discrimination. These challenges are effectively addressed through the proposed framework by leveraging multi-scale feature extraction via the shared backbone network, scale adaptation through adaptive Gaussian kernels, and interaction modeling via multi-head attention mechanisms. As a result, strong adaptability to complex real-world scenarios is demonstrated.

The advantages of the proposed model are manifested in both theoretical contributions to image processing and practical value for cultural tourism applications. From the perspective of image processing theory, the fully differentiable design enables true end-to-end inference from pixel-level input to final output, ensuring seamless gradient backpropagation and providing a solid theoretical foundation for multi-task collaborative optimization. Furthermore, the proposed bidirectional attention mechanism introduces a novel paradigm for cross-task feature interaction, offering a new methodological perspective for multi-task collaborative inference in the field of image processing. From the perspective of practical application, the model demonstrates strong real-time performance and high accuracy, making it well-suited for real-time processing of surveillance video streams in digital cultural tourism platforms. Accurate recognition of tourist behaviors and precise identification of interest regions provide reliable support for scenic area management optimization, resource allocation, and service enhancement, thereby facilitating the digital and intelligent transformation of the cultural tourism industry. Despite these advantages, certain limitations are observed. In scenarios involving severe occlusion, the accuracy of pose keypoint estimation and appearance feature extraction is degraded, leading to a decline in behavior recognition performance. In addition, for small-scale tourist instances, behavioral features and spatial information are more susceptible to background interference, which reduces the precision of interest region generation. These limitations highlight areas requiring further improvement and underscore the objectivity and rigor of the study.

To address these limitations and align with emerging trends in image processing research, future work can be pursued along two main directions. From a technical perspective, the incorporation of Transformer-based architectures can be considered to enhance long-range temporal modeling, thereby improving the capture of behavioral dependencies and increasing robustness in occluded scenarios. In addition, multimodal information fusion techniques can be explored by integrating visual features with complementary modalities such as audio and environmental context, thereby enriching feature representation and mitigating interference caused by small-scale targets and complex backgrounds. From an application perspective, the generalizability of the proposed framework can be further investigated in other public environments, such as shopping malls, transportation hubs, and urban parks. Through appropriate parameter adaptation and scenario-specific optimization, the applicability of the framework can be extended. Moreover, the integration of edge computing techniques can be explored to enable model lightweighting and improve deployment efficiency on terminal devices. Such advancements would further enhance the practical value of the framework and promote the widespread adoption of multi-task image processing techniques in the intelligent analysis of public scenarios.

5. CONCLUSIONS

In response to the critical requirements of tourist behavior recognition and interest region generation in digital cultural tourism platforms and to address the limitations of existing methods—including task separation, insufficient semantic correlation, and limited generalization capability—an end-to-end multi-task learning framework, termed TBA-IRNet, was proposed. Centered on four key innovations, a comprehensive image processing and intelligent analysis framework was constructed. Bidirectional information flow between the two tasks was achieved through a multi-task collaborative interactive attention mechanism. Behavioral semantics and temporal information were incorporated via a behavior-aware weighting strategy. Semantic modeling in complex scenarios was enhanced through a bidirectional guidance mechanism integrating spatiotemporal graphs and region-level attention. Furthermore, coordinated training across all components was ensured by a fully differentiable end-to-end optimization framework. As a result, joint optimization of tourist behavior recognition and interest region generation was realized, effectively addressing the limitations of conventional methods in cultural tourism scenarios.

The contributions of this study are reflected in both theoretical advancement and practical application. From a theoretical perspective, a novel paradigm for multi-task collaborative interaction and behavior-driven region generation has been introduced. In addition, spatiotemporal graph modeling and end-to-end optimization strategies have been systematically enhanced, providing new insights and technical references for multi-task collaborative inference in the field of image processing. From an application perspective, the proposed framework enables accurate and efficient tourist behavior recognition and automatic interest region generation, demonstrating strong adaptability to cultural tourism environments characterized by high crowd density, complex backgrounds, and diverse behavioral patterns. This capability provides effective support for intelligent management, resource allocation, and service optimization in digital cultural tourism platforms. Experimental results demonstrate that the proposed method consistently outperforms existing single-task and multi-task approaches across all key evaluation metrics, while maintaining strong real-time performance and robustness. Overall, the TBA-IRNet framework effectively addresses the lack of specialized image processing methods for cultural tourism scenarios, achieving a balance between theoretical innovation and practical applicability. These findings provide important support for the digital transformation of the cultural tourism industry and for the scenario-driven application of image processing techniques.

REFERENCES

[1] Park, S.Y., Jamieson, W. (2009). Developing a tourism destination monitoring system: A case of the Hawaii tourism dashboard. Asia Pacific Journal of Tourism Research, 14(1): 39-57. https://doi.org/10.1080/10941660902728015

[2] Sobaś, E., Gawroński, K., Gawrońska, G., Janus, B. (2017). Monitoring of tourist movement as a basis for management and protection of attractive tourist areas on the example of the Popradzki landscape park. Acta Scientiarum Polonorum. Formatio Circumiectus, 16(3): 73-88. https://doi.org/10.15576/ASP.FC/2017.16.3.73

[3] Yao, J., Wang, J., Wang, Y., Hong, F. (2025). A tourist flow monitoring and management system for scenic areas using image recognition. Traitement du Signal, 42(2): 963-973. https://doi.org/10.18280/ts.420230

[4] Bai, S., Han, F. (2020). Tourist behavior recognition through scenic spot image retrieval based on image processing. Traitement du Signal, 37(4): 619-626. https://doi.org/10.18280/ts.370410

[5] Lin, S., Zhang, H., Wang, X., Lam, J.F. (2025). The impact of spatial perception at agricultural heritage sites on tourists' carbon reduction behavior. NPJ Heritage Science, 13(1): 190. https://doi.org/10.1038/s40494-025-01677-z

[6] Yoon, S., Gwon, G.H., Lee, J.H., Jung, H.J. (2021). Three-dimensional image coordinate-based missing region of interest area detection and damage localization for bridge visual inspection using unmanned aerial vehicles. Structural Health Monitoring, 20(4): 1462-1475. https://doi.org/10.1177/1475921720918675

[7] Yeum, C.M., Choi, J., Dyke, S.J. (2019). Automated region-of-interest localization and classification for vision-based visual assessment of civil infrastructure. Structural Health Monitoring, 18(3): 675-689. https://doi.org/10.1177/1475921718765419

[8] Camprubí, R., Coromina, L. (2019). The lighting dimension of perceived tourist image: The case of Barcelona. Current Issues in Tourism, 22(19): 2342-2347. https://doi.org/10.1080/13683500.2018.1508428

[9] Zhang, X. (2025). Analysis of tourist behaviour and tourism image in Hainan Province. Transformations In Business & Economics, 24(2(65)): 572-590. https://doi.org/10.15388/Tibe.2025.24.2.26

[10] Panigrahy, A., Verma, A. (2025). Tourist experiences: A systematic literature review of computer vision technologies in smart destination visits. Journal of Tourism Futures, 11(2): 187-202. https://doi.org/10.1108/JTF-04-2024-0073

[11] Zou, Z., Cheng, Y., Qu, X., Ji, S., Guo, X., Zhou, P. (2019). Attend to count: Crowd counting with adaptive capacity multi-scale CNNs. Neurocomputing, 367: 75-83. https://doi.org/10.1016/j.neucom.2019.08.009

[12] Wu, X., Zhao, Y., Luo, J., Zhang, M., Yang, W. (2022). Bas-relief modeling from RGB monocular images with regional division characteristics. Scientific Reports, 12(1): 21692. https://doi.org/10.1038/s41598-022-24974-0

[13] Wang, J., Zou, L., Fan, C., Chi, R. (2023). Part-wise adaptive topology graph convolutional network for skeleton-based action recognition. Electronics, 12(9): 1992. https://doi.org/10.3390/electronics12091992

[14] Qiao, F., Zhu, Y., Li, G., Li, B. (2024). Dual contrastive attention-guided deformable convolutional network for single image super-resolution. Journal of Visual Communication and Image Representation, 100: 104097. https://doi.org/10.1016/j.jvcir.2024.104097

[15] Jiao, L., Hu, C., Huo, L., Tang, P. (2021). Guided-Pix2Pix: End-to-end inference and refinement network for image dehazing. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 3052-3069. https://doi.org/10.1109/JSTARS.2021.3061460

[16] Zhang, W., Ren, Z., Zhou, J., Chen, S., et al. (2024). End-to-end automatic lens design with a differentiable diffraction model. Optics Express, 32(25): 44328-44345. https://doi.org/10.1364/OE.540590

[17] Gao, H., Deng, M., Zhao, W., Zhang, D. (2022). scene adaptive segmentation for crowd counting in population heterogeneous distribution. Applied Sciences, 12(10): 5183. https://doi.org/10.3390/app12105183

[18] Wang, Y., Sun, Z., Xu, D., Wu, L., et al. (2020). A hybrid method based region of interest segmentation for continuous wave terahertz imaging. Journal of Physics D: Applied Physics, 53(9): 095403. https://doi.org/10.1088/1361-6463/ab58b6

[19] Park, Y., Kim, Y., Kim, C., Lee, G.Y., et al. (2025). End-to-end optimization of metalens for broadband and wide-angle imaging. Advanced Optical Materials, 13(9): 2402853. https://doi.org/10.1002/adom.202402853

[20] Zhang, Q., Yu, Z., Liu, X., Wang, C., Zheng, Z. (2023). End-to-end joint optimization of metasurface and image processing for compact snapshot hyperspectral imaging. Optics Communications, 530: 129154. https://doi.org/10.1016/j.optcom.2022.129154